RULE-BASED REFERENCE UPDATES AFTER R1-BASED POST REINFORCEMENT LEARNING FOR SMALL REASONING LANGUAGE MODELS.

Anonymous authors
Paper under double-blind review

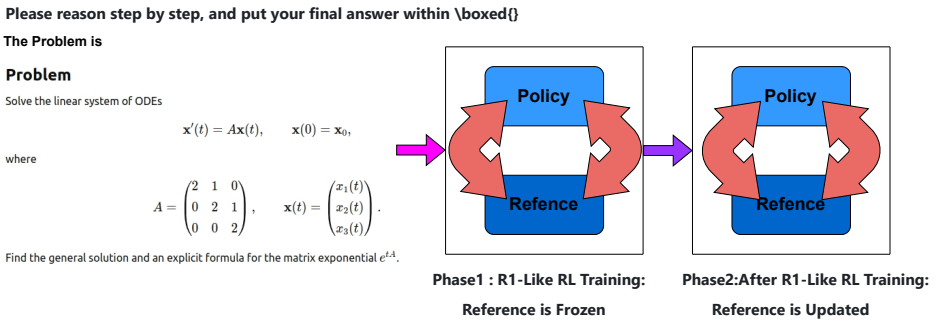

Figure 1: From Phase 1 : R1-Like Reinforcement Learning and its Curriculum Learning Variants To Phase 2: Reference Model Updates in Reinforcement Learning.

ABSTRACT

Inference scaling improves LLM reasoning, with reinforcement learning as a key driver. Although, post-training reinforcement learning and its curriculum learning variants offer significant benefits in enhancing the reasoning ability of large language models, we designate this process as Phase 1. Following this, we propose Phase 2: rule-based reference model updates in reinforcement learning after Phase 1 to explore the potential of reference model updates following R1-Like reinforcement learning. In details, we introduce a rule-based reference updates reinforcement learning approach continues to enhance the reasoning capabilities of small-sized large language models after current classical post-training reinforcement learning. In particular, a $1.5B$-parameter LLM achieves 60.2% on AIME24, 48.2% on AIME25 and 91.5% on Math500 and $1.5\% - 4\%$ score improvement on AMC, Minera and Olympia. These results, enabled by the proposed rule-based reference model updates reinforcement learning algorithm, demonstrate math reasoning capabilities comparable to O1-mini/O3-mini—achievable within a typical school laboratory setting. In addition, we open-source both the dataset and model checkpoints to support future research in large-scale reinforcement learning for LLMs.

1 INTRODUCTION

Large language models (LLMs) like OpenAI's o-series Jaech et al. (2024); OpenAI (2024; 2025c;d), DeepSeek R1 Guo et al. (2025), Claude 3.7 Anthropic (2025), Grok3 XAI (2024), and Gemini 2.5 LLC (2025) excel in tasks such as math reasoning and code generation. These models leverage large-scale reinforcement learning (RL) to acquire advanced reasoning strategies—step-by-step analysis Wei et al. (2022), self-reflection Wang et al. (2023), and backtracking Ahmadian et al. (2024). While prior RL success has relied on large base models, enhancing reasoning in small models remains challenging. To address this, we explore the reference model updates after R1-Like reinforcement learning and its curriculum learning variants approaches Guo et al. (2025); Christiano et al. (2017);

Sutton & Barto (2018); Everitt et al. (2017; 2021); Weng (2024) for improving reasoning capabilities in smaller LLMs.

In this work, we introduce the rule-based reference model updates in GRPO algorithm, a post-training reinforcement learning approach tailored for small to medium-sized LLMs after the process of R1-Like reinforcement learning and its curriculum learning variants Figure 1. Our 1.5B-parameter models trained with this method outperform larger closed- and open-source reasoning models—including OpenAI's O1-mini and O1 OpenAI (2024); Jaech et al. (2024)—on key math reasoning benchmarks, demonstrating strong reasoning capabilities.

2 Related Work

2.1 Reasoning Large Language Models

Reinforcement learning (RL) has been key to aligning 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 open-source efforts often rely on imitation learning for reasoning Yuan et al. (2024b); Yue et al. (2023); Guan et al. (2025). Recent models like OpenAI's O1 Jaech et al. (2024), DeepSeek R1 Guo et al. (2025), and Grok-3 XAI (2024) demonstrate the scalability of outcome-based RL. PRIME Cui et al. (2025) explores dense rewards, in contrast to the dominant outcome-only approaches Guo et al. (2025); Rafailov et al. (2023); Shao et al. (2024). Top models such as OpenAI's O-series Jaech et al. (2024); OpenAI (2024; 2025c;d), Claude 3.7 Anthropic (2025), and Gemini 2.5 LLC (2025) show strong math Guo et al. (2025); Jaech et al. (2024) and coding OpenAI (2025d); LLC (2025) reasoning. Long-COT models like O3 and DeepSeek-R1 OpenAI (2025c); Guo et al. (2025) benefit from RL with verifiable rewards (RLVR) Gandhi et al. (2025), avoiding costly MCTS-based data Hosseini et al. (2024); Yang et al. (2024), though they often overthink simple tasks Wang et al. (2024); Kumar et al. (2025). Efficiency-focused alternatives include latent-space optimization Hao et al. (2024); Geiping et al. (2025) and early-exit strategies Muennighoff et al. (2025); Fu et al. (2024); Zhang et al. (2024). But deep RL remains underexplored for small-scale LLMs (0.7B–1.5B) trained with limited data and resources. Besides, the updates of reference model in the reinforcement learning for enhancing small-scale LLM reasoning with limited math data is not well explored.

