Enhancing Math Reasoning in Small-sized LLMs via Preview Difficulty-Aware Intervention

Xinhan Di [*1] **JoyJiaoW** [*2]

Abstract

Reinforcement learning scaling enhances the reasoning capabilities of large language models, with reinforcement learning serving as the key technique to draw out complex reasoning. However, key technical details of state-of-the-art reasoning LLMs—such as those in the OpenAI O series, Claude 3 series, DeepMind's Gemini 2.5 series, and Grok 3 series—remain undisclosed, making it difficult for the research community to replicate their reinforcement learning training results. Therefore, we start our study from an Early Preview Reinforcement Learning (EPRLI) algorithm built on the open-source GRPO framework, incorporating difficulty-aware intervention for math problems. Applied to a $1.5B$-parameter LLM, our method achieves 50.0% on AIME24, 89.2% on Math500, 77.1% on AMC, 35.3% on Minerva and 51.9% on OBench—super-pass O1-Preview and is comparable to O1-mini within standard school-lab settings.

1. Introduction

Large language models (LLMs) like OpenAI's o-series (OpenAI, 2025a;b), DeepSeek R1 (Guo et al., 2025), Claude 3.7 (Anthropic, 2025), Grok-3 (XAI, 2024), and Gemini 2.5 (LLC, 2025) excel at complex reasoning tasks such as math and code generation. These capabilities are often acquired via large-scale reinforcement learning (RL), incorporating strategies like step-by-step reasoning (Wei et al., 2022), self-reflection (Wang et al., 2023), and backtracking (Ahmadian et al., 2024). However, enhancing reasoning in small models remains difficult. To address this, we propose a preview difficulty-aware intervention-based RL algorithm that improves the math reasoning ability in small-sized large language models. Our $1.5B$ LLM, trained with an early preview of the proposed GRPO algorithm with difficulty-aware intervention, outperforms OpenAI's O1-Preview and O1-mini (OpenAI, 2024; Jaech et al., 2024) on major math reasoning benchmarks (Guo et al., 2025; Christiano et al., 2017; Everitt et al., 2021; Weng, 2024).

2. Related Work

2.1. Reasoning Large Language Models

Reinforcement learning (RL) has been widely applied to align LLMs with human preferences (Christiano et al., 2017; Ouyang et al., 2022; Yuan et al., 2024a; Azar et al., 2024; Rafailov et al., 2023), while the open-source community has mainly relied on imitation learning (Yuan et al., 2024b; Yue et al., 2023; Guan et al., 2025) to improve reasoning. Recently, the trend has shifted toward RL, with OpenAI o1 (Jaech et al., 2024) demonstrating its promise, and later works confirming the scalability of outcome-reward-based RL (Guo et al., 2025; Qwen Team, 2024; XAI, 2024). Despite this, dense reward methods remain underexplored, as highlighted by PRIME (Cui et al., 2025), while most RL applications still use outcome reward models (ORMs) (Rafailov et al., 2023; Shao et al., 2024; Guo et al., 2025). Top-performing models—OpenAI's o-series (Jaech et al., 2024; OpenAI, 2024; 2025a;b), DeepSeek R1 (Guo et al., 2025), Claude 3.7 (Anthropic, 2025), Grok-3 (XAI, 2024), and Gemini 2.5 (LLC, 2025)—excel in reasoning tasks. However, deep reinforcement learning are not well studied for boosting reasoning in math problems in small LLMs (0.7B–1.5B) trained on limited math data with difficulty-aware intervention.

