EFFICIENT REINFORCEMENT FINETUNING VIA ADAPTIVE CURRICULUM LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement finetuning (RFT) has shown great potential for enhancing the mathematical reasoning capabilities of large language models (LLMs), but it is often sample- and compute-inefficient, requiring extensive training. In this work, we introduce ADARFT (*Adaptive Curriculum Reinforcement Finetuning*), a method that significantly improves both the efficiency and final accuracy of RFT through adaptive curriculum learning. ADARFT dynamically adjusts the difficulty of training problems based on the model's recent reward signals, ensuring that the model consistently trains on tasks that are challenging but solvable. This adaptive sampling strategy accelerates learning by maintaining an optimal difficulty range, avoiding wasted computation on problems that are too easy or too hard. ADARFT requires only a lightweight extension to standard RFT algorithms like Proximal Policy Optimization (PPO), without modifying the reward function or model architecture. Experiments on competition-level math datasets—including AMC, AIME, and IMO-style problems—demonstrate that ADARFT significantly improves both training efficiency and reasoning performance. We evaluate ADARFT across multiple data distributions and model sizes, showing that it reduces training time by up to $2\times$ and improves accuracy by a considerable margin, offering a more scalable and effective RFT framework.

1 INTRODUCTION

Reinforcement Finetuning (RFT) has emerged as a powerful technique for aligning large language models (LLMs) with task-specific goals, particularly in domains such as mathematics and code generation where correctness is well defined (DeepSeek-AI et al., 2025; OpenAI et al., 2024b). By optimizing a policy model with reward signals that reflect task success, RFT enables more targeted learning than supervised finetuning (SFT) alone. However, despite its promise, RFT remains sample-inefficient and computationally expensive. Its training involves repeated rollout generation, reward computation, and policy updates—making it costly and difficult to scale (Ahmadian et al., 2024; Kazemnejad et al., 2024; Li et al., 2024; Hu, 2025; Cui et al., 2025). Recent efforts to address RFT inefficiency have focused on algorithmic simplification (e.g., RAFT (Dong et al., 2023), GRPO (DeepSeek-AI et al., 2025), ReMax (Li et al., 2024)), and data-centric strategies (e.g., LIMO (Ye et al., 2025), LIMR (Li et al., 2025)). While these approaches improve sample or compute efficiency, they often introduce trade-offs: algorithmic simplifications may increase variance or limit stability, and static data filtering or scoring can be brittle, computationally heavy, or model-specific. Moreover, most methods' success relies on fixed datasets or training schedules, which can be suboptimal in non-uniform or imbalanced data regimes. More recently, early efforts have introduced curriculum-like ideas into RFT. Staged curricula divide training into a few manually-defined phases of increasing difficulty (Wen et al., 2025; Luo et al., 2025; Song et al., 2025), but these are coarse-grained and lack adaptivity. Other methods use online data filtering, repeatedly rolling out and pruning training samples until the model's average reward meets a target threshold (Bae et al., 2025; Yu et al., 2025). While this approach helps prevent the model from stagnating on problems that are either too easy or too difficult, it is not truly adaptive and incurs significant rollout overhead.

To address these limitations, we propose ADARFT, a reinforcement finetuning method based on adaptive curriculum learning (Bengio et al., 2009), which dynamically adjusts training set difficulty to match the model's evolving skill level. The intuition is simple: learning is most effective when tasks are neither too easy nor too hard. ADARFT formalizes this by maintaining a target difficulty level,

which increases or decreases based on recent reward feedback. At each step, the model is trained on examples closest to this target, promoting a steady progression through solvable yet challenging tasks. The full algorithm is outlined in Algorithm 1. Unlike prior work that relies on fixed stages, repeated rollouts, or model-specific data processing, ADARFT is lightweight, general, and model-agnostic. It can be directly applied on top of any standard reinforcement learning (RL) algorithms like Proximal Policy Optimization (PPO) (Schulman et al., 2017b). We evaluate ADARFT on a dataset spanning a wide range of competition-level math problems, including AMC, AIME, and IMO-style questions. Across multiple training distributions and two model sizes, ADARFT significantly improves both training efficiency and final performance. Gains are especially notable in imbalanced data regimes, where static sampling often fails. ADARFT can reduce training time by up to $2\times$, offering a practical and scalable path to more efficient RFT in structured reasoning tasks.

2 Related Work

Efficient Reinforcement Finetuning. Most RFT pipelines build on Proximal Policy Optimization (PPO) (Schulman et al., 2017b), with recent variants like RAFT (Dong et al., 2023), ReMax (Li et al., 2024), GRPO (DeepSeek-AI et al., 2025), and REINFORCE++ (Hu, 2025), aiming to reduce computational overhead by simplifying RL components. While effective, these methods often trade off stability or sample efficiency. In parallel, data-centric strategies have emerged as promising alternatives for efficient finetuning. LIMO (Ye et al., 2025) and s1 (Muennighoff et al., 2025) show that small, carefully selected supervised datasets can yield strong downstream performance, but their success hinges on manual curation, prompt engineering, and careful dataset construction, which may not generalize across tasks or models. LIMR (Li et al., 2025) and Wang et al. (2025) proposes scoring training examples based on their estimated learning impact, enabling selective finetuning with fewer samples. Yet, computing these scores requires a full training run, and the scores must be recomputed for each new model, limiting practicality and scalability. Moreover, reducing the number of training samples does not inherently translate to improved efficiency. Models still require a comparable number of optimization steps and wall-clock time to converge. In contrast, ADARFT introduces a lightweight, model-agnostic curriculum learning strategy that dynamically adjusts task difficulty based on reward feedback. This allows continuous adaptation to the model's capabilities, improving convergence speed and final accuracy without modifying the RL algorithm or requiring manual data curation.

Curriculum Learning for RL. Curriculum learning (CL) structures training by presenting tasks in an organized progression, typically from easy to hard, to enhance learning efficiency and generalization (Bengio et al., 2009). In RL, CL methods include task sorting by difficulty (Zaremba & Sutskever, 2015; Justesen et al., 2018; Wang et al., 2019), teacher-student frameworks that adaptively select tasks based on learning progress (Matiisen et al., 2017; Portelas et al., 2019), and self-play approaches that induce automatic curricula through agent competition (Sukhbaatar et al., 2018; Zhao et al., 2025). Other strategies use intermediate-goal generation in sparse-reward settings (Florensa et al., 2018), unsupervised skill discovery (Jabri et al., 2019), or knowledge transfer via progressive networks and imitation (Czarnecki et al., 2018; Rusu et al., 2022). While CL is well-studied in classical RL, its application to RFT of LLMs is still limited. Existing methods typically use staged training with hand-designed difficulty tiers (Wen et al., 2025; Luo et al., 2025; Song et al., 2025), or online filtering schemes that repeatedly sample and discard data until rewards reach a target range (Bae et al., 2025; Yu et al., 2025). These methods either lack adaptability or introduce significant computational overhead due to repeated rollouts. In contrast, ADARFT is among the first truly adaptive curriculum learning approaches for RFT: it continuously adjusts task difficulty based on the model's reward signal, enabling efficient, scalable training without fixed schedules or repeated rollouts.

3 ADARFT

We aim to improve the performance of a policy model π_θ for solving mathematical problems through adaptive curriculum learning. Fine-tuning on problems that are too easy or too hard leads to poor learning outcomes. Instead, the model should be trained on problems whose difficulty is close to the model's current capability. We frame this as an adaptive curriculum learning problem and propose

ADARFT, which adaptively adjusts the target difficulty to keep training problems within a suitable difficulty range. ADARFT is compatible with a variety of RL algorithms (e.g, GRPO, PPO); in this work, we instantiate it with PPO and refer to this variant as ADARFT (PPO).

Let D be a dataset of mathematical problems, each annotated with a precomputed difficulty score d_i. The score can be either human-annotated or model-estimated. The objective is to train a policy π_θ that improves its problem-solving ability by dynamically adjusting the training curriculum according to the model's current performance. Our proposed algorithm, ADARFT, is shown in Algorithm 1.

Algorithm 1 ADARFT – Adaptive Curriculum Reinforcement Finetuning

1: **Input:** Data source D with difficulty scores $\{d_i\}$, policy model π_θ, reward function $R(\cdot,\cdot)$, batch size B, initial target difficulty T, step size η, sensitivity α, target reward β, difficulty bounds $d_{\min}, d_{\max}$
2: Select RL algorithm $\mathcal{A}$ (e.g., PPO, GRPO, REINFORCE++)
3: **while** training is not finished **do**
4: Compute absolute differences from target difficulty: $\Delta_i = |d_i - T| \quad \forall i \in \{1, \ldots, |D|\}$
5: Sort and select top B samples closest to target difficulty: $X \leftarrow \{s_1, s_2, \ldots, s_B\}$
6: Generate responses using policy model: $G = \pi_\theta(X)$
7: Compute average reward: $R_{avg} \leftarrow \frac{1}{|X|}\sum_{i=1}^{|X|} R(X_i, G_i)$
8: Update policy: $\pi_\theta \leftarrow \mathcal{A}(\pi_\theta, X, G, R)$
9: Update and clip target difficulty: $T' \leftarrow \text{clip}(T + \eta \cdot \tanh(\alpha \cdot (R_{avg} - \beta)), d_{\min}, d_{\max})$
10: Update sampler: $T \leftarrow T'$
11: **end while**

3.1 DYNAMIC CURRICULUM SAMPLING

To construct an adaptive curriculum, we define a target difficulty T, which represents the current target difficulty level for training (more in § 3.3). ADARFT dynamically adjusts T based on the model's reward signal to maintain an optimal difficulty level for learning. At each step, the algorithm computes the absolute difference between the target difficulty and the difficulty of each problem in the dataset (Alg. 1, line 4): $\Delta_i = |d_i - T|$ for all $i \in [1, |D|]$. The batch of training problems is formed by selecting the B problems with the smallest values of Δ_i (Alg. 1, line 5), producing a batch: $X = \{s_1, s_2, \ldots, s_B\}$. This ensures that the selected problems are closest to the model's current target difficulty, focusing the learning process on problems that are neither too easy nor too hard.

