Learn Hard Problems During RL with Reference Guided Fine-tuning

Yangzhen Wu [1] **Shanda Li** [2] **Zixin Wen** [2] **Xin Zhou** [3] **Ameet Talwalkar** [2] **Yiming Yang** [2] **Wenhao Huang** [3] **Tianle Cai** [3]

Abstract

Reinforcement learning (RL) for mathematical reasoning can suffer from reward sparsity: for challenging problems, LLM fails to sample any correct trajectories, preventing RL from receiving meaningful positive feedback. At the same time, there often exists human written reference solutions along with the problem (e.g. problems from AoPS), but directly fine-tuning on these solutions offers no benefit because models often cannot imitate human proofs that lie outside their own reasoning distribution. We introduce reference-guided fine-tuning (ReGFT), a simple and effective method that utilizes human-written reference solutions to synthesize positive trajectories on hard problem and train on them before RL. For each problem, we provide the model with partial reference solution and let it to generate its own reasoning trace, ensuring the resulting trajectories remain in the model's reasoning space while still benefiting from reference guidance. Fine-tuning on these reference guided trajectories increases the number of solvable problems and produces a checkpoint that will get more positive rewards during RL. Across three benchmarks (AIME'24, AIME'25, Beyond-AIME), ReGFT consistently improves supervised accuracy, accelerates DAPO training, and raises the final performance plateau of RL. Our results show that ReGFT can effectively overcome reward sparsity and unlock stronger RL-based mathematical reasoning.

1. Introduction

Reinforcement learning (RL) has been proven effective for enhancing the reasoning capabilities of large language models (LLMs), as demonstrated by the success of OpenAI's o1 and DeepSeek's R1 (OpenAI et al., 2024; DeepSeek-AI et al., 2025). These breakthroughs establish a new paradigm in which increasing computational resources are allocated to post-training, with RL serving as the central mechanism for capability improvement.

Within this paradigm, Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as a central framework for mathematical reasoning. RLVR trains models by sampling multiple reasoning trajectories for each prompt and optimizing the model to increase the likelihood of trajectories that lead to correct solutions. Supervision is provided by an automatic, rule-based verifier that assigns rewards based solely on the correctness of the final answer, eliminating the need for human-label or reward model training. Through this process, RLVR systematically promotes reasoning paths consistent with verifiable correctness while suppressing incorrect ones, thereby eliciting and strengthening the latent reasoning capabilities present in the base model.

Despite its effectiveness, RL remains fundamentally constrained by the competence of the base model. In RLVR, learning signals are obtained only when sampled reasoning trajectories receive positive rewards from the verifier. For problems that exceed the model's current reasoning capability–such as complex problems or unseen reasoning tasks–the base model often fails to produce any correct trajectories, leading to sparse or zero rewards. In these regimes, training effectively stalls due to the absence of meaningful gradient signals. This limitation renders RL both inefficient and fragile, as substantial computation is wasted on trajectories that provide no informative feedback.

In this work, we investigate how a fixed training dataset can be utilized to maximally elicit a model's mathematical reasoning capability through reinforcement learning, with a particular focus on mitigating reward sparsity during RL. In a typical training dataset, each problem is accompanied by a reference chain-of-thought (CoT) solution, which is commonly used for supervised fine-tuning (SFT). However, directly fine-tuning on these reference solutions often yields limited performance gains, due in part to a mismatch between the model's intrinsic reasoning patterns and the structure of the provided reference, leading to poor gener-

[1]UC Berkeley [2]Carnegie Mellon University [3]Bytedance Seed. Correspondence to: Yangzhen Wu <yangzhen_wu@berkeley.edu>.

The 3rd *AI for Math Workshop* at the 43rd International Conference on Machine Learning (ICML), Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

alization (Luong et al., 2024). To address this issue, prior work has proposed REinforced Fine-Tuning (ReFT) (Luong et al., 2024), which performs on-policy sampling and applies SFT to model-generated trajectories that are verified as correct. While ReFT can alleviate reward sparsity for problems where the model is already capable of producing correct solutions, it fails to improve learning on harder problems that lie beyond the model's initial capability, where no correct trajectories can be sampled and reinforcement signals remain absent.