3. Method

3.1. Early Preview Group Relative Policy (Shao et al., 2024) Optimization(GRPO) with Difficulty-Aware Intervention

We define a discrete-time finite-horizon discounted Markov decision process (MDP) by a tuple $M = (S, A, \mathcal{P}, r, \rho_0, \gamma, H)$, where S is a state set, A is an action set, $\mathcal{P} : S \times A \times S \to \mathbb{R}_+$ is the transition probability

[*]Equal contribution [1]AI Lab, Giant Network, Shanghai, China [2]DeepReason, Shanghai, China. Correspondence to: JoyJiaoW <deepreasoninggo@gmail.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

distribution, $\gamma \in [0,1]$ is a discount factor, and H the horizon. Our objective is to find a stochastic policy π_θ that maximizes the expected discounted return within the MDP, $\eta(\pi_\theta) = \mathbb{E}_\tau \left[\sum_{t=0}^{H} \gamma^t r(s_t, a_t)\right]$. We use $\tau = (s_0, a_0, \ldots)$ to denote the entire state-action trajectory, where $s_0 \sim \rho_0(s_0), a_t \sim \pi_\theta(a_t|s_t), s_{t+1} \sim \mathcal{P}(s_{t+1}|s_t, a_t)$.

In this work, we propose a method to learn a hierarchical policy and efficiently adapt all the levels in the hierarchy to perform a new task. We study hierarchical policies composed of a higher level, or manager $\pi_{\theta_{high}}(a_{t^{high}}|s_{t^{high}})$, and a lower level, or sub-policy $\pi_{\theta_{low}}(a_{t^{low}}|s_{t^{low}})$. Both the higher level and the lower level take actions in the environment directly. The manager typically operates at a lower frequency than the sub-policies.

It is important to note that the hierarchical architecture in our preview version framework is composed of L discrete levels, where each level is indexed by $l \in \{0,1\}$. In this configuration, a higher-level policy corresponds to high $=$ $l+1$ and its corresponding lower-level policy is defined as low $= l$. This structure allows for top-down coordination, in which, higher levels guide the behavior and planning strategies of the lower ones.

$$\mathcal{J}_{\text{GRPO}^{Her}}(\theta) = \prod_{l=1}^{2} \mathbb{E}_{q^l \sim D_q^l,\, \{o_i^l\}_{i=1}^{G^l} \sim \pi_\theta(\cdot|q^l)}$$
$$\prod_{l=1}^{L} \left[\frac{1}{\sum_{i^l=1}^{G^l} |o_i^l|} \sum_{i^l=1}^{G^l} \sum_{j^l=1}^{|o_i^l|} \min \left(\frac{\pi_\theta(o_i^l \mid q^l)}{\pi_{\theta_{\text{old}}}(o_i^l \mid q^l)} A_{i^l, j^l}^l, \right. \right. \quad (1)$$
$$\left. \left. \text{clip} \left(\frac{\pi_\theta(o_i^l \mid q^l)}{\pi_{\theta_{\text{old}}}(o_i^l \mid q^l)}, 1 - \varepsilon_{\text{low}}^l, 1 + \varepsilon_{\text{high}}^l \right) A_{i^l, j^l}^l \right) \right]$$

In the early preview version of the proposed Group Relative Policy (Shao et al., 2024) Optimization(GRPO) with Difficulty-Aware Intervention, we adopt a simplification of the underlying Markov Decision Process (MDP). Specifically, we assume that the policies across the high and low levels share the same parameterization. This simplification leads to the expression $\pi_{\theta_{\text{high}}}(a_{t^{\text{high}}} \mid s_{t^{\text{high}}}) = \pi_{\theta_{\text{low}}}(a_{t^{\text{low}}} \mid s_{t^{\text{low}}}) = \pi_\theta$, where both high- and low-level policies are treated uniformly under the shared policy π_θ. This unified parameterization not only reduces the model complexity but also facilitates efficient training and inference within the hierarchical structure.