3.2 POLICY UPDATE

The selected batch X is used to train the policy model π_θ, which generates responses: $G = \pi_\theta(X)$. A reward signal is computed based on the correctness of the model's output (Alg. 1, line 7): $R_i = 1$ if the response is correct, and $R_i = 0$ if the response is incorrect. The average reward over the batch is computed as (Alg. 1, line 7): $R_{avg} = \frac{1}{|X|}\sum_{i=1}^{|X|} R(X_i, G_i)$. The policy can then be updated using a reinforcement learning algorithm $\mathcal{A}$ such as PPO, GRPO, or REINFORCE++ (Alg. 1, line 8): $\pi_\theta \leftarrow \mathcal{A}(\pi_\theta, X, G, R)$.

3.3 TARGET DIFFICULTY UPDATE

To adapt the curriculum dynamically, the target difficulty is updated based on the average reward. If the model performs well on the current difficulty level (high reward), the target difficulty increases, making the training problems harder. Conversely, if the model performs poorly, the target difficulty decreases. This dynamic update mechanism lies at the core of ADARFT's curriculum adaptation strategy. The update rule (Alg. 1, line 9) is defined as:

$$T' = \text{clip}(T + \eta \cdot \tanh(\alpha \cdot (R_{avg} - \beta)), d_{\min}, d_{\max})$$

Here, η, α, β are hyperparameters: η is the step size for adjusting the target difficulty, α controls the sensitivity of the update, and β is the target reward level, representing the desired success rate. The $\tanh$ function ensures smooth updates and prevents large jumps in difficulty by saturating for large deviations, while the "clip" function constrains the target difficulty within the valid range

$[d_{\min}, d_{\max}]$. These bounds can be manually specified or automatically derived from the training set, for example by taking the minimum and maximum of the difficulty scores $\{d_i\}$. Intuition and guidance for selecting these hyperparameters are discussed in Section 3.4 and 4.3.

3.4 Theoretical Justification for Target Reward β

A key component of AdaRFT is its adaptive curriculum mechanism, which steers training toward a target reward level β. Intuitively, we aim to train on examples that are neither trivially easy nor prohibitively hard. In this light, setting $\beta = 0.5$, corresponding to a success rate of roughly 50%, naturally aligns with this goal. This section formalizes that intuition by analyzing the relationship between reward variance and learnability in RFT with binary rewards.

In entropy-regularized reinforcement learning, the optimal policy π^* can be expressed relative to a reference policy π_{init} as (Korbak et al., 2022; Go et al., 2023; Rafailov et al., 2023):

$$\pi^*(y \mid x) = Z(x)\pi_{\text{init}}(y \mid x) \exp\left(\frac{1}{\tau} r(x, y)\right) \tag{1}$$

where τ is the inverse temperature parameter controlling entropy regularization, and $Z(x)$ is the partition function that normalizes the action probability. The corresponding optimal value function and the partition function is given by (Schulman et al., 2017a; Richemond et al., 2024):

$$V^*(x) := \tau \log \mathbb{E}_{y \sim \pi_{\text{init}}(\cdot|x)} \left[\exp\left(\frac{1}{\tau} r(x, y)\right)\right] \quad \text{and} \quad Z(x) = \exp\left(\frac{1}{\tau} V^*(x)\right) \tag{2}$$

We can then take the expectation of the log-ratio between the optimal policy and the initial policy with respect to $y \sim \pi_{\text{init}}(\cdot \mid x)$, leading to (Haarnoja et al., 2017; Schulman et al., 2017a):

$$\mathbb{E}_{y \sim \pi_{\text{init}}(\cdot|x)} \left[\log \frac{\pi^*(y \mid x)}{\pi_{\text{init}}(y \mid x)}\right] = \frac{1}{\tau} \mathbb{E}_{\pi_{\text{init}}}[r(x, y)] - \frac{1}{\tau} V^*(x) \tag{3}$$

Since the left-hand side can be interpreted as the negative reverse KL divergence between π_{init} and π^* (Rafailov et al., 2024), Bae et al. (2025) show that when the reward $r(x, y)$ with $y \sim \pi_{\text{init}}(\cdot \mid x)$ is Bernoulli, the KL divergence is lower-bounded by the reward variance:

$$D_{\text{KL}}(\pi_{\text{init}} \| \pi^*) \geq \frac{p(x)(1 - p(x))}{2\tau^2} \tag{4}$$

where $p(x)$ is the model's success rate on prompt x. This implies that the lower bound on the KL divergence, and consequently the gradient magnitude during policy updates, is proportional to the reward variance, which is maximized when $p(x) = 0.5$. In other words, training on prompts that the model succeeds on roughly half the time may yield the strongest learning signal. In Section 5 and Appendix 5.3, we conduct an ablation study by varying the target reward β, demonstrating that setting $\beta = 0.5$ consistently leads to the best performance, supporting the hypothesis that training on prompts with a success rate near 50% provides the most informative learning signal.

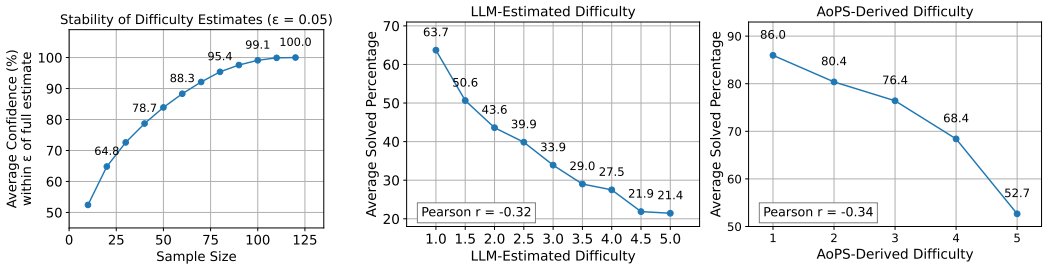
4 Experiments

4.1 Difficulty Estimation

Accurate estimation of problem difficulty is critical for AdaRFT. For difficulty estimation, we select the Qwen 2.5 MATH 7B model (Qwen et al., 2025) because it demonstrates a balanced solving ability. A model that is too strong (e.g., OpenAI o1 (OpenAI et al., 2024b), DeepSeek R1 (DeepSeek-AI et al., 2025)) would solve most problems on the first attempt, leading to poor discrimination between easy and hard problems. Conversely, a model that is too weak (e.g., LLaMA 3.3 1B (Grattafiori et al., 2024)) would fail to solve most problems even after multiple attempts, limiting the signal required for curriculum adaptation. For each problem, the difficulty score is computed as:

$$d_i = 100 \times \left(1 - \frac{\text{number of successful attempts on problem } i}{n}\right)$$

where n is the number of attempts per problem. In our setup, we use $n = 128$.



(a) Average confidence that subsampled difficulty estimates fall within ±0.05 of the full-sample estimate.

(b) Correlation between average solved percentage and two types of difficulty labels: (left) LLM-estimated difficulty and (right) AoPS-derived difficulty levels.

Figure 1: Evaluation of difficulty estimation: (a) Stability of difficulty scores under subsampling of model rollouts; (b) Correlation between labeled difficulty levels and average solved percentage.

To evaluate the stability of our difficulty estimation process, we simulate how confidence varies with different numbers of samples. For each problem, we treat the full set of 128 rollouts as the ground-truth difficulty estimate and compute how often sub-sampled estimates fall within a tolerance of $\epsilon = 0.05$. Specifically, we run 10 random sampling trials per sample size and average the confidence across all problems in the dataset. As shown in Figure 1a, even with as few as 64 samples, the estimated difficulty remains within ± 0.05 of the full estimate over 90% of the time. With just 40 samples, the confidence remains around 80%. These results indicate that accurate and robust difficulty estimation can be achieved with significantly fewer rollouts, reducing the computational burden of large-scale curriculum construction.

To further validate the reliability of our difficulty estimates, we examined their alignment with the difficulty levels provided in the MATH dataset. The MATH dataset comprises 12,500 competition-level mathematics problems sourced from contests such as the American Mathematics Competitions (AMC) and the American Invitational Mathematics Examination (AIME). Each problem is categorized into one of five difficulty levels, following the classification system used by the Art of Problem Solving (AoPS) community.[1] In this system, level 1 denotes the easiest problems, while level 5 represents the most difficult. As shown in Figure 1b, there is a clear downward trend in the average solve rate as the labeled difficulty level increases, ranging from 86.0% at level 1 to 52.7% at level 5. Specifically, the AoPS-derived difficulty levels yield a Pearson correlation of $r = -0.34$ ($p < 0.05$) with model success rates. This negative correlation indicates that the model's empirical performance aligns well with the intended difficulty stratification, reinforcing the utility of both the labeled difficulty levels and our estimation approach in guiding curriculum learning. To further streamline the difficulty estimation process, we also prompted GPT-4o (`gpt-4o-0806`) (OpenAI et al., 2024a) to assign difficulty levels to the DeepScaleR dataset based on the AoPS rubric. Each problem was presented to GPT-4o with a request to rate its difficulty according to AoPS guidelines (the full prompt is shown in Appendix B.3). This approach provides a lightweight and scalable alternative to rollout-based estimation. As shown in Figure 1b, GPT-4o's difficulty ratings also correlate well with the model success rates, with a Pearson correlation of $r = -0.32$ ($p < 0.05$), making it a practical proxy for curriculum scheduling when computational resources are constrained.

4.2 Dataset

We use the DeepScaleR dataset (Luo et al., 2025) as the training set. DeepScaleR compiles problems from multiple sources, including AIME from 1984 to 2023 and AMC prior to 2023. The dataset also includes problems from the Omni-MATH (Gao et al., 2024) and Still datasets (Team, 2025), which feature problems from various national and international math competitions. This results in a diverse and challenging training set, covering a wide range of mathematical domains and difficulty levels.