2.2 Hierarchical Reinforcement Learning and Curriculum Learning

Hierarchical Reinforcement Learning (HRL) supports temporal abstraction and efficient exploration Sutton & Barto (2018); Nachum et al. (2018), using frameworks like options Sutton & Barto (2018); Bacon et al. (2017); Harutyunyan et al. (2018); Klissarov et al. (2017); Kaelbling (1993); Gao et al. (2024b); Dayan & Hinton (1993a); Salter et al. (2022) and goal-conditioned feudal models Dayan & Hinton (1993b); Vezhnevets et al. (2017). Techniques like transition relabeling Nachum et al. (2018); Levy et al. (2018) and leveraging demonstrations Rajeswaran et al. (2018); Nair et al. (2018); Hester et al. (2018); Shiarlis et al. (2018); Fox et al. (2017); Kipf et al. (2019); Zhang et al. (2020); Pertsch et al. (2020); Chane-Sane et al. (2021); Kreidieh et al. (2020); Singh et al. (2021), behavior priors Salter et al. (2022), and action primitives Dalal et al. (2021); Nasiriany et al. (2022) further boost learning. However, the reference model updates in Hierarchical Reinforcement Learning and not well studied for further increase the reasoning ability of large language models after R1-Like reinforcement learning and it's curriculum learning variants. Therefore, we start to explore the reference model updates after the phase of R1-Like reinforcement learning and its curriculum learning variants.

3 Method

Although R1-Like reinforcement learning and its curriculum learning variants offer significant benefits in enhancing the reasoning ability of large language models, we designate this process as Phase 1. Following this, we propose Phase 2: rule-based reference model updates in reinforcement learning after Phase 1 to explore the potential of reference model updates following R1-Like reinforcement learning Figure 2. The details of the two phases are outlined below.

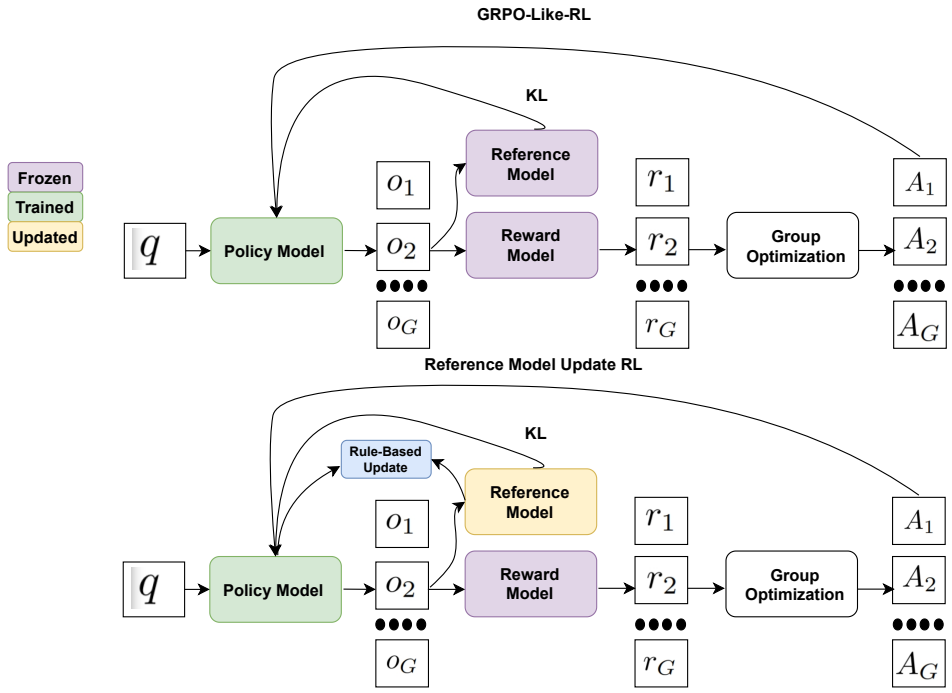

Figure 2: Description of Phase 1: GRPO Reinforcement learning and it's Variants with Frozen Reference Model. Phase 2: Reinforcement learning with Updated Reference Model in GRPO.

3.1 Phase 1(Preliminary): LLM Reasoning Via GRPO Plus(GRPO+) and VariantsYu et al. (2025)

3.1.1 Group Relative Policy OptimizationShao et al. (2024)

Compared to Proximal Policy Optimization (PPO) Schulman et al. (2017), Group-Relative Policy Optimization (GRPO) Shao et al. (2024) eliminates the value function and estimates the advantage in a group-relative manner.

For a specific question-answer pair (q, a), the behavior policy $\pi_{\theta_{\text{old}}}$ samples a group of G individual responses $\{o_i\}_{i=1}^{G}$. Then, the advantage of the i-th response is calculated by normalizing the group-level rewards $\{R_i\}_{i=1}^{G}$ as follows:

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(\{R_i\}_{i=1}^{G})}{\text{std}(\{R_i\}_{i=1}^{G})}. \tag{1}$$

Similar to PPO, GRPO adopts a clipped objective, together with a directly imposed KL penalty term:

$$\begin{aligned} &\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta^{\text{old}}}(\cdot|q)} \\ &\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\min\left(r_{i,t}(\theta)\hat{A}_{i,t},\right.\right.\right. \\ &\left.\left.\left.\text{clip}\Big(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\Big)\hat{A}_{i,t}\right) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\right)\right] \end{aligned} \tag{2}$$

where

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi\_\theta_{\text{old}}(o_{i,t} \mid q, o_{i,<t})}. \tag{3}$$