To overcome this limitation, we propose Reference-Guided Finetuning (ReGFT), which explicitly leverages reference solutions to synthesize training trajectories for hard problems that the model cannot solve independently. Instead of directly fine-tuning on reference CoTs, ReGFT uses reference solutions as guidance to elicit model-generated reasoning traces that are aligned with the model's own reasoning process while remaining solution-consistent. The model is then fine-tuned on a mixture of these reference-guided trajectories and its self-generated reasoning traces. This process substantially increases the model's pass rate on hard training problems, resulting in denser and more informative reward signals during subsequent reinforcement learning. In experiments, we found that

- As shown in Figure 2, models initialized with ReGFT consistently outperform the raw checkpoint throughout RL training across all benchmarks, exhibiting both faster performance gains in early stages and higher final accuracy, indicating that reference-guided fine-tuning effectively alleviates reward sparsity and improves optimization efficiency.
- Comparative results show that directly fine-tuning on reference solutions alone is insufficient; instead, learning benefits critically depend on model-derived reasoning trajectories that are guided by references but remain aligned with the model's own generation process, which is essential for robust learning and generalization.
- Analysis of `pass@k` under increasing inference budgets reveals that RL can improves models' `pass@k` performance. And ReGFT yields more stable and sustained improvements for all compute budget, while gains from ReFT are primarily limited to low-k regimes and do not scale as effectively.

2. Related Work

This work studies a central question in reinforcement learning for language models: *How can RL enable a model to solve problems that the base model is originally unable to solve?*

Recent studies (Yue et al., 2025; Wu et al., 2026) show that naïve RL often fails to expand a model's intrinsic problem-solving frontier. In particular, they observe that RL training can exhibit `pass@k` saturation as k increases and may even underperform the model's raw inference performance on problems that are initially unsolvable. These results suggest that standard RL primarily amplifies existing behaviors rather than inducing genuinely new reasoning capabilities.

Motivated by this limitation, a growing body of work modifies the RL pipeline to enable learning beyond the model's original competence—specifically, to improve performance on hard or previously unsolvable problems. Existing approaches can be broadly grouped into three complementary directions.

Scaling RL and Adaptive Sampling. Several works attribute the limitations of naïve RL to insufficient or poorly allocated exploration rather than a fundamental inability of RL to improve reasoning. These methods improve RL either by prolonging training horizons (Liu et al., 2025; Hu et al., 2025) or by adaptively reallocating rollout budgets toward harder and more informative instances (Li et al., 2025b; Yang et al., 2025b). Concretely, they scale exploration breadth via increased rollouts per example, extend training depth to allow new strategies to emerge, and design difficulty-aware sampling schemes that focus updates on problems with stronger learning signals. Together, these approaches suggest that scaling and redistributing exploration during RL can improve learning on problems where uniform, short-horizon RL shows limited progress.

Question Augmentation during RL. Another line of work improves learning on hard problems by modifying the question distribution during RL to provide more informative training signals. (Li et al., 2025a) augments difficult questions with partial solutions or scaffolding, increasing the likelihood of sampling rewarding trajectories and improving sample efficiency. (Chen et al., 2025) similarly introduces self-generated or externally provided hints for initially unsolvable questions, transforming zero-reward cases into learnable instances. These approaches demonstrate that augmenting hard questions during RL can help models make progress when standard RL receives little or no learning signal.