[1] https://artofproblemsolving.com/wiki/index.php/AoPS_Wiki:Competition_ratings

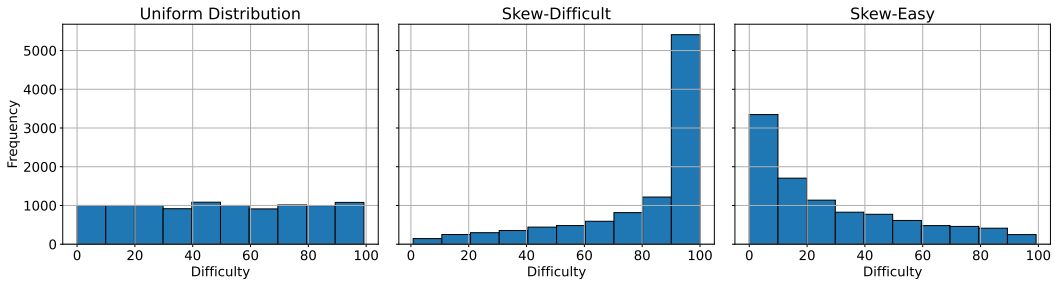


Figure 2: Difficulty distribution for different training sets: Uniform, Skew-Difficult, and Skew-Easy. Each training set contains 10,000 samples.

In practice, we do not have control over the exact difficulty distribution of the data collected for training. This motivates our investigation into how different difficulty distributions influence AdaRFT. To this end, we construct three distinct distributions from the DeepScaleR dataset. The first is a `skew-difficult` distribution, where most problems are challenging. The second is a `skew-easy` distribution, where most problems are relatively easy. The third is a `uniform` distribution, where problems are evenly balanced across all difficulty levels, ensuring a consistent representation of easy, moderate, and hard problems. Each of these three distributions includes 10,000 samples. The data distribution for each setting is shown in Figure 2.

For evaluation, we use six benchmark datasets to assess the model's performance across different levels of difficulty and mathematical reasoning. The first benchmark, MATH 500 (Lightman et al., 2023), is a subset of the MATH dataset (Hendrycks et al., 2021) containing 500 representative problems designed to test a model's general mathematical capability. GSM8K (Cobbe et al., 2021) is a set of grade-school math problems. OlympiadBench (He et al., 2024) includes a collection of problems from Olympiad-level mathematics and physics competitions. Minerva Math (Lewkowycz et al., 2022) is a curated set of undergraduate-level math problems that assess complex mathematical reasoning and symbolic manipulation. AMC 23 and AIME 24 include problems from the 2023 American Mathematics Competitions and the 2024 American Invitational Mathematics Examination, respectively. Since AMC 23 contains only 40 problems and AIME 24 only 30, we report accuracy as the average over 8 sampled responses per problem to ensure stable estimates. Together, these datasets span elementary, high school, and advanced competition-level math, providing a comprehensive evaluation of the model's reasoning abilities.

4.3 Training Setup

We trained two models on the three difficulty-based distributions of the DeepScaleR dataset described in Section 4.2: Qwen 2.5 7B and Qwen 2.5 MATH 1.5B. This setup allows us to evaluate the effectiveness of AdaRFT on models with different initial performance levels when exposed to skew-difficult, skew-easy, and uniform problem distributions. All models were trained using four different approaches: (1) the standard PPO algorithm, (2) AdaRFT (PPO), our method that integrates adaptive curriculum learning with PPO (see Section 3), (3) PPO with filtered data, a baseline that trains PPO on data filtered by pass@k accuracy, and (4) PPO with a fixed curriculum schedule.

For the data filtering baseline (3), following prior work (Bae et al., 2025; Hu et al., 2025; Zyphra, 2025), we first run a pass@40 analysis for each combination of model and data distribution. We then discard examples that are either too easy or too hard, removing all problems with solved rates $\leq 10\%$ or $\geq 90\%$. This restricts training to problems of intermediate difficulty. However, this procedure removes a large fraction of the data, including many potentially informative examples. In addition, because difficulty is defined using pass@k metrics, the filtering must be recomputed whenever the model or the data distribution changes.

For the fixed curriculum baseline (4), we follow the approach of prior work (Parashar et al., 2025; Team et al., 2025). In this setting, the difficulty of sampled problems follows a predetermined schedule that increases linearly over training steps. Suppose training runs for n total steps. We then

define a target difficulty $T(s)$ at step s by

$$T(s) = T_{\min} + \frac{T_{\max} - T_{\min}}{n} \cdot s,$$

and at each step we sample problems whose estimated difficulty matches this target. Unlike AdaRFT, this schedule increases difficulty at a fixed rate regardless of how quickly or slowly the model learns.

The training batch size was set to $B = 1024$, with the target reward β set to 0.5 to promote learning at a balanced success rate. The sensitivity parameter α and step size η were tuned using a validation set to ensure stable curriculum updates. We set $\alpha = 2$, $\eta = 50$, and the initial target difficulty $T = 0$. The step size η acts as a scaling factor between the reward signal and the difficulty metric. Since the difficulty metric ranges from 0 to 100 and the reward ranges from 0 to 1, a target reward $\beta = 0.5$ implies that the maximum reasonable adjustment to the difficulty metric should be around 50. Therefore, we set $\eta = 50$ to scale the reward signal appropriately to the difficulty range. The sensitivity parameter $\alpha = 2$ controls the slope of the $\tanh$ function. Setting α to 2 makes the $\tanh$ function behave approximately linearly when the difference between the average reward and the target reward is small. The intuition behind using the $\tanh$ function is that when the average reward is close to the target reward, a roughly linear adjustment is appropriate. However, when the average reward deviates significantly from the target reward, linear adjustments may be too large, leading to instability. The $\tanh$ function smooths out these adjustments, allowing for more controlled changes when the difference is large while maintaining sensitivity when the difference is small. Both models were trained on 8 A100 GPUs for approximately 100 steps. The implementation details can be found in Appendix B.

5 Results and Analysis

We evaluate the performance of standard PPO and AdaRFT (PPO) across multiple training setups and two model sizes: Qwen 2.5 MATH 1.5B and Qwen 2.5 7B. Figure 3 presents the learning curves averaged across six benchmarks, while Table 1 and 2 provide a detailed breakdown of accuracy and training efficiency. On average, models trained with AdaRFT (PPO) outperform their PPO-only counterparts in both final accuracy and training efficiency. This improvement is particularly notable in non-uniform data distributions, where curriculum adaptation is most beneficial.

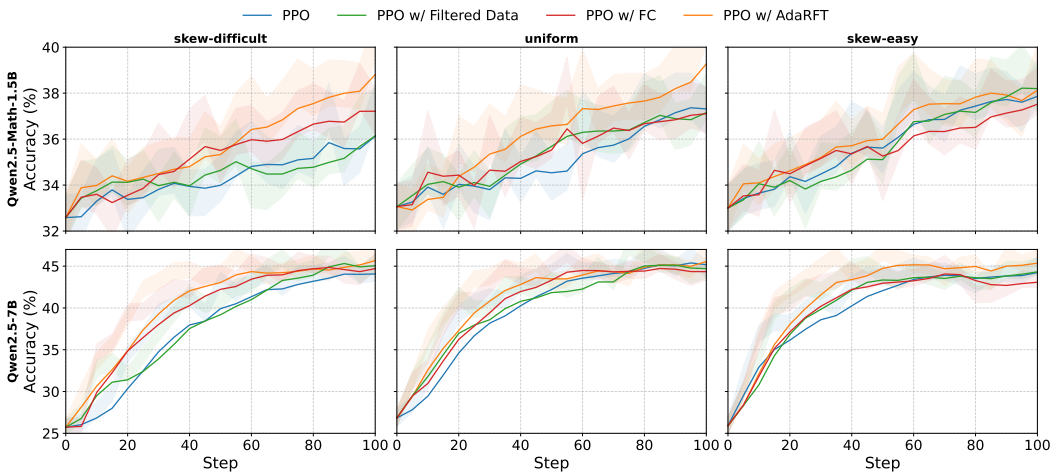


Figure 3: Performance comparison of PPO, PPO with filtered data, PPO with fixed curriculum (PPO w/ FC), and AdaRFT (PPO) across different setups (uniform, skew-easy, skew-difficult). Accuracy is the average of MATH 500, GSM8K, AIME 24, AMC 23, OlympiadBench, and Minerva Math. Compared with baselines, AdaRFT improves both the accuracy and training efficiency. For clarity, curves are exponentially smoothed.

5.1 TRAINING EFFICIENCY

As shown in Figure 3 and Table 1, models trained with ADARFT consistently require fewer training steps to match the performance of those trained with standard PPO, PPO on filtered data, and PPO with a fixed curriculum schedule. Specifically, we report how many additional steps are needed for PPO variants to match the performance of ADARFT at step 60 for Qwen 2.5 Math 1.5B, and step 40 for Qwen 2.5 7B. Because models are evaluated only every 5 training steps, we apply exponential smoothing with a smoothing parameter of 0.3 to the accuracy curves to reduce variance and obtain stable estimates of performance over time. The shaded areas in Figure 3 represent the raw, unsmoothed accuracy $\pm 1\%$, offering a visual cue for the typical fluctuation in evaluation accuracy. For Qwen 2.5 Math 1.5B, standard PPO requires 43 extra steps (+71.7%) in the skew-difficult setting and 34 steps (+56.7%) in the uniform setting to match ADARFT's performance. PPO with filtered training data requires even more: +49 steps (81.7%) and +52 steps (86.7%) in the respective settings. In the skew-easy scenario, PPO requires +16 steps (26.7%), while PPO with filtered data needs +21 steps (35.0%) to catch up to ADARFT. The efficiency gains remain significant with the larger Qwen 2.5 7B model. In the skew-difficult setting, PPO and PPO with filtered data require +24 steps (60.0%) and +25 steps (62.5%), respectively. PPO with a fixed curriculum schedule also follows this trend suggesting that while fixed curricula can modestly improve training efficiency, their inability to adapt the the model's evolving learning dynamics limits their convergence speed relative to ADARFT.