It is also worth noting that GRPO Shao et al. (2024) computes the objective at the sample-level. To be exact, GRPO first calculates the mean loss within each generated sequence, before averaging the loss of different samples. Such difference may have an impact on the performance of the algorithm. where μ_R and σ_R are the mean and standard deviation of the rewards in the group:

3.1.2 GROUP RELATIVE POLICY OPTIMIZATION PLUS(GRPO+)

The advanced Group Relative Policy Optimization algorithm Yu et al. (2025) is then developed. It samples a group of outputs $\{o_i\}_{i=1}^{G}$ for each question q paired with the answer a, and optimizes the policy via the following objective:

$$\begin{aligned}
&\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim D_q,\, \{o_i\}_{i=1}^{G}\sim\pi_\theta(\cdot|q)} \\
&\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{j=1}^{|o_i|}\min\left(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}A_{i,j},\right.\right. \\
&\left.\left.\text{clip}\left(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)},\, 1-\varepsilon_{\text{low}},\, 1+\varepsilon_{\text{high}}\right)A_{i,j}\right)\right]
\end{aligned} \tag{4}$$

where

$$A_{i,j} = \frac{r_i - \text{mean}(\{r_i\}_{i=1}^{G})}{\text{std}(\{r_i\}_{i=1}^{G})} \tag{5}$$

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi\_\theta_{\text{old}}(o_{i,t} \mid q, o_{i,<t})}. \tag{6}$$

Then, the key enhancements are represented as the following:

3.1.3 ENHANCEMENTS

Four kinds of enhancements to reasoning-focused reinforcement learning are applied. The KL penalty Kullback & Leibler (1951) is removed due to large divergence in long-CoT training. To improve exploration, *Clip-Higher* relaxes PPO's upper clipping bound Schulman et al. (2017). Dynamic sampling filters extreme-reward samples to maintain gradient diversity. Lastly, a token-level policy gradient KimiTeam et al. (2025) emphasizes longer responses for finer reward attribution.

3.1.4 CURRICULUM GRPO-LIKE REINFORCEMENT LEARNING.

Then, curriculum reinforcement learning is applied as an post-training strategy to further enhance the reasoning capabilities of large language models. The central idea is to structure the training process in a progressive manner, where the model is exposed to reasoning problems of gradually increasing difficulty rather than being overwhelmed by the full complexity of the task at once. At the early stages of training, the model begins with relatively simple reasoning instances that require shorter chains of thought, enabling it to establish a solid foundation of problem-solving skills and response structures. As training progresses, the complexity of the problems is systematically increased in a step-by-step curriculum, encouraging the model to develop more sophisticated reasoning strategies and adapt to diverse scenarios. This incremental exposure not only improves the robustness of reasoning but also optimizes the efficiency of the model's thinking process. By controlling both the difficulty of the reasoning tasks and the expected length of the responses within each sub-curriculum stage, the method guides the model to generate more concise and effective reasoning traces. As a result, the model learns not only to solve increasingly challenging problems but also to refine the quality, brevity, and accuracy of its outputs, ultimately leading to stronger reasoning performance with improved generalization across tasks.

3.2 PHASE 2: RULE-BASED REFERENCE MODEL UPDATES IN GROUP RELATIVE POLICY OPTIMIZATION(GRPO)

3.2.1 BATCH-SIZE-LEVEL TRAINING DATASET COLLECTION

In this phase of reinforcement learning, training dataset collection focuses on gathering data samples that can further strengthen the reasoning abilities of the policy model after the initial optimization stage, which includes R1-like reinforcement learning and curriculum-based variants. Unlike the earlier phase, where the goal is to build foundational reasoning skills and gradually increase task difficulty, this stage emphasizes the large-scale construction of high-quality reasoning trajectories to drive the next round of policy updates. The dataset is collected in batches, with each batch following the collection rules represented as the following:

$$D = threshold_2 < f_D(pass1@avg(rollout_n)) < threshold_1 \tag{7}$$

In detail, at each training step, we collect only a subset of each batch from the training dataset. For each sample, if the score of $rollout_n$ lies within the range $(threshold_2, threshold_1)$, the sample is retained in the batch. Samples with scores outside this range are discarded from the dataset. In the implementation, $threshold_2 = 0.2$ and $threshold_2 = 0.8$.

3.2.2 UPDATES IN STEP-WISE LEVEL

After the training dataset is collected at the batch-size level, we update the reference model in the GRPO training at the step-wise level. A rule-based update strategy is applied, in which the weights of the frozen reference model are replaced with the weights of the current actor model at specific training steps in Phase 2. In detail, at each training step i, we define an *updating score*, denoted as

$$\Delta\big(pass1@avg(rollout_n^i)\big), \tag{8}$$

which is calculated as the difference between $pass1@avg(rollout_n^i)$ and $pass1@avg(rollout_n^{i-1})$. Here, $pass1@avg(rollout_n^i)$ represents the $pass1@avg$ score of $rollout_n$ at training step i, with n fixed at 8. The update timestep $update_{i^*}$ is then chosen as the step i^* where the updating score achieves its maximum value within the interval from step 0 to step t_i, where t_i ¡ t_{total} , the total number of training steps in one epoch of Phase 2 reinforcement learning.

$$\begin{aligned} Update_{i^*} &= \max_{0 \le i \le t_i} \Delta\big(pass1@avg(rollout_n^i)\big) \\ \Delta\big(pass1@avg(rollout_n^i)\big) &= pass1@avg(rollout_n^i) - pass1@avg(rollout_n^{i-1}) \end{aligned} \tag{9}$$