Interleaving SFT and RL. A third direction combines supervised fine-tuning (SFT) and RL to address the difficulty of discovering correct trajectories on hard problems. (Ma et al., 2025) explicitly alternates between online SFT and RL, using SFT to introduce solutions on the hardest questions and RL to reinforce and generalize these behaviors. In contrast, (Fu et al., 2025) proposes a single-stage objective that jointly optimizes supervised and reinforcement

losses to improve training stability without phase switching. (Zhou et al., 2026) augments RL with self-generated explanations as auxiliary supervision to guide exploration toward structured reasoning trajectories, while (Zhang et al., 2025b) anchors RL rollouts to partial expert prefixes when self-generated trajectories fail, ensuring the presence of successful rollouts and densifying the reward signal.

Our Difference. Our work differs from the above approaches in both *timing* and *methodology*. While prior methods modify RL itself—by injecting supervised objectives during RL training or guiding exploration with expert-prefixed rollouts—we instead focus on improving the model's competence *before* RL. Specifically, we investigate whether targeted SFT prior to RL can raise the model's baseline capability on hard problems such that subsequent RL naturally receives positive rewards on instances that were previously unsolvable. This perspective is supported by observations that stronger mid-training leads to more effective downstream RL (Zhang et al., 2025a), and is orthogonal to the choice of RL algorithm.

The closest related works are ReFT (Luong et al., 2024) and RL Teachers (Cetin et al., 2025). ReFT fine-tunes models on self-generated correct traces to increase the probability of producing correct trajectories on hard problems, while RL Teachers distill trajectories generated by an RL-trained teacher model. In contrast, our method interleaves the model's own reasoning with expert solutions available in the training data to synthesize trajectories for hard problems *prior to RL*, thereby improving the model's initial competence and enabling more effective reward-driven learning in subsequent RL stages.

3. Approach

On-policy finetuning with model-generated trajectories (ReFT). We first consider REinforced Fine-Tuning (ReFT) (Luong et al., 2024), which applies supervised finetuning to correct reasoning trajectories generated by the model itself. By restricting supervision to on-policy, verified trajectories, ReFT aligns training signals with the model's intrinsic reasoning patterns and avoids the mismatch often observed when directly fine-tuning on human-written reference solutions. This approach is effective for problems where the model is already capable of producing correct trajectories, as it increases the likelihood of such trajectories and improves the model's readiness for subsequent reinforcement learning. However, ReFT fundamentally depends on the availability of model-generated correct solutions and therefore does not apply to harder problems where the base model fails to produce any correct trajectories.

Reference-Guided Finetuning for Unsolved Problems (ReGFT). To extend supervised learning beyond the model's initial capability boundary, we propose *Reference-Guided Finetuning* (ReGFT). ReGFT is designed to improve the model's ability to generate correct reasoning trajectories on problems that are initially unsolvable by the base model. Similar to prior work that assumes access to reference solutions in the training data (Zhang et al., 2025b; Li et al., 2025a; Chen et al., 2025; Cetin et al., 2025), we leverage these references to construct effective and informative training signals for hard problems.

Concretely, ReGFT synthesizes training trajectories by prompting the model with partial reference solutions as hints. These hints provide high-level structural guidance about the solution while requiring the model to reason from the beginning using its own generative process, rather than copying or completing a reference trajectory. This design preserves the model's intrinsic reasoning style and avoids the brittleness associated with directly fine-tuning on full reference solutions. Unlike BREAD, which injects expert prefixes during RL rollouts, and RL Teachers, which distill trajectories generated by an RL-trained teacher model, ReGFT operates entirely in a pre-RL supervised stage and focuses on improving the model's initial competence on hard problems. As validated in later sections, maintaining self-generated reasoning is crucial for effective generalization and for enabling downstream reinforcement learning. In practice, we fine-tune the model on a mixture of self-generated correct trajectories (as in ReFT) and reference-guided trajectories constructed with such hints, which increases the model's pass rate on hard problems and yields a stronger initialization for subsequent RL. An overview of the ReGFT pipeline is shown in Figure 1. We train only on hard problems, defined as those with less than 25% accuracy under the original model when sampled 16 times. Easier problems are excluded to avoid overfitting and to focus learning on cases where the model lacks sufficient competence.