3.1.1. Reformulation

Then, we propose the reformulation of the early preview group relative policy (Shao et al., 2024) optimization(GRPO) with difficulty-aware intervention (EPRLI) samples a group of outputs $\{o_i^l\}_{i=1}^{G^l}$ for each question q_i^l paired with the answer a^l, $l = \{0,1\}$ and optimizes the policy via the objective represented as equation 1, where L is the total number of levels in the early preview group relative policy (Shao et al., 2024) optimization(GRPO) with difficulty-aware intervention (EPRLI) algorithm 1. Similarly, l denotes the index of the lever in the $L = 2$ hierarchy, which means there are total $L = 2$ hierarchy in EPRLI.

3.1.2. Implementation

Then, we implement our proposed preview early preview group relative policy optimization(GRPO) with Difficulty-Aware Intervention(EPRLI) of reasoning LLMs. Particularly, the implementation is made to take the difficulty of the reasoning tasks in accordance with the hierarchy in the proposed early preview EPRLI. The details of the proposed implementations(Algorithm 1) is represented as the following:

Early Preview EPRLI(Algorithm 1) Implementation In the implementation of the Early Preview Group Relative Policy Optimization(GRPO) with Difficulty-Aware Intervention framework, the hierarchy is structured with a total of two levels specifically designed to tackle mathematical reasoning problems. This two-level hierarchical design is crucial to effectively manage the complexity inherent in such tasks. More precisely, the quality values at different levels satisfy the relationships $Q^1 < Q^2$, Additionally, the maximum allowable sequence lengths follow a strict increasing order such that $\boldsymbol{Len^2_{max}} < \boldsymbol{Len^1_{max}}$. This setup enables the preview algorithm to handle math reasoning problems of varying lengths/difficulties while maintaining the preview hierarchical learning.

Furthermore, the hierarchical policies H^1, H^2 are designed with a dominant influence order such that $H^1 \gg H^2$, meaning the top-level policy has the greater guiding power in the reasoning process, while the subsequent levels exert progressively less influence. So, optimizing this reasoning trajectory through carefully controlled hierarchical interactions.

4. Experiment

To investigate the effectiveness of the two proposed implementations of the early preview hierarchical GRPO on the reasoning capabilities of large language models (LLMs), we conduct a series of experiments. The experiments are designed to provide a comparative analysis against the state-of-the-art reasoning-oriented LLMs of different parameters, in particular, DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025), STILL-3-1.5B-Preview (RUC-AIBOX, 2025), DeepScaler-1.5B-Preview (Luo et al., 2025), FastCuRL-1.5B-Preview (Chen et al., 2025) with 1.5B parameters, Qwen3-4B (Yang et al., 2025), DeepSeek-R1-Distill-Qwen-

Algorithm 1 Early Preview EPRLI: Early Preview Group Relative Policy Optimization(GRPO) with Difficulty-Aware Intervention

```
Require: initial policy model $\pi_\theta$; reward model $\{R^l\}$; task prompts $\{\mathcal{D}^l\}$ with corresponding difficulty
    level $\{\mathcal{Q}^l\}$; hyperparameters $\{\varepsilon^l_{\text{low}}\}, \{\varepsilon^l_{\text{high}}\}$, $l = 1, 2, \ldots, L$, $\boldsymbol{Q^{l-1} > Q^l}$. Length Reward $\{\mathcal{K}^l\}$ with
    corresponding max length $\{\boldsymbol{Len^l_{\mathbf{max}}}\}$, $\boldsymbol{Len^{l-1}_{\mathbf{max}} = Len^l_{\mathbf{max}}}$.
Ensure: $\pi_\theta$
for $l = 1, \ldots, L$ do
    for step $= 1, \ldots, M$ do
        Sample a batch $\mathcal{D}^l_b$ from $\mathcal{D}^l$
        Update the old policy model $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
        Sample $G^l$ outputs $\{o^l_i\}_{i^l=1}^{G^l} \sim \pi_{\theta_{\text{old}}}(\cdot \mid q^l)$ for each question $q^l \in \mathcal{D}^l_b$
        Compute rewards $\{{r^l}_{i^l}\}_{i^l=1}^{G^l}$ for each sampled output $o^l_i$ by running $R^l$
        Filter out $o^l_i$ and add the remaining to the dynamic sampling buffer.
        if buffer size $n^l_b < N^l$ then
            continue
        end if
        For each $o^l_i$ in the buffer, compute $\hat{A}^l_{i^l,t^l}$ for the $t^l$-th token of $o^l_i$
    end for
    for iteration $= 1, \ldots, \mu^l$ do
        Update the policy model $\pi_\theta$ by maximizing the GRPO+ objective combining with Length Reward
    $\boldsymbol{\mathcal{K}^l}$
    end for
end for=0
```