In the implementation, to further enhance the stability of Phase 2 during reference model updating, we introduce an additional evaluation step before committing to any update. Specifically, at the timestep $Update_{i^*}$, the actor model is evaluated on the AIME24 benchmark to obtain its performance score under the metric $pass1@avg16$. This additional evaluation serves as a safeguard to ensure that only meaningful improvements are transferred to the reference model. If, and only if, the actor model achieves a higher $pass1@avg16$ score on AIME24 compared to the current reference model, then the update is applied by replacing the weights of the reference model with those of the actor model. Otherwise, the reference model remains unchanged, thereby preventing unnecessary or unstable updates that could harm the learning process. By incorporating this rule-based evaluation mechanism, the system ensures that reference model updates reflect genuine progress in reasoning performance rather than transient fluctuations, ultimately leading to a more stable and reliable training process in Phase 2 reinforcement learning.

4 EXPERIMENT

To investigate the effectiveness of the proposed rule-based reference updating GRPO in Phase 2 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-1.5B Guo et al.

Table 1: Model Performance Comparison

Model	MATH500	AIME24	AIME25	AMC	Minerva	Olympia
Close-Source						
O1-Preview	85.5	44.6	–	–	–	–
O1-Mini	90.0	70.0	–	–	–	–
O1	90.4	71.5	–	–	–	–
Claude Sonnet	82.2	23.3	–	–	–	–
Open-Source-Large						
DeepSeek-R1	97.3	79.8	–	–	–	
Qwen3-235B	94.6	85.7	–	–	–	
Llama 4 Behemoth	95.0	78.0	–	–	–	
Kimi-1.5	96.2	77.5	–	–	–	
Qwen 2.5-72B	83.1	30.0	–	–	–	
Phi4-Reasoning-14B	–	81.3	–	–	–	
Llama 4 Maverick	18.0	64.0	–	–	–	
Open-Source-4B/7B						
MIMO-7B	95.8	68.2	–	–	–	–
DeekSeek-Qwen-Distill-7B	92.8	55.5	–	–	–	–
Qwen3-4B	-	73.8	–	–	–	–
Open-Source-1.5B						
DeepSeek-R1-QWEN-1.5B	82.8	28.8	–	62.9	26.5	43.3
STILL-3-1.5B-Preview	84.4	32.5	–	66.7	29.0	45.4
FastCuRL-1.5B-Preview	88.0	43.1	–	74.2	31.6	50.4
FastCuRL-1.5B-V2	89.3	47.5	–	77.0	32.8	53.3
Diff-Aware-1.5B-Preview	89.2	50.0	33.0	77.1	35.3	51.9
FastCuRL-1.5B-V2	89.3	47.5	30.0	77.0	34.7	54.5
FastCuRL-1.5B-V3	90.5	49.6	32.9	78.5	34.7	53.3
FastCuRL-1.5B-V3+(Ours)	91.5	53.6	39.2	79.6	35.2	57.9
OpenNemotron-1.5B	85.9	54.0	41.7	75.8	26.3	-56.2
OpenNemotron-1.5B+(Ours)	88.0	60.2	48.2	78.2	29.2	59.1

(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-7B Guo et al. (2025), MIMO-7B Xiaomi LLM-Core Team (2025) with middle-sized parameters, Llama 4 Maverick AI (2025c), Phi4-Reasoning-14B Abdin et al. (2025), Qwen 2.5-72B Team (2024), Kimi-1.5 Team (2025a), Llama 4 Behemoth AI (2025b), 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 methods.

4.1 Experiment Setup and Evaluation

Setup We use DeepSeek-R1-Distill-Qwen-1.5B Guo et al. (2025) as our base model, a 1.5B parameter distilled model. Training is conducted using the AdamW optimizer Loshchilov & Hutter (2019) with a constant learning rate of 1×10^{-6}. For roll-outs, we set the temperature to 0.6 and sample 16 responses per prompt, appending "Let's think step by step and output the final answer within `\boxed{}`." to each problem without a system prompt.

Benchmarks and Dataset We evaluate our model across five math reasoning benchmarks: MATH500 Hendrycks et al. (2021), AIME2024 AI-MO (2024a), AMC2023 AI-MO (2024b), Minerva Lewkowycz et al. (2022), and OlympiadBench He et al. (2024b). For fair comparison, the