Reinforcement learning from improved initialization Starting from checkpoints produced by ReFT or ReGFT, we can apply reinforcement learning to further improve reasoning performance. In practice, ReGFT checkpoints tend to be stronger and are more capable of generating trajectories that receive positive rewards on previously unsolvable questions, making them a better starting point for RL.

Among existing reinforcement learning algorithms for reasoning (Schulman et al., 2017; Shao et al., 2024; Yu et al., 2025b; Zheng et al., 2025), we adopt DAPO (Decoupled Clip and Dynamic sAmpling Policy Optimization) (Yu et al., 2025a), a variant of GRPO(Shao et al., 2024) as our reinforcement learning algorithm. DAPO is a strong and well-validated method that improves training stability and

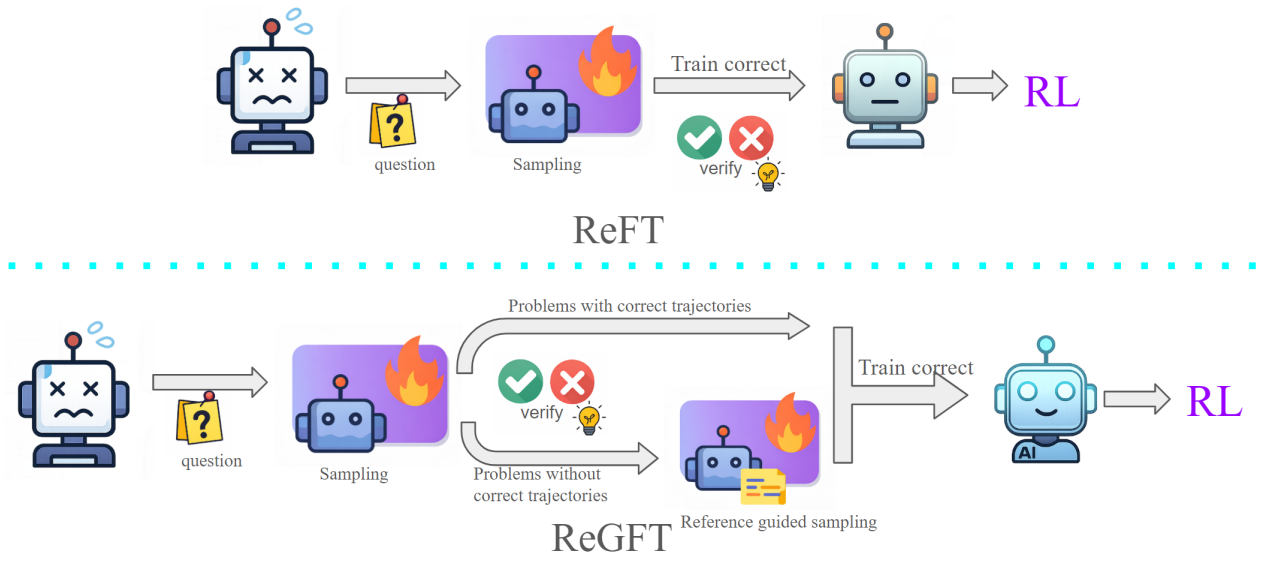


Figure 1. Comparison of ReFT and ReGFT. **Top:** ReFT fine-tunes the model using verified correct trajectories obtained from standard sampling. **Bottom:** ReGFT additionally applies reference-guided sampling to recover hard problems without correct trajectories.

sample efficiency through two key designs: decoupled clipping, which enables flexible reinforcement of high-quality trajectories without collapsing policy diversity, and dynamic sampling, which ensures that each update is based on reward-diverse trajectory groups and thus provides informative learning signals even under sparse rewards.

We deliberately choose DAPO to demonstrate that the benefits of ReGFT are orthogonal to advances in reinforcement learning algorithms. Despite DAPO already incorporating dynamic sampling to mitigate uninformative gradients, initializing training from ReGFT checkpoints consistently leads to further improvements. This shows that ReGFT enables sample more corrected generated trajectories themselves, rather than relying on improved optimization or sampling strategies during RL. Consequently, ReGFT complements even state-of-the-art RL methods like DAPO and can be combined with them to yield additional gains.