7B (Guo et al., 2025), MIMO-7B (Xiaomi LLM-Core Team, 2025) with middle-sized parameters, Llama 4 Maverick (AI, 2025b), Phi4-Reasoning-14B (Abdin et al., 2025), Qwen 2.5-72B (Team, 2024), Kimi-1.5 (Team, 2025a), Llama 4 Behemoth (AI, 2025a), Qwen3-235B (Team, 2025b), DeepSeek-R1 (Guo et al., 2025) with large-sized parameters, and closed-source reasoning models such as Claude 3.7 Sonnet (Standard) (Anthropic, 2025), O1, O1-Mini (OpenAI, 2024a), and O1-Preview (OpenAI, 2024b), enabling a thorough evaluation of the proposed early preview methods.

4.1. Experiment Setup

We choose DeepScaler-1.5B-Preview-16k (Luo et al., 2025) as our base model, which is a $1.5B$ parameter model. We utilize the AdamW (Loshchilov & Hutter, 2019) optimizer with a constant learning rate of 1×10^6 for optimization. For the roll-outs, we set the temperature to 0.6 and sample 16 responses per prompt. In this experiment, 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.

4.2. Benchmarks

Math Reasoning Benchmark To better evaluate the trained model, we have selected five benchmarks to assess its performance: MATH 500 (Hendrycks et al., 2021), AIME 2024 (AI-MO, 2024a), AMC 2023 (AI-MO, 2024b), Minerva Math (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024b).

4.3. Dataset and Evaluation Metric

Math Reasoning Dataset The training dataset is consisted of $40K$ problems with two difficulty levels. Particularly, it is consisted of AIME (of America, 2024) (American Invitational Mathematics Examination) problems (1984-2023), AMC (of America, 2025) (American Mathematics Competition) problems (prior to 2023), Omni-MATH (Gao et al., 2024) dataset and Still dataset (RUC-AIBOX, 2025). We set the maximum generation length for the models to 32768 tokens and leverage 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: $\text{PASS@}1 = \frac{1}{k}\sum_{i=1}^{k} p_i$.

4.4. Math Reasoning Experiments

The proposed hierarchical reasoning model is evaluated against both open-source and closed-source state-of-the-art reasoning models, including O1-Preview (OpenAI, 2024b), O1-Mini (OpenAI, 2024a), O1 (OpenAI, 2024a), Claude 3.7 Sonnet (Anthropic, 2025), and others. As shown in Table 1, our $1.5B$ model achieves impressive performance