Table 2: Combined Model Rankings with Updated MATH500—AIME24—AIME25 Scores

MATH-500		**AIME24**		**AIME25**	
Model	Acc.	Model	Acc.	Model	Acc.
Gemini 2.5 Pro Exp	95.2%	O3-Pro	93.0%	o3-Mini	86.5%
O3	94.6%	Gemini 2.5 Pro Exp	92.0%	Gemini2.5-Pro-Exp	85.8%
Qwen 3 (235B)	94.6%	DeepSeek-R1-0528	91.4%	o3	85.3%
Grok 3 Mini Fast High Reasoning	94.2%	O3 Mini	86.5%	Grok3-Mini-Fast	85.0%
O4 Mini	94.2%	Gemini 2.5 Pro Exp	85.8%	Qwen3-(235B)	84.0%
DeepSeek R1	92.2%	O3	85.3%	o4-Mini	83.7%
O3 Mini	91.8%	Grok 3 Mini Fast High Reasoning	85.0%	DeepSeek-R1	74.0%
Gemini 2.5 Flash Preview (Thinking)	91.8%	Qwen 3 (235B)	84.0%	o1	71.5%
Claude 3.7 Sonnet (Thinking)	91.6%	O4 Mini	83.7%	Grok3-Mini-Low-Reasoning	70.6%
Gemini 2.5 Flash Preview	91.6%	Qwen-3-30B-A3B	65.8%	Phi-4 Reasoning	65.8%
O1	90.4%	O1	71.5%	GPT-4.1	66.3%
Ours-1.5B	91.5%	Grok 3 Mini Fast Low Reasoning	70.6%	DeepSeek R1	74.0%
Grok 3 Beta	89.8%	**Ours-1.5B**	60.2%	Grok3	58.7%
DeepSeek V3(03/24/2025)	88.6%	Grok 3 Beta	58.7%	**Ours-1.5B**	48.2%
Gemini 2.0 Flash(001)	88.0%	DeepSeek V3	52.2%	DeepSeek-R1-Distill-70B	46.7%
GPT4.1 Mini	88.0%	GPT 4.1 mini	49.4%	Gemini 2.0 Flash (001)	29.8%
GPT4.1	87.2%	Claude 3.7 Sonnet (Thinking)	44.6%	Claude 3.7 Sonnet (Non-thinking)	22.5%
Mistreal Medium 3	87.0%	Mistreal Medium 3	42.3%	Gemini 1.5 Pro (002)	18.7%
LLama4 Maveric	85.2%	GPT4.1	39.8%	Gemini 1.5 Flash (002)	17.3%
Gemini 2.0 Flash Think Exp	84.6%	Gemini 2.0 Flash (001)	29.8%		
Gemini 1.5 Pro (002)	82.8%	DeepSeek V3	27.5%		
DeepSeek V3	80.4%	GPT4.1 nano	27.3%		

Table 3: Ablation Studies

Model	**MATH500**	**AIME24**	**AIME25**	**AMC**	**Minerva**	**Olympia**
Close-Source						
Base1-FastCuRL-1.5B-V3	89.3	47.5	30.0	77.0	34.7	54.5
W/O Data Collection	89.3	47.5	30.0	77.0	34.7	54.5
W/O Reference Update	89.0	47.2	30.0	76.0	34.2	54.9
Ours	91.5	53.6	39.2	79.6	35.2	57.9
Base2-ReasoningNemotron-1.5B	85.9	54.0	41.7	75.8	26.3	56.2
W/O Data Collection	86.1	54.2	41.9	76.1	27.1	56.3
W/O Reference Update	85.8	54.1	41.9	76.0	26.9	56.8
Ours	88.0	60.2	48.2	78.2	29.2	59.1

training dataset is the same with the baseline models Wasi Uddin Ahmad (2025); Song et al. (2025). It includes problems of varied difficulty, comprising AIME of America (2024), AMC of America (2025), Omni-MATH Gao et al. (2024a), and Still RUC-AIBOX (2025). Similarly, the parameter setting such as the temperature, the top_p, the top_k, the max length of thinking tokens is settingas the same with the baseline models Wasi Uddin Ahmad (2025); Song et al. (2025).

Evaluation Metric We adopt PASS@1 as the evaluation metric. Using a temperature of 0.6 and top-$p = 1.0$, we generate $k = 16$ responses per question. PASS@1 is then computed as: PASS@1 $= \frac{1}{k}\sum_{i=1}^{k} p_i$.

4.2 Math Reasoning Experiments

Math Benchmarks The proposed rule-based reference model updates in phase 2 is evaluated against top open- and closed-source models, including Gemini-2.5-Pro DeepMind (2025a), O3-Mini OpenAI (2025a), Grok-3-Mini (High) xAI (2025a), Qwen3-235B-A22B Team (2025c), and others. As shown in Table 3, our 1.5B model achieves strong performance across benchmarks: 56.3 Pass@1 on AIME24 Jia (2025), 90.5 on MATH500 HuggingFaceH4 (2025), 78.6 on AMC23 of America (2023), 34.7 on Minerva Dyer & Gur-Ari (2022), and 55.5 on OlympiadBench He et al. (2024a),

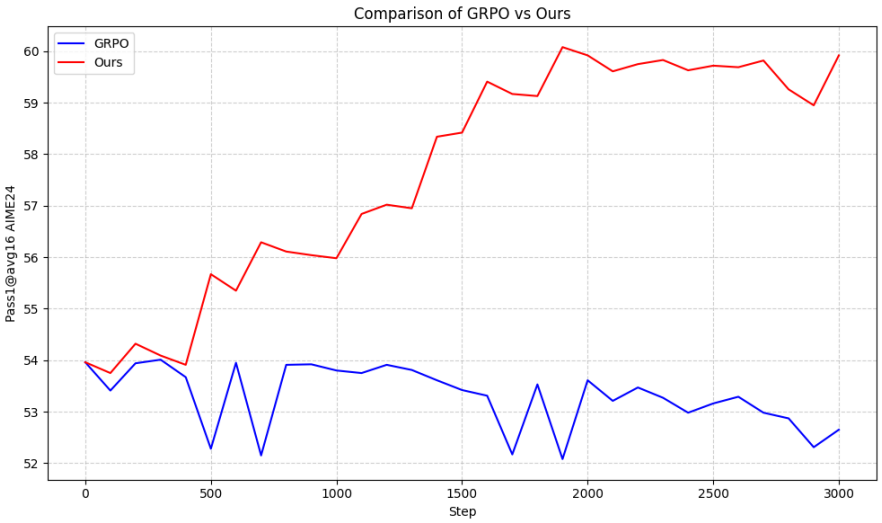

Figure 3: Comparison of Reference Frozen(GRPO) and Proposed Reference Updates(Oues) showing Pass1@avg16 AIME24 values over training steps.

demonstrating robust reasoning ability across diverse math tasks. For fair comparison, the length, the temperature and other hyper-parameters are set the same with the baseline1(FastCuRL-1.5B-V3 Song et al. (2025)) and baseline2(ReasoningNemotron-1.5B Wasi Uddin Ahmad (2025)) in the testing.