4. Experiments

4.1. Experimental Setup

Model We select Qwen3-4B-2507-Instruct (Yang et al., 2025a) as our base model. This choice reflects a careful balance between model capability and feasible training compute. At the 4B scale, Qwen3-4B is the strongest open-source instruction-tuned model that our available compute can support for full reinforcement learning experiments.

Qwen3-4B-Instruct exhibits strong reasoning performance while maintaining robust instruction-following ability, achieving competitive results on challenging benchmarks such as AIME. These properties make it particularly well suited for studying verifier-based reinforcement learning on hard reasoning tasks. In contrast, other popular model families at comparable or larger scales are less suitable under our setting: prior work shows that Qwen2.5 models can be substantially influenced by spurious reward signals (Shao et al., 2025), while LLaMA 3 models struggle to reach competitive performance on our hardest benchmarks (Grattafiori et al., 2024).

Overall, Qwen3-4B-Instruct provides the strongest and most reliable foundation within our compute budget, enabling controlled and meaningful evaluation of our proposed methods.

Datasets We use OmniMath (Gao et al., 2024), a benchmark of 4,428 Olympiad-level mathematics problems, each paired with a verified reference solution, as our training dataset. OmniMath is particularly well suited to our study for two reasons. First, its high difficulty leads to severe reward sparsity: for many problems, standard model rollouts fail to produce any correct solutions even with extensive sampling, making it a natural testbed for methods targeting sparse- or zero-reward reinforcement learning. Second, the availability of reference solutions enables controlled investigation of reference-guided strategies, allowing us to evaluate how auxiliary supervision can improve learning without relying on additional external data. For evaluation, we use the widely adopted AIME24&25 benchmarks and BeyondAIME (ByteDance Seed, 2025), which is more challenging than AIME and remains unsaturated by frontier models, consisting of 100 questions.

Implementation Details We use verl (Sheng et al., 2024) as our training framework for both SFT and RL. All DAPO training runs converge in approximately two days on 64 H20 GPUs. For evaluation, we use SGLang (Zheng et al., 2024) to accelerate inference. We set the maximum generation length to 16,384 tokens, which is the default for Qwen3-4B-Instruct on reasoning tasks and is rarely reached in practice, ensuring fair comparisons across methods. During generation, we use a temperature of 0.7 and top-p sampling with $p = 0.9$, following standard practice for reasoning-focused benchmarks.

4.2. Main Results

4.2.1. Reference-Guided Finetuning Enhances RL Training

To evaluate whether reference-guided fine-tuning improves downstream reinforcement learning, we perform DAPO-based RL starting from both the original raw checkpoint and the ReGFT checkpoint. During training, we sample 64 responses per prompt to promote sufficient exploration of potentially correct reasoning trajectories; the effect of varying this sampling budget is analyzed separately in Section 4.2.4.

As shown in Figure 2, models initialized with ReGFT consistently outperform those initialized from the raw checkpoint across all three benchmarks throughout the entire RL process. Notably, the ReGFT-initialized models achieve higher accuracy at convergence and exhibit substantially faster improvement during the early and middle stages of training. This behavior indicates that ReGFT provides a stronger starting point for reinforcement learning, enabling more effective progress with fewer RL updates.

These results suggest that reference-guided fine-tuning alleviates the reward sparsity that commonly hinders reinforcement learning on hard reasoning tasks. By increasing the likelihood that sampled trajectories contain correct ones on hard problems, ReGFT supplies RL with more informative learning signals. As a result, RL converges more rapidly and reaches a higher performance ceiling, demonstrating that ReGFT improves not only optimization efficiency but also the attainable final performance under a strong RL algorithm.