In addition to improved sample efficiency, ADARFT also achieves faster average training time per step across nearly all settings, as reported in Table 1. Though PPO with filtered data can sometimes offer marginal gains in per-step time (e.g., skew-easy setups), it still falls behind in convergence speed. This is largely due to the fact that easier problems require fewer tokens to solve. For example, an arithmetic reasoning question from GSM8K might require only around 200 tokens for Qwen 2.5 7B to reach a correct answer, whereas a competition-level math problem from AIME could require around 2000 tokens, a $10\times$ difference in rollout length. The total token length affects multiple components of the training step, including the rollout itself and the subsequent PPO update. While PPO update time does not scale linearly with sequence length due to batching and attention computation patterns, longer sequences still incur higher compute costs. As a result, curriculum learning's tendency to prioritize shorter, easier problems early in training leads to shorter sequences on average, reducing per-step compute and improving overall training throughput. These results underscore that ADARFT is both sample-efficient and compute-efficient, delivering faster and more cost-effective training.

Table 1: Average time per step (in seconds) at step 100 and extra steps required to match ADARFT's performance at step 60 (for Qwen 2.5 Math 1.5B) or step 40 (for Qwen 2.5 7B), across different setups and methods.

Model	Setup	Method	Avg Step Time (s)	Extra Steps (%)	Extra Steps	Extra Time (s)
Qwen2.5 Math 1.5B	skew-difficult	ADARFT	**122.24**	**0.0%**	**+0**	**0.00**
		PPO	132.95	71.7%	+43	5716.85
		PPO (w/ Filter)	128.20	81.7%	+49	6281.80
		PPO (w/ FC)	130.91	26.7%	+16	2094.56
	uniform	ADARFT	**121.31**	**0.0%**	**+0**	**0.00**
		PPO	126.82	56.7%	+34	4311.88
		PPO (w/ Filter)	126.35	86.7%	+52	6570.20
		PPO (w/ FC)	126.40	80.0%	+48	6067.20
	skew-easy	ADARFT	120.52	**0.0%**	**+0**	**0.00**
		PPO	121.15	26.7%	+16	1938.40
		PPO (w/ Filter)	**115.12**	35.0%	+21	2417.52
		PPO (w/ FC)	121.99	58.3%	+35	4269.65
Qwen2.5 7B	skew-difficult	ADARFT	**239.92**	**0.0%**	**+0**	**0.00**
		PPO	246.21	60.0%	+24	5909.04
		PPO (w/ Filter)	254.22	62.5%	+25	6355.50
		PPO (w/ FC)	243.12	22.5%	+9	2188.08
	uniform	ADARFT	**234.16**	**0.0%**	**+0**	**0.00**
		PPO	243.82	32.5%	+13	3169.66
		PPO (w/ Filter)	263.11	57.5%	+23	6051.53
		PPO (w/ FC)	240.62	17.5%	+7	1684.34
	skew-easy	ADARFT	247.44	**0.0%**	**+0**	**0.00**
		PPO	235.27	50.0%	+20	4705.40
		PPO (w/ Filter)	**233.13**	42.5%	+17	3963.21
		PPO (w/ FC)	240.66	57.5%	+23	5535.18

5.2 Model Performance

In addition to improving efficiency, AdaRFT (PPO) also improves the final model performance. As shown in Table 2, at the end of training (step 100), AdaRFT yields consistent improvements in final accuracy across all configurations. The reported averages reflect accuracy across six diverse benchmarks: GSM8K, MATH 500, OlympiadBench, Minerva Math, AMC 23, and AIME 24. On skew-difficult data, the Qwen 2.5 Math 1.5B model improves from 37.41% (PPO) to 40.48% (AdaRFT (PPO)), a gain of over 3 percentage points in average accuracy. Similar improvements appear in the uniform setting, where AdaRFT (PPO) reaches 41.11%, compared to 37.20% with PPO. Even on skew-easy data, where the baseline performs well, curriculum learning still improves performance, reaching 39.18% versus 38.46%. For the larger Qwen 2.5 7B model, final accuracy gains are also consistent, though slightly more modest. In the skew-difficult setting, AdaRFT (PPO) improves from 44.17% to 46.83%. In the uniform setting, accuracy rises from 44.70% to 46.92%, and in the skew-easy case, from 45.07% to 45.94%. These results show that AdaRFT is effective even for stronger models, enhancing both stability and peak performance.

Table 2: Accuracy (%) at step 100 for every model, setup, and benchmark. AdaRFT in this table refers to AdaRFT instantiated with PPO, i.e., AdaRFT (PPO).

Model	Setup	Method	GSM8K	MATH 500	Olympiad Bench	Minerva Math	AMC 23 (Avg@8)	AIME 24 (Avg@8)	Average
Qwen 2.5 Math 1.5B	skew-difficult	AdaRFT	**74.00**	**66.40**	**20.36**	**15.07**	**55.00**	**12.08**	**40.48**
		PPO	69.67	64.60	20.65	12.87	47.50	9.17	37.41
		PPO (w/ Filter)	71.65	62.40	20.06	**15.07**	45.00	9.17	37.22
		PPO (w/ FC)	72.55	66.40	20.95	14.34	45.00	4.06	37.22
	uniform	AdaRFT	**74.53**	**66.20**	**21.99**	14.34	**57.50**	**12.08**	**41.11**
		PPO	71.95	65.20	21.10	**15.81**	42.50	6.67	37.20
		PPO (w/ Filter)	72.63	65.80	20.21	13.60	45.00	10.00	37.87
		PPO (w/ FC)	71.95	65.60	19.61	14.71	42.50	9.27	37.27
	skew-easy	AdaRFT	73.62	66.20	19.91	**13.97**	**55.00**	9.17	**39.18**
		PPO	72.71	**67.40**	19.17	**13.97**	45.00	**12.50**	38.46
		PPO (w/ Filter)	**74.75**	65.20	**20.36**	13.60	45.00	10.00	38.15
		PPO (w/ FC)	72.40	66.20	20.06	13.24	50.00	6.67	38.09
Qwen 2.5 7B	skew-difficult	AdaRFT	**90.98**	71.40	**25.85**	22.43	**52.50**	**15.83**	**46.83**
		PPO	89.69	71.20	23.33	23.53	50.00	11.25	44.17
		PPO (w/ Filter)	88.48	**72.20**	24.37	25.00	50.00	12.08	45.35
		PPO (w/ FC)	89.92	72.00	24.52	25.37	47.50	13.96	45.54
	uniform	AdaRFT	**90.14**	72.60	**24.96**	24.26	**55.00**	14.58	**46.92**
		PPO	89.31	72.40	23.63	**25.37**	42.50	**15.00**	44.70
		PPO (w/ Filter)	89.08	**74.40**	23.18	22.43	45.00	13.33	44.57
		PPO (w/ FC)	89.84	72.40	23.92	23.90	42.50	13.33	44.32
	skew-easy	AdaRFT	**90.14**	72.60	**25.56**	23.16	**50.00**	**14.17**	**45.94**
		PPO	89.39	**73.60**	23.33	**24.26**	47.50	13.33	45.07
		PPO (w/ Filter)	89.31	71.60	24.22	23.90	47.50	13.33	44.98
		PPO (w/ FC)	89.46	72.20	24.37	23.90	37.50	13.33	43.46

5.3 Ablation on Target Reward β

To better understand the role of the target reward β in AdaRFT, we perform an ablation study varying β in the target difficulty update rule. Recall that β controls the target average reward the model is expected to achieve and implicitly steers the curriculum: lower values prioritize easier problems, while higher values shift the curriculum toward more challenging samples. We train a Qwen 2.5 Math 1.5B model on the uniform data distribution with AdaRFT (PPO) using three different values of β: 0.2, 0.5, and 0.8. For comparison, we also include standard PPO without AdaRFT (denoted as "w/o AdaRFT") as a baseline.

As shown in Figure 4, the model trained with $\beta = 0.5$ achieves the highest accuracy throughout training. This supports our theoretical motivation in Section 3.4: maximizing reward variance, which occurs when success rate ≈ 0.5, provides the strongest learning signal. Models with $\beta = 0.2$ and $\beta = 0.8$ underperform likely due to curriculum misalignment: $\beta = 0.8$ overly focuses on easy problems, while $\beta = 0.2$ overemphasizes difficult ones, both of which limit the model's capacity to generalize. The reward and difficulty curves align with the accuracy outcomes discussed above. The $\beta = 0.5$ configuration maintains a stable reward near 0.5, reflecting balanced difficulty exposure. In contrast, $\beta = 0.8$ results in overly high reward (i.e., easy samples), while $\beta = 0.2$ maintains a reward around 0.2 for most of training, indicating the model is repeatedly presented with overly

difficult problems. As expected, response length is the shortest for $\beta = 0.8$ and longest for $\beta = 0.2$, consistent with the idea that longer responses correlate with problem complexity.

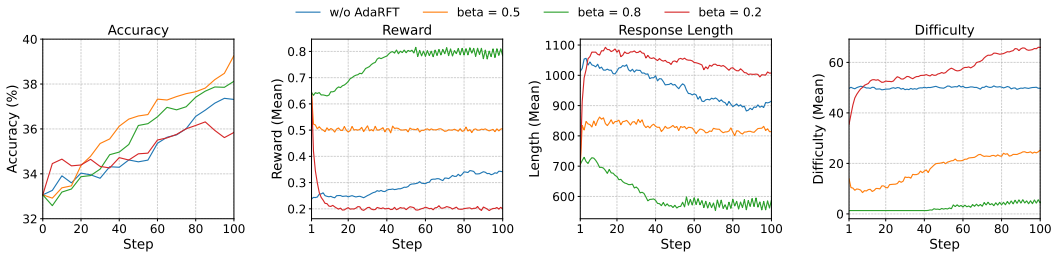


Figure 4: Ablation on β in AdaRFT: we compare model accuracy, average reward, response length, and mean difficulty under $\beta = 0.2$, $\beta = 0.5$, and $\beta = 0.8$, along with standard PPO (w/o AdaRFT).