We evaluate our reasoning model against leading open- and closed-source models, including Gemini-2.5-Pro DeepMind (2025a), O3-Mini OpenAI (2025a), Grok-3-Mini (High) xAI (2025a), and Qwen3-235B Team (2025c). As shown in Table 3, our 1.5B model achieves strong results across diverse benchmarks: 60.2 on AIME24 Jia (2025), 91.5 on MATH500 HuggingFaceH4 (2025), 79.6 on AMC23 of America (2023), 39.2 on Minerva Dyer & Gur-Ari (2022), and 59.1 on OlympiadBench He et al. (2024a), demonstrating robust general mathematical reasoning.

For the Base1-FastCuRL-1.5B-V3 baseline, our approach achieves increases across all evaluation benchmarks. Notably, on MATH500, our model improves accuracy from 90.5% to 91.5%, which corresponds to an approximate 1.00% relative gain. On AIME24, the improvement is more significant, rising from 49.6% to 53.6%, a 4.00% increase. The gain on AIME25 is also notable, moving from 32.9% to 39.2%, which is a 7.00% relative improvement. Additional improvements can be seen on AMC (1.10%), Minerva (a slight decrease of 0.50%), and Olympia (4.50%). Compared with Base2-OpenReasoning-1.5B, our model shows even stronger gains. On MATH500, accuracy improves from 85.9% to 88.0%, representing a 3.10% improvement. On AIME24, the increase is from 54.0% to 60.2%, a 5.80% gain. Similarly, on AIME25, our model boosts performance from 41.7% to 48.2%, which corresponds to a 6.50% relative improvement. Further improvements are observed on AMC (2.40%) and Minerva (2.90%).

Three Normal Math Leaderboards On a range of competitive benchmarks, our model demonstrates outstanding performance by ranking the 11th place on both the Math500 , 14th AIME24 leaderboards and 15th AIME25 leaderboards. In particular, on the Math500 HuggingFaceH4 (2025) benchmark, Ours-1.5B delivers exceptional results by outperforming several prominent models in the field. These include Grok 3 Beta xAI (2025b)(89.8%), DeepSeek V3 (03/24/2025) AI (2025a)(88.6%), Gemini 2.0 Flash (001) DeepMind (2025b)(88.0%). Similarly, on the AIME24 Jia (2025) benchmark, which evaluates problem-solving capabilities in a highly challenging math competition setting, Ours-1.5B again achieves superior results. It surpasses DeepSeek V3 (03/24/2025) AI (2025a) (52.2%), GPT-4.1 Mini OpenAI (2025b) (49.4%). Besides, on the AIME25 Jia (2025) benchmark, which evaluates problem-solving capabilities in a highly challenging math competition setting, Ours-1.5B again achieves superior results. It's score is almost the same with DeepSeek-

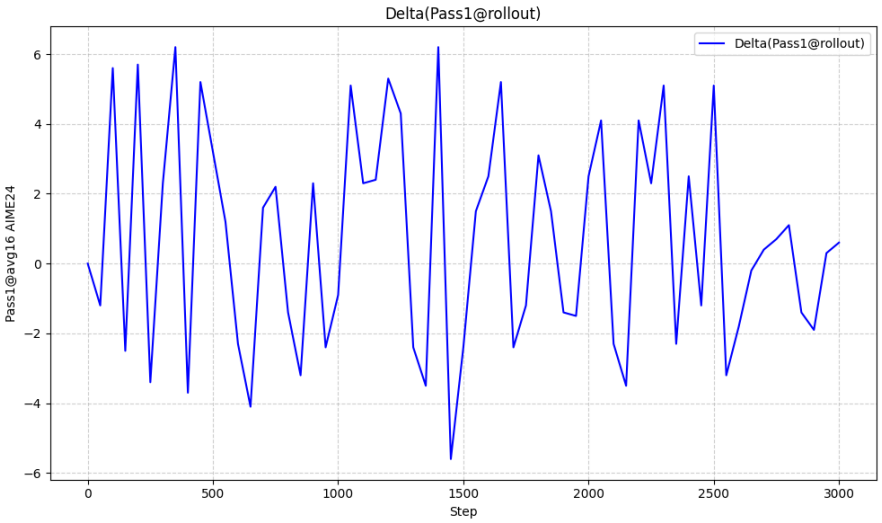

Figure 4: Details of showing $\Delta Pass1@avgrollout_{16}$ values over training steps for the phase 2.

R1-Distll-70B AI (2025a) and surpasses Gemini 2.0 Flash (001) AI (2025a) (29.8.2%), Claude 3.7 Sonnet (Non-thinking) 22.5% xAI (2025a).

Ablation Studies The ablation study in Table 3 investigates the contributions of individual components within two base models—Base1-FastCuRL-1.5B-V3 and Base2-Nemotron-Research-Reasoning-Qwen-1.5B across six benchmark datasets. For Base1, removing either Batch-Wise Data Collection or Step-Wise Reference Model Updation led to stagnation or a slight degradation in performance across tasks such as AIME24, AIME25, and Olympia, suggesting their necessity for reasoning enhancement. Notably, the full model outperforms all ablated variants, achieving 53.6% on AIME24 and 39.2% on AIME25. Similarly, in Base2, removing Batch-Wise Data Collection or Step-Wise Reference Model Collection leads to minor drops in performance.