4.2.2. Impact of reference-Guided Demonstrations

ReGFT differs from ReFT in that it incorporates reference-guided demonstrations in addition to self-generated correct trajectories, whereas ReFT relies exclusively on the model's own successful rollouts. To isolate the contribution of human guidance, we perform DAPO-based reinforcement learning starting from both ReFT and ReGFT checkpoints. As shown in Figure 2 and Figure 3, initializing RL from the ReFT checkpoint already leads to faster early-stage improvements compared to the raw model. This confirms that ReFT effectively densifies reward signals by increasing the probability of sampling correct trajectories, thereby improving optimization efficiency in the initial phase of RL. However, despite this advantage, ReFT consistently underperforms ReGFT at convergence across all benchmarks, and even show inferior performance compared with raw DAPO on BeyondAIME(ByteDance Seed, 2025) benchmark.

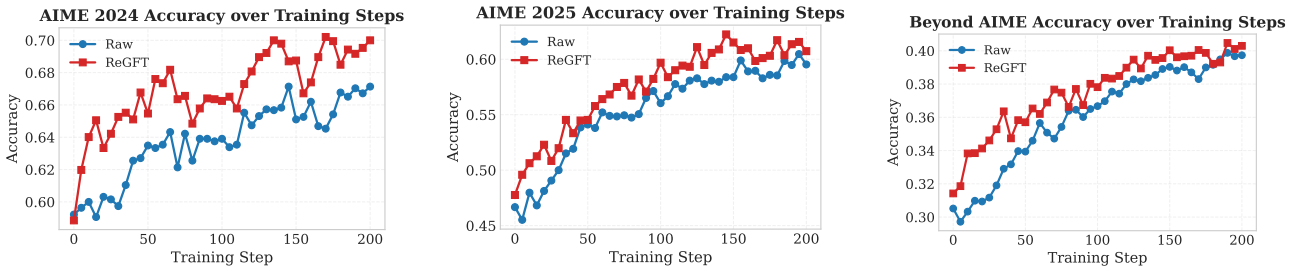
The persistent performance gap indicates that ReFT primarily accelerates learning without fundamentally expanding the model's reasoning capabilities. In contrast, ReGFT achieves higher asymptotic accuracy, suggesting that reference-guided demonstrations play a critical role beyond optimization speed. By introducing reasoning patterns that the model cannot reliably discover through self-exploration alone, human reference guidance enables ReGFT to extend the model's competence to previously unsolvable problems. Consequently, ReGFT not only improves training efficiency but also raises the achievable performance limit after reinforcement learning.

4.2.3. Necessity of Model-Derived Reasoning

To further investigate how reference solutions should be incorporated, we compare ReGFT with a baseline that directly fine-tunes the model on raw human-written reference solutions, without requiring the model to generate its own reasoning. This ablation examines whether exposure to correct human reasoning alone is sufficient to improve downstream RL performance.

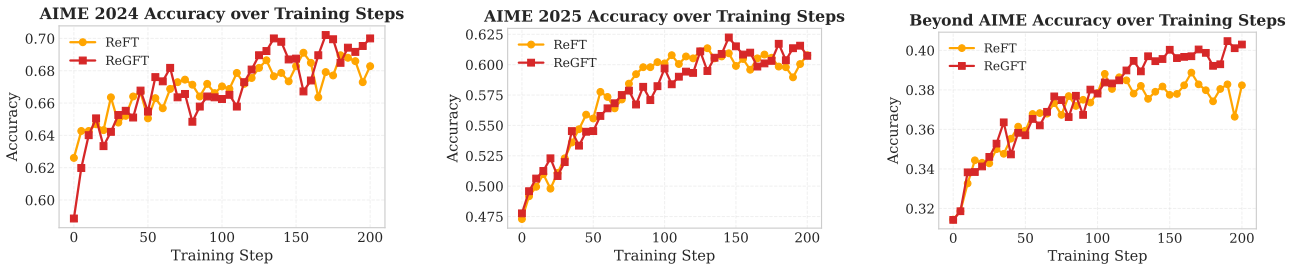
As shown in Figure 4, directly fine-tuning on unprocessed human solutions leads to substantially weaker performance, both during supervised training and after reinforcement learning. This outcome indicates that simply providing correct reasoning traces is insufficient: the model struggles to internalize reasoning patterns that are not expressed in its own generative style or latent reasoning space.