Table 1. Model Performance Comparison

Model	**MATH500**	**AIME24**	**AMC**	**Minerva**	**OBench**	**Avg.**
Close-Source						
O1-Preview (OpenAI, 2024b)	85.5	44.6	–	–	–	–
O1-Mini (OpenAI, 2024a)	90.0	70.0	–	–	–	–
O1 (OpenAI, 2024a)	90.4	71.5	–	–	–	–
Claude 3.7 Sonnet (Standard) (Anthropic, 2025)	82.2	23.3	–	–	–	–
Open-Source-Large						
DeepSeek-R1 (Guo et al., 2025)	97.3	79.8	–	–	–	–
Qwen3-235B (Team, 2025b)	94.6	85.7	–	–	–	–
Llama 4 Behemoth (AI, 2025a)	95.0	78.0	–	–	–	–
Kimi-1.5 (Team, 2025a)	96.2	77.5	–	–	–	–
Qwen 2.5-72B (Team, 2024)	83.1	30.0	–	–	–	–
Phi4-Reasoning-14B (Abdin et al., 2025)	–	81.3	–	–	–	–
Llama 4 Maverick (AI, 2025b)	18.0	64.0	–	–	–	–
Open-Source-4B/7B						
MIMO-7B (Xiaomi LLM-Core Team, 2025)	95.8	68.2	–	–	–	–
DeekSeek-Qwen-Distill-7B (Guo et al., 2025)	92.8	55.5	–	–	–	–
Qwen3-4B (Yang et al., 2025)	-	73.8	–	–	–	–
Open-Source-1.5B						
DeepSeek-R1-Distill-QWEN-1.5B (Guo et al., 2025)	82.8	28.8	62.9	26.5	43.3	48.9
STILL-3-1.5B-Preview (RUC-AIBOX, 2025)	84.4	32.5	66.7	29.0	45.4	51.6
DeepScaler-1.5B-Preview (Luo et al., 2025)	87.8	43.1	73.6	30.2	50.0	57.0
FastCuRL-1.5B-Preview (Chen et al., 2025)	88.0	43.1	74.2	31.6	50.4	57.5
Ours2-1.5B Algorithm 1	**89.2**	**50.0**	**77.1**	**35.3**	**51.9**	**60.7**

across multiple benchmarks: 50.0 Pass@1 on AIME24 (Jia, 2025), 89.2 on MATH500 (HuggingFaceH4, 2025), 74.7 on AMC23 (of America, 2023), 35.3 on Minerva (Dyer & Gur-Ari, 2022), and 51.9 on OlympiadBench (He et al., 2024a). These results demonstrate the model's robust general reasoning ability across various mathematical and competition-level tasks.

Notably, the reinforcement learning training strategy with preview difficulty-aware intervention enables our $1.5B$ model to outperform the current best-performing $1.5B$ reasoning model by 6.9 points on AIME24 (Jia, 2025), 1.4 points on MATH500 (HuggingFaceH4, 2025), 1.1 on AMC23 (of America, 2023), 4.1 on Minerva (Dyer & Gur-Ari, 2022), and 1.9 on OlympiadBench (He et al., 2024a) —averaging a 3.7-point gain overall. Furthermore, it surpasses several larger parameter models, including O1-Preview (OpenAI, 2024b), and is comparable with O1-2024-12-17 (Low) (OpenAI, 2024a).

5. Discussion

We initiate an exploration of reinforcement learning for improving the reasoning capabilities of large language models (LLMs) by introducing a preview difficulty-aware intervention strategy based on reinforcement learning, specifically tailored for mathematical problem-solving tasks. Despite being applied to a relatively small-scale math dataset, our approach demonstrates reasoning ability improvements: our 1.5B parameter model not only surpasses OpenAI's O1-Preview (OpenAI, 2024b) but also approaches the performance level of the stronger O1-Mini model (OpenAI, 2024a).

We plan to further develop the framework to support both small- and mid-sized models, with a longer-term goal of developing a unified reasoning agent that can excel across domains, including mathematical and coding tasks. To foster transparency and accelerate progress in this area, we commit to open-sourcing to provide the community with tools and benchmarks to advance the study of reasoning in LLMs under resource-efficient settings.

References

Abdin, M., Agarwal, S., Awadallah, A., Balachandran, V., Behl, H., Chen, L., de Rosa, G., Gunasekar, S., Javaheripi, M., Joshi, N., Kauffmann, P., Lara, Y., Mendes, C. C. T., Mitra, A., Nushi, B., Papailiopoulos, D., Saarikivi, O., Shah, S., Shrivastava, V., Vineet, V., Wu, Y., Yousefi, S., and Zheng, G. Phi-4-reasoning technical report. https://arxiv.org/abs/2504.21318, 2025. arXiv preprint arXiv:2504.21318.