6 When does curriculum learning help?

Our findings show that curriculum learning provides the greatest benefits under two key conditions: (1) imbalanced training distributions, and (2) limited model capacity. In skewed distributions, particularly the skew-difficult settings, standard PPO often struggles to gain traction early in training due to insufficient reward signals. AdaRFT mitigates this by initially sampling easier problems, enabling the model to bootstrap capabilities before tackling harder content. Conversely, the benefits of AdaRFT are less pronounced when the model is strong enough or the data is already well-balanced. In both cases, the model is either already exposed to a representative distribution of task difficulties or finds most problems challenging enough, thus reducing the need for dynamic difficulty adjustment. In addition, we conducted further experiments detailed in Appendix A, including evaluations on datasets with more extreme difficulty distributions (A.2), difficulty estimation using an LLM-based judge (A.4), and instantiations of AdaRFT with alternative RL algorithms (GRPO, REINFORCE++) (A.3). Across all these settings, AdaRFT consistently demonstrates effectiveness, highlighting its robustness to diverse data distributions, compatibility with various RL algorithms, and flexibility with different difficulty metrics.

It is important to note that manual data curation and task scheduling carefully designed for a specific model could potentially achieve similar results by selecting a training sequence that aligns with the model's learning capacity (Yu et al., 2025; Shen et al., 2025; Chen et al., 2025; Zeng et al., 2025). However, such methods require significant human effort, domain knowledge, and often need to be re-tuned for each new model or task. Fixed curriculum schedules face similar limitations: because they rely on a predetermined ordering of difficulty, they cannot adjust when the model learns faster or slower than expected. In contrast, AdaRFT requires no manual data curation or model-specific preprocessing. The curriculum automatically adapts to the model's reward signal during training, making it broadly applicable across model scales and training distributions. This automatic adaptability not only saves engineering effort but also improves scalability and robustness in real-world training pipelines. Moreover, AdaRFT is particularly advantageous in fixed data settings, as it can tailor the difficulty schedule to match the capabilities of any model, whether strong or weak, without altering the dataset, providing a unified solution that generalizes across model architectures and skill levels.

7 Conclusion

We propose AdaRFT, an adaptive curriculum learning strategy for reinforcement finetuning (RFT) that dynamically matches problem difficulty to a model's evolving skill level. By adjusting a target difficulty based on reward feedback, AdaRFT improves both sample and compute efficiency without modifying the reward function or underlying RL algorithm. Experiments across multiple data regimes and model sizes show consistent gains in convergence speed and final accuracy, especially in imbalanced training distributions. This lightweight, scalable approach highlights the value of curriculum-aware training for efficient and robust alignment in structured reasoning tasks.

REFERENCES

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, 2024. URL https://arxiv.org/abs/2402.14740.

Sanghwan Bae, Jiwoo Hong, Min Young Lee, Hanbyul Kim, JeongYeon Nam, and Donghyun Kwak. Online difficulty filtering for reasoning oriented reinforcement learning. *arXiv preprint arXiv:2504.03380*, 2025.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 41–48, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Zhipeng Chen, Yingqian Min, Beichen Zhang, Jie Chen, Jinhao Jiang, Daixuan Cheng, Wayne Xin Zhao, Zheng Liu, Xu Miao, Yang Lu, Lei Fang, Zhongyuan Wang, and Ji-Rong Wen. An empirical study on eliciting and improving r1-like reasoning models, 2025. URL https://arxiv.org/abs/2503.04548.

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.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards, 2025. URL https://arxiv.org/abs/2502.01456.

Wojciech Marian Czarnecki, Siddhant M. Jayakumar, Max Jaderberg, Leonard Hasenclever, Yee Whye Teh, Simon Osindero, Nicolas Heess, and Razvan Pascanu. Mix&match - agent curricula for reinforcement learning, 2018. URL https://arxiv.org/abs/1806.01780.

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.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment, 2023. URL https://arxiv.org/abs/2304.06767.

Carlos Florensa, David Held, Xinyang Geng, and Pieter Abbeel. Automatic goal generation for reinforcement learning agents, 2018. URL https://arxiv.org/abs/1705.06366.

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, 2024. URL https://arxiv.org/abs/2410.07985.

Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. Aligning language models with preferences through f-divergence minimization. *arXiv preprint arXiv:2302.08215*, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe

Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pp. 1352–1361. PMLR, 2017.

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

bench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024. URL https://arxiv.org/abs/2402.14008.

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. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models, 2025. URL https://arxiv.org/abs/2501.03262.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model, 2025. URL https://arxiv.org/abs/2503.24290.

Allan Jabri, Kyle Hsu, Ben Eysenbach, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Unsupervised curricula for visual meta-reinforcement learning, 2019. URL https://arxiv.org/abs/1912.04226.

Niels Justesen, Ruben Rodriguez Torrado, Philip Bontrager, Ahmed Khalifa, Julian Togelius, and Sebastian Risi. Illuminating generalization in deep reinforcement learning through procedural level generation, 2018. URL https://arxiv.org/abs/1806.10729.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment, 2024. URL https://arxiv.org/abs/2410.01679.

Tomasz Korbak, Hady Elsahar, Germán Kruszewski, and Marc Dymetman. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. *Advances in Neural Information Processing Systems*, 35:16203–16220, 2022.

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, 2022. URL https://arxiv.org/abs/2206.14858.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. Limr: Less is more for rl scaling, 2025. URL https://arxiv.org/abs/2502.11886.

Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models, 2024. URL https://arxiv.org/abs/2310.10505.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

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

Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. Teacher-student curriculum learning, 2017. URL https://arxiv.org/abs/1707.00183.

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, 2025. URL https://arxiv.org/abs/2501.19393.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian,

Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie

Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024a. URL https://arxiv.org/abs/2410.21276.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024b. URL https://arxiv.org/abs/2412.16720.

Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, and Shuiwang Ji. Curriculum reinforcement learning from easy to hard tasks improves llm reasoning, 2025. URL https://arxiv.org/abs/2506.06632.

Rémy Portelas, Cédric Colas, Katja Hofmann, and Pierre-Yves Oudeyer. Teacher algorithms for curriculum learning of deep rl in continuously parameterized environments, 2019. URL https://arxiv.org/abs/1910.07224.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang,

Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q*: Your language model is secretly a q-function. *arXiv preprint arXiv:2404.12358*, 2024.

Pierre Harvey Richemond, Yunhao Tang, Daniel Guo, Daniele Calandriello, Mohammad Gheshlaghi Azar, Rafael Rafailov, Bernardo Avila Pires, Eugene Tarassov, Lucas Spangher, Will Ellsworth, et al. Offline regularised reinforcement learning for large language models alignment. *arXiv preprint arXiv:2405.19107*, 2024.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks, 2022. URL https://arxiv.org/abs/1606.04671.

John Schulman, Xi Chen, and Pieter Abbeel. Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440*, 2017a.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017b. URL https://arxiv.org/abs/1707.06347.

Wei Shen, Guanlin Liu, Zheng Wu, Ruofei Zhu, Qingping Yang, Chao Xin, Yu Yue, and Lin Yan. Exploring data scaling trends and effects in reinforcement learning from human feedback, 2025. URL https://arxiv.org/abs/2503.22230.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

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

Sainbayar Sukhbaatar, Zeming Lin, Ilya Kostrikov, Gabriel Synnaeve, Arthur Szlam, and Rob Fergus. Intrinsic motivation and automatic curricula via asymmetric self-play, 2018. URL https://arxiv.org/abs/1703.05407.

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, Weixin Xu, 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, Zonghan Yang, and Zongyu Lin. Kimi k1.5: Scaling reinforcement learning with llms, 2025. URL https://arxiv.org/abs/2501.12599.

RUCAIBox STILL Team. Still-3-1.5b-preview: Enhancing slow thinking abilities of small models through reinforcement learning. 2025. URL https://github.com/RUCAIBox/Slow_Thinking_with_LLMs.

Rui Wang, Joel Lehman, Jeff Clune, and Kenneth O. Stanley. Paired open-ended trailblazer (poet): Endlessly generating increasingly complex and diverse learning environments and their solutions, 2019. URL https://arxiv.org/abs/1901.01753.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Lucas Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, Weizhu Chen, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Reinforcement learning for reasoning in large language models with one training example, 2025. URL https://arxiv.org/abs/2504.20571.

Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, Haosheng Zou, Yongchao Deng, Shousheng Jia, and Xiangzheng Zhang. Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond, 2025. URL https://arxiv.org/abs/2503.10460.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning, 2025. URL https://arxiv.org/abs/2502.03387.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Wojciech Zaremba and Ilya Sutskever. Learning to execute, 2015. URL https://arxiv.org/abs/1410.4615.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild, 2025. URL https://arxiv.org/abs/2503.18892.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data, 2025. URL https://arxiv.org/abs/2505.03335.

Zyphra. Zr1-1.5b, a small but powerful reasoning model for math and code, 2025. URL https://www.zyphra.com/post/introducing-zr1-1-5b-a-small-but-powerful-math-code-reasoning-model.

A Further Discussion

A.1 Training Dynamics: AdaRFT vs. Fixed Curriculum vs. PPO

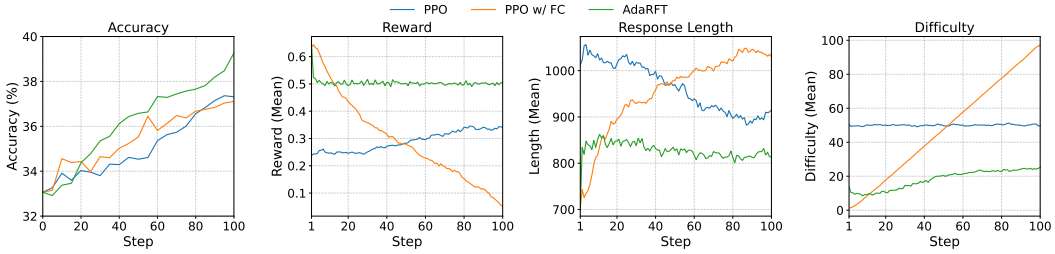


Figure 5: Training dynamics of standard PPO, PPO with a fixed curriculum (PPO w/ FC), and AdaRFT on the Qwen 2.5 Math 1.5B model under the uniform data distribution. We plot accuracy, reward, response length, and average difficulty of sampled training problems over training steps. Curves are exponentially smoothed ($\alpha = 0.3$) for clarity.