In contrast, ReGFT prompts the model to produce its own reasoning trajectories under partial reference guidance, resulting in solutions that are both correct and distributionally aligned with the model's inference process. These model-derived trajectories serve as an effective interface between human reference demonstrations and reinforcement learning. The results demonstrate that model-generated reasoning is a necessary component for successful knowledge transfer, and that ReGFT provides a principled mechanism for incorporating reference guidance without breaking this alignment.



(a) AIME 2024 *(b)* AIME 2025 *(c)* Beyond AIME

Figure 2. Reinforcement learning performance over training steps on three challenging benchmarks. Models initialized with **ReGFT** consistently achieve higher accuracy, faster convergence, and a superior final performance compared to the raw checkpoint, demonstrating that reference-guided fine-tuning provides a stronger initialization for RL.



(a) AIME 2024 *(b)* AIME 2025 *(c)* Beyond AIME

Figure 3. Reinforcement learning performance comparison between **ReFT** and **ReGFT** across three benchmarks. While both initializations accelerate early-stage RL, ReGFT consistently achieves higher final accuracy, highlighting the contribution of reference-guided demonstrations beyond self-generated trajectories.

4.2.4. Inference-Time Scaling Performance

We evaluate inference-time scaling by measuring `pass@k` accuracy. For each problem, we sample $N = 1024$ independent generations and estimate `pass@k` as

$$\text{pass@k} \;=\; 1 - \frac{\binom{N-c}{k}}{\binom{N}{k}},$$

where c denotes the number of correct samples among the N generations.

As shown in Figure 5, reinforcement learning consistently improves inference-time scaling behavior. In contrast to prior observations that report early `pass@k` saturation (Yue et al., 2025), we observe substantial gains across a wide range of k. We attribute this to the difference of the base model, as well as we have a larger sampling budget during RL.

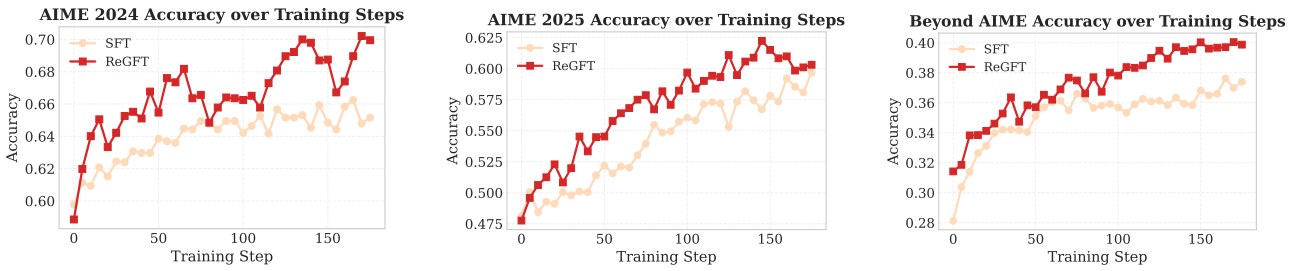
Across all benchmarks, **ReGFT + DAPO** exhibits the strongest and most consistent scaling performance. It dominates ReFT- and raw-based baselines on AIME 2025 and Beyond-AIME, and outperforms them across both small and large k regimes on AIME 2024. Notably, ReGFT maintains a clear performance margin as k increases, indicating improved coverage of the solution space rather than gains limited to a small number of samples.