Ahmadian, A., Cremer, C., Gallé, M., Fadaee, M., Kreutzer, J., Pietquin, O., Üstün, A., and Hooker, S. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024. URL https://arxiv.org/abs/2402.14740.

AI, M. Llama 4: The beginning of a new era of natively multimodal ai, 2025a. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2025-05-21.

AI, M. Llama 4 maverick: A natively multimodal mixture-of-experts model. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025b. Accessed: 2025-05-21.

AI-MO. Aime 2024. https://huggingface.co/datasets/AI-MO/aimo-validation-aime, 2024a.

AI-MO. Amc 2023. https://huggingface.co/datasets/AI-MO/aimo-validation-amc, 2024b.

Anthropic. Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/claude/sonnet, 2025. Accessed 2025-05-13.

Anthropic. Claude 3.7 sonnet: The first hybrid reasoning model, 2025. URL https://www.anthropic.com/news/claude-3-7-sonnet. Accessed: 2025-05-21.

Azar, M. G., Rowland, M., Piot, B., Guo, D., Calandriello, D., Valko, M., and Munos, R. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, 2024. abs/2310.12036.

Chen, Y., Yang, N., Luo, Z., He, J., Zhang, M., and Wang, L. Fastcurl: Curriculum reinforcement learning with progressive context extension for efficient training of r1-like reasoning models. *arXiv preprint arXiv:2503.17287*, 2025. URL https://arxiv.org/abs/2503.17287.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. 2017.

Cui, G., Yuan, L., Wang, Z., Wang, H., Li, W., He, B., Fan, Y., Yu, T., Xu, Q., Chen, W., et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

Dyer, E. and Gur-Ari, G. Minerva: Solving quantitative reasoning problems with language models. *Google Research Blog*, 2022. Accessed: 2025-05-20.

Everitt, T., Hutter, M., Kumar, R., and Krakovna, V. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *arXiv preprint arXiv:2105.00901*, 2021.

Gao, B., Song, F., Yang, Z., Cai, Z., Miao, Y., Dong, Q., Li, L., Ma, C., Chen, L., Xu, R., Tang, Z., Wang, B., Zan, D., Quan, S., Zhang, G., Sha, L., Zhang, Y., Ren, X., Liu, T., and Chang, B. Omni-math: A universal olympiad level mathematic benchmark for large language models, 2024. URL https://arxiv.org/abs/2410.07985.

Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., and Yang, M. Rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*, 2025.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., and et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

He, C., Luo, R., Bai, Y., Hu, S., Thai, Z., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., Liu, J., Qi, L., Liu, Z., and Sun, M. 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)*, pp. 3828–3850, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

He, C., Luo, R., Bai, Y., Hu, S., Thai, Z., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., Liu, J., Qi, L., Liu, Z., and Sun, M. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the ACL (Long Papers)*, pp. 3828–3850, Bangkok, Thailand, 2024b. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, Virtual Event, 2021. OpenReview.net. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

HuggingFaceH4. Math-500: A subset of the math benchmark. https://huggingface.co/datasets/HuggingFaceH4/MATH-500, 2025. Accessed: 2025-05-20.

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., and Carney, A. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Jia, M. Aime 2024 dataset. https://huggingface.co/datasets/Maxwell-Jia/AIME_2024, 2025. Accessed: 2025-05-20.

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., and Misra, V. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, June 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/18abbeef8cfe9203fdf9053c9c4fe191-Paper-Conference.pdf. ArXiv preprint arXiv:2206.14858.