To further contextualize the role of adaptive curricula, we analyze the training dynamics of fixed curriculum (PPO w/ FC), standard PPO, and AdaRFT when training the Qwen 2.5 Math 1.5B model on the uniform distribution. Figure 5 illustrates the evolution of accuracy, reward, response length, and sampled difficulty across the first 100 training steps.

The accuracy curves highlight the central tradeoff of fixed curricula. PPO w/ FC initially converges slightly faster than standard PPO, benefiting from early exposure to easier problems. However, it ultimately underperforms compared to AdaRFT because its difficulty schedule increases independently of the model's actual learning progress. As a result, the model is pushed into harder problem regimes too quickly. This misalignment is clearly visible in the reward curve. Early in training, PPO w/ FC achieves rewards well above 0.5, indicating that the model is initially exposed to problems that are easier than its current capability. However, as training progresses, the fixed curriculum increases difficulty at a rate that outpaces the model's learning speed. As a result, the reward drops below 0.5 and continues declining, showing that the model is increasingly confronted with problems it cannot yet solve. This mismatch limits the effectiveness of the updates and ultimately leads to slower convergence compared to an adaptive curriculum.

The response length patterns reinforce this interpretation. Because longer responses typically correspond to more difficult reasoning tasks, the rapid increase in response length under PPO w/ FC shows that the curriculum escalates difficulty faster than the model can adapt. In contrast, standard PPO maintains more stable lengths but lacks the structured progression necessary for efficient learning. AdaRFT, by comparison, keeps response lengths moderate and gradually increasing, consistent with its difficulty traces: the curriculum raises problem difficulty only when the model's reward stays near the target value, ensuring that the model always receives examples that are challenging but solvable.

This particular dynamic suggest a general underlying issue: since the model's perceived difficulty changes over the course of training, a fixed curriculum cannot remain aligned, and different model-dataset combinations may experience extended periods of being over-challenged or under-challenged. This mismatch limits the effectiveness of updates and leads to slower convergence compared to an adaptive curriculum. In contrast, AdaRFT is designed to handle precisely this challenge: by adjusting the difficulty schedule based on the model's reward signal, it continually matches training difficulty to the model's evolving capability, ensuring sustained learning progress throughout training. Without the ability to adjust to real-time model performance, fixed schedules risk either overwhelming the model or wasting compute on overly easy tasks. AdaRFT avoids both extremes by maintaining the average reward near 0.5, automatically pacing the introduction of harder problems as the model improves. This adaptive alignment between task difficulty and model capability leads to smoother reward trajectories, controlled response lengths, and ultimately more stable and efficient learning.

A.2 Data Difficulty on Model Performance

To better understand the effect of data difficulty on model performance, we introduce two additional data distributions: easy-extreme and hard-extreme. Unlike the skew-difficult and skew-easy distributions, which still include a mix of difficulty levels, the easy-extreme and hard-extreme sets consist exclusively of the most polarized examples. Specifically, easy-extreme contains only the easiest samples with difficulty levels no greater than 15, while hard-extreme includes only the hardest samples with difficulty levels of at least 97. Each of these extreme distributions consists of approximately 8,000 samples, providing a focused and controlled evaluation of model behavior under minimal or maximal difficulty conditions. We trained a Qwen 2.5 7B model on each of the two extreme distributions using PPO, and compared their performance to models trained on the uniform distribution with PPO (Uniform) and with AdaRFT instantiated with PPO (Uniform + AdaRFT), as described in Section 5. The results are presented in Figure 6. The key takeaway is that training on only overly easy or hard problems fails to provide useful learning signals, reinforcing the need for AdaRFT to adaptively steer models toward challenges matched to their current ability.

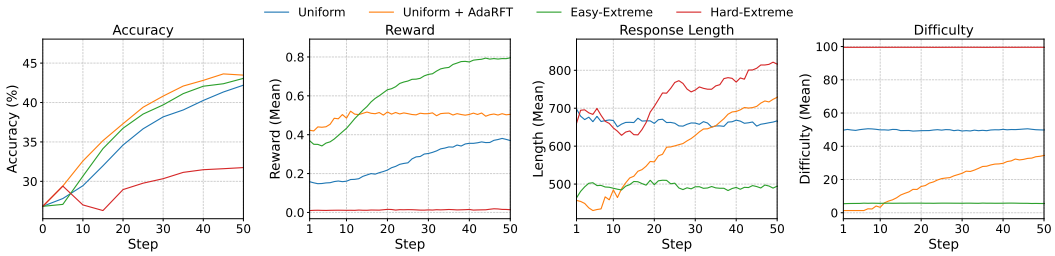


Figure 6: Performance comparison of Qwen 2.5 7B trained on different data distributions using PPO (Uniform, Easy-Extreme, Hard-Extreme) and AdaRFT instantiated with PPO (Uniform + AdaRFT). For clarity, curves are exponentially smoothed ($\alpha = 0.3$) to reduce noise.

Accuracy. The leftmost panel of Figure 6 shows that uniform + AdaRFT achieves the highest overall accuracy throughout training, outperforming both uniform and the two extreme settings. This highlights the effectiveness of AdaRFT in guiding the model through an optimal difficulty progression. In contrast, hard-extreme struggles significantly, with a flat and lower trajectory, indicating that exposing the model only to very difficult problems limits learning progress. This suggests that without a gradual exposure strategy, models trained on only the hardest problems are unable to bootstrap their capabilities effectively.

Reward. The reward trends provide important clues about learning dynamics. The easy-extreme setup achieves the fastest reward improvement during early training, surpassing both uniform and hard-extreme. In particular, easy-extreme consistently operates in a reward range between 0.4 and 0.6 during early training, which corresponds to a success rate that is both challenging and attainable. In contrast, the reward of the uniform and hard-extreme setup lingers below 0.2 in early training, leading to slower learning. This suggests that training on problems with intermediate difficulty—those that are neither trivially easy nor prohibitively hard—provides the most effective learning signal. Notably, AdaRFT is explicitly designed to exploit this insight: by setting the target reward $\beta = 0.5$, we encourage the model to train on problems that match this "productive struggle" zone. As shown by the uniform + AdaRFT curve, the algorithm successfully maintains an average reward near 0.5 throughout training, allowing the model to learn at an optimal pace. Notably, while the uniform setup eventually reaches a reward of nearly 0.5 by step 50, it does not result in faster learning. This is likely because the model is already fairly well trained by that stage, so the additional reward signal contributes less to further improvement. In contrast, the hard-extreme model receives almost no reward signal for most of the training, while the uniform setup shows slower and more gradual reward accumulation.

Response Length. The response length panel reveals how the complexity of generated solutions evolves during training. The hard-extreme model consistently produces the longest responses, with length increasing steadily, reflecting the higher complexity and reasoning depth required by the

hardest problems. In contrast, the easy-extreme setup maintains short and stable responses, consistent with its simpler problem set. The uniform and uniform + AdaRFT setups fall between these two extremes. Notably, uniform + AdaRFT shows a gradual increase in response length over time. This trend aligns with the behavior of the curriculum learning algorithm: as the model improves, it is exposed to increasingly difficult problems, which naturally demand more elaborate reasoning and longer solutions. This dynamic suggests that response length can serve as a useful proxy for problem difficulty and reasoning complexity during training.

Difficulty. Finally, the difficulty panel illustrates how problem difficulty evolves under each setup. The easy-extreme and hard-extreme curves remain flat, confirming that these datasets contain only problems from the tail ends of the difficulty spectrum (i.e., ≤ 15 and ≥ 97, respectively). The uniform curve is centered around 50, as expected, while uniform + AdaRFT shows a steady increase in difficulty over time. This adaptive progression confirms that curriculum learning effectively steers the model from easier to harder problems, aligning difficulty with the model's evolving capabilities.

A.3 AdaRFT with Diverse RL Algorithms

To evaluate the generality of AdaRFT beyond PPO, we trained the Qwen 2.5 Math 1.5B model on a skew-difficult data distribution using two alternative reinforcement learning algorithms: REINFORCE++ and GRPO (see implementation details in Appendix B). As shown in Figure 7, AdaRFT significantly improves both the convergence speed and final accuracy across these variants. Across both cases, the adaptive curriculum acts orthogonally to the underlying optimization method. These results reinforce the plug-and-play nature of AdaRFT: it consistently enhances sample efficiency and policy robustness across algorithmic choices, making it broadly applicable in diverse reinforcement finetuning pipelines. Notably, this generalization holds without any additional tuning or algorithm-specific modifications, underscoring the practical utility of curriculum-aware training in both lightweight and computation-heavy RFT settings.

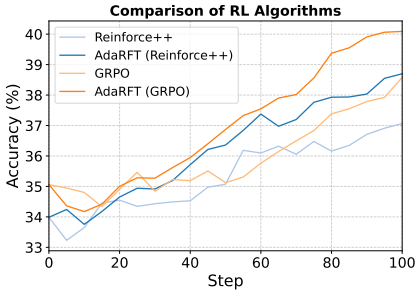


Figure 7: Comparison between models trained with and without AdaRFT using REINFORCE++ and GRPO.

A.4 Training on LLM-Estimated Difficulty

In addition to rollout-based difficulty estimation, we explore an alternative strategy that uses LLM-judged difficulty levels to guide curriculum construction. As described in Section 4.1, we prompt GPT-4o (`gpt-4o-0806`) to assign difficulty levels to math problems in the DeepScaleR dataset according to the AoPS rubric. This approach offers a lightweight and scalable alternative to computing pass@k success rates from model rollouts, making it especially attractive in low-resource scenarios.