Key Visualization of Training Process Visual details of the training process are shown in Figure3 and Figure4 for better comparison of the frozen reference model and updated reference model in the phase 2 and the values of updating score(in percentage%) in the phase 2. In details, in the phase 2 the pass1@avg16 score of frozen reference model(bule line) and updated reference model(red line) AIME24 is represented in Figure3. The updating score of the updated reference model is represented in Figure4.

5 Discussion

We explore reinforcement learning of reference model updates to enhance the reasoning capabilities of large language models (LLMs) in mathematics. In Phase 2, our framework introduces rule-based updates to improve step-by-step problem-solving. We evaluate this approach on a small-scale math dataset using a 1.5B-parameter model.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, Vibhav Vineet, Yue Wu, Safoora Yousefi, and Guoqing Zheng. Phi-4-reasoning technical report. https://arxiv.org/abs/2504.21318, 2025. arXiv preprint arXiv:2504.21318.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 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.

DeepSeek AI. Deepseek v3 (03/24/2025): Enhanced performance and capabilities, March 2025a. URL https://api-docs.deepseek.com/news/news250325. Accessed: 2025-05-20.

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

Meta AI. Llama 4 maverick: A natively multimodal mixture-of-experts model. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025c. 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.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, 2024. abs/2310.12036.

Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Elliot Chane-Sane, Cordelia Schmid, and Ivan Laptev. Goal-conditioned reinforcement learning with imagined subgoals. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. TBD, 2021.

Yifan Chen, Nick Yang, Zhihao Luo, Jiayi He, Ming Zhang, and Lei Wang. 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.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. 2017.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

Murtaza Dalal, Deepak Pathak, and Ruslan Salakhutdinov. Accelerating robotic reinforcement learning via parameterized action primitives. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://arxiv.org/abs/2110.15360.

Peter Dayan and Geoffrey E. Hinton. Feudal reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 271–278. Morgan Kaufmann Publishers, Inc., 1993a.

Peter Dayan and Geoffrey E. Hinton. Feudal reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 271–278, 1993b. doi: 10.1108/IR-08-2017-0143. URL http://www.cs.toronto.edu/~fritz/absps/dh93.pdf.

Google DeepMind. Gemini 2.5 pro: Our most intelligent ai model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/, March 2025a. Accessed: 2025-05-20.

Google DeepMind. Gemini 2.0 flash (001): Next-generation multimodal ai model. https://ai.google.dev/gemini-api/docs/models/gemini, February 2025b. Accessed: 2025-05-20.

Ethan Dyer and Guy Gur-Ari. Minerva: Solving quantitative reasoning problems with language models. *Google Research Blog*, 2022. URL https://research.google.com/blog/2022/06/minerva-solving-quantitative-reasoning-problems-with-language-models/. Accessed: 2025-05-20.

Tom Everitt, Victoria Krakovna, Laurent Orseau, Marcus Hutter, and Shane Legg. Reinforcement learning with a corrupted reward channel. *arXiv preprint arXiv:1711.00391*, 2017.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *arXiv preprint arXiv:2105.00901*, 2021.

Roy Fox, Sanjay Krishnan, Ion Stoica, and Ken Goldberg. Multi-level discovery of deep options. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1665–1674. PMLR, 2017. URL https://proceedings.mlr.press/v70/fox17a.html.

Yichao Fu, Junda Chen, Siqi Zhu, Zheyu Fu, Zhongdongming Dai, Aurick Qiao, and Hao Zhang. Efficiently serving llm reasoning programs with certaindex. *arXiv preprint arXiv:2412.20993*, 2024.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

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. Omni-math: A universal olympiad level mathematic benchmark for large language models, 2024a. URL https://arxiv.org/abs/2410.07985.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, et al. A framework for few-shot language model evaluation. https://zenodo.org/records/12608602, 2024b.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*, 2025.

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

Dong Guo, Dongrui Yang, Hao Zhang, Jinjun Song, Rui Zhang, Rui Xu, Qizhe Zhu, Sheng Ma, Peng Wang, Xia Bi, 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.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Anna Harutyunyan, Peter Vrancx, Pierre-Luc Bacon, Doina Precup, and Ann Nowé. Learning with options that terminate off-policy. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3201–3208, 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11740.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 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/.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 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.

Dan Hendrycks, Collin Burns, Steven Basart, Antonia Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 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.

Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Andrew Sendonaris, Gabriel Dulac-Arnold, Ian Osband, John P. Agapiou, Joel Z. Leibo, and Audrunas Gruslys. Deep q-learning from demonstrations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3223–3230, 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11794.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. V-star: Training verifiers for self-taught reasoners. *arXiv preprint arXiv:2402.06457*, 2024.

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

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, and Alex Carney. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

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

Leslie Pack Kaelbling. Learning to achieve goals. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1094–1098, 1993.

KimiTeam, 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.

Thomas Kipf, Yujia Li, Hanjun Dai, Vinicius Zambaldi, Alvaro Sanchez-Gonzalez, Edward Grefenstette, Pushmeet Kohli, and Peter Battaglia. Compile: compositional imitation learning and execution. In *International Conference on Machine Learning*, pp. 3418–3428. PMLR, 2019.

Martin Klissarov, Pierre-Luc Bacon, Jean Harb, and Doina Precup. Learning options end-to-end for continuous action tasks. In *NeurIPS Hierarchical Reinforcement Learning Workshop*, 2017. URL https://arxiv.org/abs/1712.00004.