LLC, G. Gemini 2.5 Flash Preview (Thinking). https://ai.google.dev/gemini-api/docs/changelog, April 2025. Preview: gemini-2.5-flash-preview-04-17; Accessed: 2025-05-13.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, 2019. OpenReview.net. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Luo, M., Tan, S., Wong, J., Shi, X., Tang, W., Roongta, M., Cai, C., Luo, J., Zhang, T., Li, E., Popa, R. A., and Stoica, I. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. https://github.com/agentica-project/deepscaler, 2025. Notion Blog.

of America, M. A. 2023 american mathematics competitions (amc 8, amc 10, amc 12). https://maa.org/student-programs/amc, 2023. Accessed: 2025-05-20.

of America, M. A. Aime24: American invitational mathematics examination 2024, 2024. URL https://www.maa.org/math-competitions/aime. Accessed: 2025-05-20.

of America, M. A. American mathematics competitions, 2025. URL https://maa.org/student-programs/amc/. Accessed: 2025-05-20.

OpenAI. Introducing openai o1: A reasoning-focused ai model, 2024a. URL https://openai.com/o1. Accessed: 2025-05-21.

OpenAI. Introducing openai o1-preview: A reasoning-focused ai model, 2024b. URL https://openai.com/index/introducing-openai-o1-preview/. Accessed: 2025-05-21.

OpenAI. Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/, 2024. Accessed 2025-05-13.

OpenAI. Openai o1-2024-12-17 (low reasoning effort). https://openai.com/index/o1-and-new-tools-for-developers/, December 2024a. Accessed: 2025-05-20.

OpenAI. Introducing openai o1-preview: A new series of reasoning models, September 2024b. URL https://openai.com/index/introducing-openai-o1-preview/. Accessed: 2025-05-20.

OpenAI. o3-mini. https://platform.openai.com/docs/models, 2025a. Accessed: 2025-05-13.

OpenAI. o4-mini. https://platform.openai.com/docs/models/o4-mini, April 2025b. Accessed: 2025-05-13.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., and et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning. https://qwenlm.github.io/blog/qwq-32b-preview/, 2024.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

RUC-AIBOX. Still-3-preview-rl-data: A dataset for hierarchical reinforcement learning in large language models, 2025. URL https://huggingface.co/datasets/RUC-AIBOX/STILL-3-Preview-RL-Data. Accessed: 2025-05-20.

Shao, Q., Wang, P., Zhu, R., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., and Wu, Yiming, e. a. Deepseek-math: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Team, K. Kimi k1.5: Scaling reinforcement learning with llms, 2025a. URL https://arxiv.org/abs/2501.12599. Accessed: 2025-05-21.

Team, Q. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115. Accessed: 2025-05-21.

Team, Q. Qwen3-235b-a22b: A 235b parameter mixture-of-experts model, 2025b. URL https://huggingface.co/Qwen/Qwen3-235B-A22B. Accessed: 2025-05-21.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain-of-thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., and et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. 2022.

Weng, L. Reward hacking in reinforcement learning. https://lilianweng.github.io/lil-log/2024/11/01/reward-hacking.html, 2024.

XAI. Grok 3 beta — the age of reasoning agents. https://x.ai/news/grok-3, 2024.

Xiaomi LLM-Core Team. Mimo: Unlocking the reasoning potential of language model – from pretraining to post-training, 2025. URL https://arxiv.org/abs/2505.07608.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Yuan, L., Cui, G., Wang, H., Ding, N., Wang, X., Deng, J., Shan, B., Chen, H., Xie, R., Lin, Y., Liu, Z., Zhou, B., Peng, H., Liu, Z., and Sun, M. Advancing llm reasoning generalists with preference trees. *arXiv preprint*, 2024a.

Yuan, L., Li, W., Chen, H., Cui, G., Ding, N., Zhang, K., Zhou, B., Liu, Z., and Peng, H. Free process rewards without process labels. *arXiv preprint*, 2024b. URL https://arxiv.org/abs/2412.01981.

Yue, X., Qu, X., Zhang, G., Fu, Y., Huang, W., Sun, H., Su, Y., and Chen, W. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.