To assess the effectiveness of this strategy, we train a Qwen 2.5 Math 1.5B model on the skew-difficult distribution using AdaRFT (PPO) with two curriculum schedules: one based on rollout-derived pass@k difficulty, and the other guided by GPT-4o's difficulty ratings. Since the LLM-judged difficulty is on a scale of 1 to 5 (rather than 0 to 100), we set the step size hyperparameter $\eta = 2.5$ to align the difficulty adjustment magnitude with the reward signal. All other hyperparameters are kept unchanged. As shown in Figure 8, both curriculum strategies outperform standard PPO without curriculum learning. While rollout-based difficulty estimation yields the strongest gains, the LLM-judged curriculum still provides a noticeable improvement over the baseline.

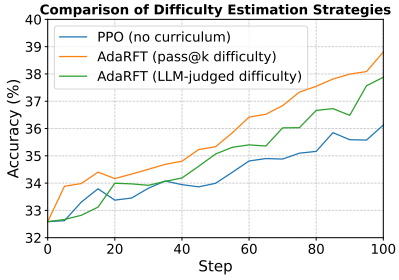


Figure 8: Comparison of different difficulty estimation strategies.

These results demonstrate that AdaRFT remains effective even when the difficulty signal is derived from heuristic or approximate sources like LLM judgments. Although less precise than empirical

pass@k metrics, the LLM-based difficulty still provides enough structure to enable meaningful curriculum adaptation. This makes it a practical fallback when rollout computation is too costly, and suggests that future work could explore hybrid approaches that combine lightweight heuristics with periodic empirical calibration.

B Implementation Details

B.1 Training Configuration

We trained both the actor and critic models using the PPO algorithm on a single node with 8 A100 GPUs. Each model was trained for approximately 100 optimization steps using the `veRL` library (Sheng et al., 2024). We used two model variants: `Qwen2.5-7B` and `Qwen2.5-MATH-1.5B`. The latter has a shorted context window, so we adjusted the max response length and the sequence parallel size accordingly.

Table 3 summarizes the core hyperparameter settings used across all three algorithms: PPO, GRPO, and REINFORCE++. We highlight both shared defaults and algorithm-specific overrides, including KL treatment modes, rollout settings, and critic configurations.

B.2 Deriving the Target-Difficulty Update Rule from a Linear Mapping

A central component of our curriculum mechanism is the update of the target difficulty T based on the model's observed reward performance. While the final update rule (Eq. 13) involves a hyperbolic tangent, it is in fact a smooth and stabilized variant of a standard linear mapping between reward space and difficulty space. We derive it here for clarity.

Step 1: Linear mapping between two intervals. The classical linear interpolation formula for mapping a value $v \in [x, y]$ to a target interval $[a, b]$ is

$$v' = a + \frac{(v - x)(b - a)}{y - x}. \tag{5}$$

If we directly map the average reward $R_{\text{avg}} \in [r_{\min}, r_{\max}]$ to the difficulty range $[d_{\min}, d_{\max}]$, we obtain

$$T_{\text{naive}}(R_{\text{avg}}) = d_{\min} + \frac{(R_{\text{avg}} - r_{\min})(d_{\max} - d_{\min})}{r_{\max} - r_{\min}}. \tag{6}$$

In our main setting, $r_{\min} = 0$, $r_{\max} = 1$, and $[d_{\min}, d_{\max}] = [0, 100]$, so the naive mapping simplifies to

$$T_{\text{naive}}(R_{\text{avg}}) = 100\, R_{\text{avg}}. \tag{7}$$

Step 2: Mapping reward deviation instead of absolute reward. For curriculum learning, we do not wish to reassign a new difficulty level at every step. Instead, we aim to *adjust* the current target difficulty depending on whether the model is performing above or below a desired target success rate β. We therefore consider the deviation

$$\delta = R_{\text{avg}} - \beta. \tag{8}$$

Given $R_{\text{avg}} \in [0, 1]$, the deviation satisfies $\delta \in [r_{\min} - \beta,\ r_{\max} - \beta] = [-\beta,\ 1 - \beta]$. With the common choice $\beta = 0.5$, this becomes $\delta \in [-0.5, 0.5]$.

Applying the linear mapping rule equation 5 from the deviation range $[-0.5,\ 0.5]$ to a symmetric difficulty-change interval $[-\Delta, \Delta]$ yields

$$\Delta T_{\text{lin}}(\delta) = -\Delta + \frac{(\delta - (-0.5))(\Delta - (-\Delta))}{0.5 - (-0.5)} = 2\Delta\, \delta. \tag{9}$$

Thus the naive linear controller becomes

$$T'_{\text{lin}} = T + 2\Delta\, (R_{\text{avg}} - \beta). \tag{10}$$

This already captures the desired behavior: difficulty increases when performance exceeds the target, decreases when performance falls short, and remains stable when $R_{\text{avg}} = \beta$.

Step 3: Stabilizing the update via a smooth saturating nonlinearity. A purely linear controller may cause excessively large changes when the reward deviation is large or noisy. To obtain a stable update rule, we replace the linear term with a smooth, odd, saturating nonlinearity. The hyperbolic tangent is a natural choice: it behaves linearly near zero (which recovers the linear mapping) and saturates as its argument grows.

We therefore define a smoothed difficulty adjustment

$$\Delta T(\delta) = \eta \cdot \tanh\big(\alpha(R_{\text{avg}} - \beta)\big). \tag{11}$$

Here, η sets the maximum update magnitude and α controls the sensitivity around the target reward. For small deviations, $\tanh(z) \approx z$, so locally

$$\Delta T(\delta) \approx \eta\alpha(R_{\text{avg}} - \beta), \tag{12}$$

recovering a linear controller with effective slope $\eta\alpha$ while ensuring global boundedness.

Step 4: Clipping to the valid difficulty range. To ensure the target difficulty remains within the observed range of the data, we apply a final clipping:

$$T' = \text{clip}\big(T + \eta \cdot \tanh\big(\alpha(R_{\text{avg}} - \beta)\big),\ d_{\min},\ d_{\max}\big)\,. \tag{13}$$

The full update rule equation 13 is therefore a direct, smoothed generalization of a naive linear mapping between reward deviations and difficulty adjustments. It preserves the intuitive behavior of the linear controller near the target reward, while the saturating nonlinearity and clipping ensure stable, bounded, and data-consistent curriculum updates.

Because the reward is bounded in $[0, 1]$ and the difficulty metric spans $[0, 100]$, we set the step size $\eta = 50$ to align their scales. The modulation parameter $\alpha = 2$ ensures smooth and controlled progression throughout training.

B.3 Prompt for Difficulty Estimation Using LLM as a Judge

The prompt used for difficulty estimation (as described in Section 4.1) is shown in Table 4, Table 5, and Table 6. The descriptions of the difficulty scales and examples are sourced from the AoPS Wiki.[2] Although GPT-4o is prompted to rate problem difficulty on a scale from 1 to 10, we found that over 95% of the problems fall within the range of 1 to 5. Therefore, we clip the scores and use a revised scale from 1 to 5. In addition to integer scores, we also allow half-point increments such as 1.5, 2.0, and 2.5 for finer-grained difficulty estimation.

C The Use of Large Language Models for ICLR 2026

In this ICLR submission, large language models (LLMs) were used solely as writing aids for grammar correction, wording refinement, and text polishing. They were not employed for idea generation, technical contributions, or any aspect of the research beyond enhancing readability and clarity.

[2] https://artofproblemsolving.com/wiki/index.php/AoPS_Wiki:Competition_ratings

Table 3: Comparison of training hyperparameters for PPO, GRPO, and REINFORCE++ using the veRL library. Shared defaults are used unless overridden.

Category	Parameter	PPO	GRPO	REINFORCE++
Algorithm-Specific Settings				
General	Advantage estimator	GAE	GRPO	REINFORCE++
	Gamma (γ)	1.0	—	—
	Lambda (λ)	1.0	—	—
	Batch size	1024	1024	1024
	Max prompt length	1024	1024	1024
	Gradient checkpointing	Enabled	Enabled	Enabled
Actor	Learning rate	1×10^{-6}	1×10^{-6}	1×10^{-6}
	Mini-batch size	1024	1024	1024
	Dynamic batch size	Enabled	Enabled	Enabled
	KL penalty role	Reward	Loss	Loss
	KL loss type	Fixed	Low-variance KL	MSE
	KL loss coefficient (β)	0.001	0.001	0.001
	Entropy coefficient	0.001	0.001	0
	Clip ratio	0.2	0.2	0.2
	Gradient clipping	1.0	1.0	1.0
	Sequence parallel size	Model-specific	Model-specific	Model-specific
Rollout	Backend	vLLM	vLLM	vLLM
	Tensor model parallel size	2	2	2
	Rollouts per sample	1	8	1
	Nucleus sampling p	1.0	1.0	1.0
	GPU memory utilization	0.5	0.5	0.5
	Sampling temperature	1.0	1.0	1.0
Critic	Warmup steps	0	—	—
	Learning rate	1×10^{-5}	—	—
	Sequence parallel size	Model-specific	—	—
Model-Specific Overrides (shared across all algorithms)				
Qwen2.5-7B	Max response length	8000	8000	8000
	Sequence parallel size	2	2	2
	Max token length / GPU	8000	8000	8000
Qwen2.5-MATH-1.5B	Max response length	3000	3000	3000
	Sequence parallel size	1	1	1
	Max token length / GPU	16000	16000	16000
ADARFT *Parameters*				
Curriculum Learning	Target reward (β)	0.5	0.5	0.5
	Sensitivity (α)	2	2	2
	Step size (η)	50	50	50
	Initial difficulty (T)	0	0	0

Prompt for Difficulty Estimation (Part 1)

 # Math Problem
{problem}

 # Your Task
You are a subject matter expert in mathematics tasked with evaluating the difficulty level of individual math problems. Your assessment should be objective and based on a detailed difficulty scale provided below. Your judgment will help calibrate and categorize problems for use in educational settings or assessments. Be thorough, fair, and consistent in your evaluation.