Abdul Rahman Kreidieh, Samyak Parajuli, Nathan Lichtlé, Yiling You, Rayyan Nasr, and Alexandre M. Bayen. Inter-level cooperation in hierarchical reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 2000–2002, 2020. URL https://dl.acm.org/doi/10.5555/3398761.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951. doi: 10.1214/aoms/1177729694. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-22/issue-1/On-Information-and-Sufficiency/10.1214/aoms/1177729694.full.

Abhinav Kumar, Jaechul Roh, Ali Naseh, Marzena Karpinska, Mohit Iyyer, Amir Houmansadr, and Eugene Bagdasarian. Overthink: Slowdown attacks on reasoning llms. *arXiv e-prints*, pp. arXiv:2502, 2025.

Andrew Levy, Robert Platt, and Kate Saenko. Hierarchical actor-critic. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=SJ3rcZ0cK7.

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. 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.

Google LLC. 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.

Ilya Loshchilov and Frank Hutter. 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.

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

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. S1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Ofir Nachum, Honglak Lee, Shane Gu, and Sergey Levine. Data-efficient hierarchical reinforcement learning. *Advances in Neural Information Processing Systems*, 31:3738–3748, 2018.

Ashvin Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. In *Advances in Neural Information Processing Systems*, pp. 1–13, 2018.

Soroush Nasiriany, Huihan Liu, and Yuke Zhu. Augmenting reinforcement learning with behavior primitives for diverse manipulation tasks. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7477–7484, 2022. URL https://arxiv.org/abs/2110.03655.

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

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

Mathematical Association of America. 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 o3-mini: A cost-efficient reasoning model. https://openai.com/index/openai-o3-mini/, January 2025a. Accessed: 2025-05-20.

OpenAI. Gpt-4.1 mini: A high-performance language model, April 2025b. URL https://openai.com/index/gpt-4-1/. Accessed: 2025-05-20.

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

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Karl Pertsch, Youngwoon Lee, and Joseph J. Lim. Accelerating reinforcement learning with learned skill priors. In *Conference on Robot Learning (CoRL)*, pp. 944–957, 2020. URL https://arxiv.org/abs/2010.11944.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In *Proceedings of Robotics: Science and Systems (RSS)*. Robotics: Science and Systems Foundation, 2018.

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.

Sasha Salter, Markus Wulfmeier, Dhruva Tirumala, et al. Mo2: Model-based offline options. In *Conference on Lifelong Learning Agents*, pp. 902–919, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Qizhe Shao, Peng Wang, Ronghang Zhu, Rui Xu, Jian Song, Xia Bi, Hao Zhang, Meng Zhang, Yichong Li, and et al. Wu, Yiming. 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.

Konstantinos Shiarlis, Markus Wulfmeier, Shaun Salter, Shimon Whiteson, and Ingmar Posner. Taco: Learning task decomposition via temporal alignment for control. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 4654–4663. PMLR, 2018.

Avi Singh, Huihan Liu, Gaoyue Zhou, Tianhe Yu, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Parrot: Data-driven behavioral priors for reinforcement learning. In *9th International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2011.10024.

Mingyang Song, Mao Zheng, Zheng Li, Wenjie Yang, Xuan Luo, Yue Pan, and Feng Zhang. Fastcurl: Curriculum reinforcement learning with stage-wise context scaling for efficient training r1-like reasoning models, 2025. URL https://arxiv.org/abs/2503.17287.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

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

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

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

Qwen Team. Qwen3-235b-a22b: A 235b-parameter mixture-of-experts language model. https://huggingface.co/Qwen/Qwen3-235B-A22B, April 2025c. Accessed: 2025-05-20.

Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. In *International Conference on Machine Learning*, pp. 3540–3549. PMLR, 2017.

P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, and Z. Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the ACL (Long Papers)*, pp. 9426–9439, 2024.

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

Somshubra Majumdar Aleksander Ficek Siddhartha Jain Jocelyn Huang Vahid Noroozi Boris Ginsburg Wasi Uddin Ahmad, Sean Narenthiran. OpenCodeReasoning: Advancing Data Distillation for Competitive Coding. 2025. URL https://arxiv.org/abs/2504.01943.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, 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.

Lilian Weng. 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.

xAI. Grok 3 mini (high reasoning). https://x.ai/blog/grok-3, April 2025a. Accessed: 2025-05-20.

xAI. Grok 3 beta: The age of reasoning agents, February 2025b. URL https://x.ai/news/grok-3. Accessed: 2025-05-20.

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

A. Yang, B. Zhang, B. Hui, B. Gao, B. Yu, C. Li, D. Liu, J. Tu, J. Zhou, J. Lin, and et al. Qwen 2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Qingcai Yu, Zewei Zhang, Ronghang Zhu, Yilun Yuan, Xiaofu Zuo, Yalong Yue, Tian Fan, Guoliang Liu, Liang Liu, Xiaolin Liu, and et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou, Hao Peng, Zhiyuan Liu, and Maosong Sun. Advancing llm reasoning generalists with preference trees. *arXiv preprint*, 2024a.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. *arXiv preprint*, 2024b. URL https://arxiv.org/abs/2412.01981.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Hu Jinfang, Zhiyuan Liu, and Bowen Zhou. Ultramedical: Building specialized generalists in biomedicine. *arXiv preprint*, 2024.

Tianren Zhang, Shangqi Guo, Tian Tan, Xiaolin Hu, and Feng Chen. Generating adjacency-constrained subgoals in hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9814–9826, 2020. URL https://proceedings.neurips.cc/paper/2020/file/f5f3b8d720f34ebebceb7765e447268b-Paper.pdf.

A Appendix

You may include other additional sections here.