 # Difficulty Scale
1: Problems strictly for beginner, on the easiest elementary school or middle school levels (MOEMS, MATHCOUNTS School, AMC 8 1-10, AMC 10 1-10, easier AMC 12 1-5, and others that involve standard techniques introduced up to the middle school level), most traditional middle/high school word problems.
1.5: Problems for stronger beginner students, on the level of the middling problems in most middle school contests (AMC 8 11-20, harder AMC 10 1-10, AMC 12 1-5, and those others that force students to apply their school-level knowledge to slightly more challenging problems), traditional middle/high school word problems with more complex problem solving.
2: For motivated beginners, harder questions from the previous categories (AMC 8 21-25, MATHCOUNTS Chapter (Sprint 21-30, Target 6-8), MATHCOUNTS States/Nationals, AMC 10 11-15, AMC 12 5-10, easiest AIME 1-3)
2.5: More advanced beginner problems, hardest questions from previous categories (Harder AMC 8 21-25, harder MATHCOUNTS States questions, AMC 10 16-20, AMC 12 11-15, usual AIME 1-3)
3: Early intermediate problems that require more creative thinking (harder MATHCOUNTS National questions, AMC 10 21-25, AMC 12 15-20, hardest AIME 1-3, usual AIME 4-6).
4: Intermediate-level problems (AMC 12 21-25, hardest AIME 4-6, usual AIME 7-10).
5: More difficult AIME problems (11-13), simple proof-based Olympiad-style problems (early JBMO questions, easiest USAJMO 1/4).
6: High-leveled AIME-styled questions (14/15). Introductory-leveled Olympiad-level questions (harder USAJMO 1/4 and easier USAJMO 2/5, easier USAMO and IMO 1/4).
7: Tougher Olympiad-level questions, may require more technical knowledge (harder USAJMO 2/5 and most USAJMO 3/6, extremely hard USAMO and IMO 1/4, easy-medium USAMO and IMO 2/5).
8: High-level Olympiad-level questions (medium-hard USAMO and IMO 2/5, easiest USAMO and IMO 3/6).
9: Expert Olympiad-level questions (average USAMO and IMO 3/6).
9.5: The hardest problems appearing on Olympiads which the strongest students could reasonably solve (hard USAMO and IMO 3/6).
10: Historically hard problems, generally unsuitable for very hard competitions (such as the IMO) due to being exceedingly tedious, long, and difficult (e.g. very few students are capable of solving on a worldwide basis).

Table 4: Prompt for difficulty estimation using LLM as a judge.

Prompt for Difficulty Estimation (Part 2)

 # Examples

For reference, here are some sample problems from each of the difficulty levels 1-10:

<1: Jamie counted the number of edges of a cube, Jimmy counted the numbers of corners, and Judy counted the number of faces. They then added the three numbers. What was the resulting sum? (2003 AMC 8, Problem 1)

1: How many integer values of x satisfy $|x| < 3\pi$? (2021 Spring AMC 10B, Problem 1)

1.5: A number is called flippy if its digits alternate between two distinct digits. For example, 2020 and 37373 are flippy, but 3883 and 123123 are not. How many five-digit flippy numbers are divisible by 15? (2020 AMC 8, Problem 19)

2: A fair 6-sided die is repeatedly rolled until an odd number appears. What is the probability that every even number appears at least once before the first occurrence of an odd number? (2021 Spring AMC 10B, Problem 18)

2.5: A, B, C are three piles of rocks. The mean weight of the rocks in A is 40 pounds, the mean weight of the rocks in B is 50 pounds, the mean weight of the rocks in the combined piles A and B is 43 pounds, and the mean weight of the rocks in the combined piles A and C is 44 pounds. What is the greatest possible integer value for the mean in pounds of the rocks in the combined piles B and C? (2013 AMC 12A, Problem 16)

3: Triangle ABC with $AB = 50$ and $AC = 10$ has area 120. Let D be the midpoint of $\overline{AB}$, and let E be the midpoint of $\overline{AC}$. The angle bisector of $\angle BAC$ intersects $\overline{DE}$ and $\overline{BC}$ at F and G, respectively. What is the area of quadrilateral $FDBG$? (2018 AMC 10A, Problem 24)

3.5: Find the number of integer values of k in the closed interval $[-500, 500]$ for which the equation $\log(kx) = 2\log(x + 2)$ has exactly one real solution. (2017 AIME II, Problem 7)

4: Define a sequence recursively by $x_0 = 5$ and

$$x_{n+1} = \frac{x_n^2 + 5x_n + 4}{x_n + 6}$$

for all nonnegative integers n. Let m be the least positive integer such that

$$x_m \leq 4 + \frac{1}{2^{20}}.$$

In which of the following intervals does m lie?

(A) $[9, 26]$ **(B)** $[27, 80]$ **(C)** $[81, 242]$ **(D)** $[243, 728]$ **(E)** $[729, \infty)$ (2019 AMC 10B, Problem 24 and 2019 AMC 12B, Problem 22)

4.5: Find, with proof, all positive integers n for which $2^n + 12^n + 2011^n$ is a perfect square. (USAJMO 2011/1)

5: Find all triples (a, b, c) of real numbers such that the following system holds:

$$a + b + c = \frac{1}{a} + \frac{1}{b} + \frac{1}{c},$$

$$a^2 + b^2 + c^2 = \frac{1}{a^2} + \frac{1}{b^2} + \frac{1}{c^2}.$$

(JBMO 2020/1)

5.5: Triangle ABC has $\angle BAC = 60°$, $\angle CBA \leq 90°$, $BC = 1$, and $AC \geq AB$. Let H, I, and O be the orthocenter, incenter, and circumcenter of $\triangle ABC$, respectively. Assume that the area of pentagon $BCOIH$ is the maximum possible. What is $\angle CBA$? (2011 AMC 12A, Problem 25)

6: Let $\triangle ABC$ be an acute triangle with circumcircle ω, and let H be the intersection of the altitudes of $\triangle ABC$. Suppose the tangent to the circumcircle of $\triangle HBC$ at H intersects ω at points X and Y with $HA = 3, HX = 2$, and $HY = 6$. The area of $\triangle ABC$ can be written in the form $m\sqrt{n}$, where m and n are positive integers, and n is not divisible by the square of any prime. Find $m + n$. (2020 AIME I, Problem 15)

Table 5: Prompt for difficulty estimation using LLM as a judge.

Prompt for Difficulty Estimation (Part 3)

6.5: Rectangles BCC_1B_2, CAA_1C_2, and ABB_1A_2 are erected outside an acute triangle ABC. Suppose that

$$\angle BC_1C + \angle CA_1A + \angle AB_1B = 180^\circ.$$

Prove that lines B_1C_2, C_1A_2, and A_1B_2 are concurrent. (USAMO 2021/1, USAJMO 2021/2)
7: We say that a finite set $\mathcal{S}$ in the plane is balanced if, for any two different points A, B in $\mathcal{S}$, there is a point C in $\mathcal{S}$ such that $AC = BC$. We say that $\mathcal{S}$ is centre-free if for any three points A, B, C in $\mathcal{S}$, there is no point P in $\mathcal{S}$ such that $PA = PB = PC$.
Show that for all integers $n \geq 3$, there exists a balanced set consisting of n points. Determine all integers $n \geq 3$ for which there exists a balanced centre-free set consisting of n points. (IMO 2015/1)
7.5: Let $\mathbb{Z}$ be the set of integers. Find all functions $f : \mathbb{Z} \to \mathbb{Z}$ such that

$$xf(2f(y) - x) + y^2 f(2x - f(y)) = \frac{f(x)^2}{x} + f(yf(y))$$

for all $x, y \in \mathbb{Z}$ with $x \neq 0$. (USAMO 2014/2)
8: For each positive integer n, the Bank of Cape Town issues coins of denomination $\frac{1}{n}$. Given a finite collection of such coins (of not necessarily different denominations) with total value at most most $99 + \frac{1}{2}$, prove that it is possible to split this collection into 100 or fewer groups, such that each group has total value at most 1. (IMO 2014/5)
8.5: Let I be the incentre of acute triangle ABC with $AB \neq AC$. The incircle ω of ABC is tangent to sides BC, CA, and AB at D, E, and F, respectively. The line through D perpendicular to EF meets ω at R. Line AR meets ω again at P. The circumcircles of triangle PCE and PBF meet again at Q.
Prove that lines DI and PQ meet on the line through A perpendicular to AI. (IMO 2019/6)
9: Let k be a positive integer and let S be a finite set of odd prime numbers. Prove that there is at most one way (up to rotation and reflection) to place the elements of S around the circle such that the product of any two neighbors is of the form $x^2 + x + k$ for some positive integer x. (IMO 2022/3)
9.5: An anti-Pascal triangle is an equilateral triangular array of numbers such that, except for the numbers in the bottom row, each number is the absolute value of the difference of the two numbers immediately below it. For example, the following is an anti-Pascal triangle with four rows which contains every integer from 1 to 10.

$$\begin{array}{ccccccc} & & & 4 & & & \\ & & 2 & & 6 & & \\ & 5 & & 7 & & 1 & \\ 8 & & 3 & & 10 & & 9 \end{array}$$

Does there exist an anti-Pascal triangle with 2018 rows which contains every integer from 1 to $1 + 2 + 3 + \cdots + 2018$? (IMO 2018/3)
10: Prove that there exists a positive constant c such that the following statement is true: Consider an integer $n > 1$, and a set $\mathcal{S}$ of n points in the plane such that the distance between any two different points in $\mathcal{S}$ is at least 1. It follows that there is a line ℓ separating $\mathcal{S}$ such that the distance from any point of $\mathcal{S}$ to ℓ is at least $cn^{-1/3}$.
(A line ℓ separates a set of points S if some segment joining two points in $\mathcal{S}$ crosses ℓ.) (IMO 2020/6)

 # Return format
Please return the corresponding difficulty scale (integer) in \box{ }

Table 6: Prompt for difficulty estimation using LLM as a judge.