In contrast, while ReFT improves `pass@1` over raw DAPO, these gains diminish as k grows. We hypothesize that this behavior arises because ReFT primarily reinforces the model's existing successful trajectories, which can bias exploration and limit the diversity of solutions sampled at inference time. By incorporating external reference guidance during pre-RL finetuning, ReGFT expands the model's underlying competence on hard problems, leading to consistently higher `pass@k` under increased test-time compute. These results highlight the effectiveness of ReGFT in enabling robust inference-time scaling beyond what can be achieved by self-reinforcement alone.

4.3. Additional Experimental Evidence

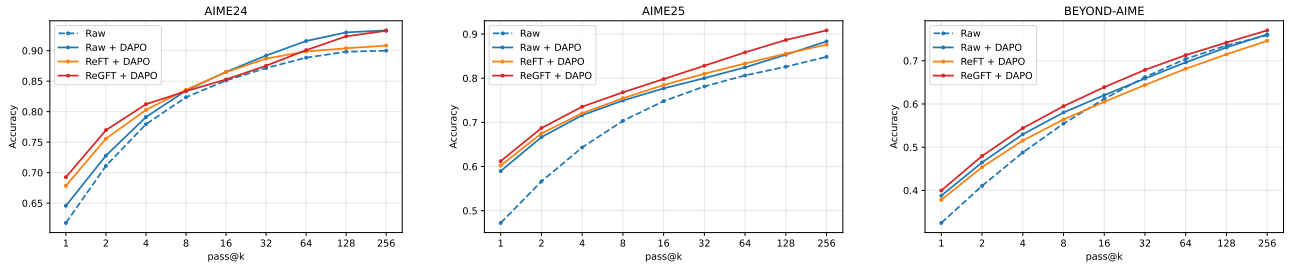
4.3.1. Reference-Guided Generation Enables the Model to Solve Previously Unsolvable Problems

We evaluate whether reference-guided generation can expand the set of problems for which the model is able to produce correct reasoning trajectories on the OmniMath dataset. We compare two sampling strategies: normal sampling and reference-guided sampling. For each problem, we sample 64 independent reasoning trajectories under each strategy. In the reference-guided condition, the model is provided with the first 80% sentences of the human-written



(a) AIME 2024 *(b)* AIME 2025 *(c)* Beyond AIME

Figure 4. Accuracy comparison between direct SFT on raw human reference solutions and **ReGFT**. Models trained directly on reference solutions fail to achieve competitive RL performance, underscoring the importance of model-derived reasoning trajectories.



(a) AIME 2024 *(b)* AIME 2025 *(c)* Beyond AIME

Figure 5. Inference-time scaling performance (`pass@k`) of raw and RL-trained checkpoints across benchmarks. Solid lines denote DAPO-trained models, while dashed line indicate the raw pre-RL checkpoint.

reference solution as contextual guidance, while the remaining 20%—which typically contains the final answer—is withheld. This design aims to avoid trivial answer copying, although in practice we found that the model almost always derives its own reasoning independently even when exposed to the full reference solution.

Across the 4,428 problems in OmniMath, standard sampling solves 68.58% of the problems, whereas reference-guided sampling solves 70.82%. More importantly, reference-guided sampling enables the model to solve an additional 5.85% of problems that are never solved under standard sampling, indicating that reference guidance can unlock reasoning trajectories that are otherwise inaccessible to the model. Conversely, 3.61% of problems are solved only by standard sampling. In ReGFT, we train on correct solutions from both sampling strategies on these hard problems, thereby maximizing the probability of correct solution generation.

Despite access to human-written guidance, performance remains well below 100%. We identify two primary factors contributing to this gap. First, the model's ability to interpret and internalize complex human reasoning is limited, particularly when reference solutions rely on advanced mathematical tools or domain-specific insights that fall outside the model's training distribution. Second, the rule-based verifier used in OmniMath cannot reliably assess open-ended or proof-style solutions, leading to false negatives even when the model's reasoning is substantively correct. These limitations highlight both the challenges of reference-guided generation and the constraints imposed by current automatic verification methods.

4.3.2. Performance of the Fine-Tuned Models

We evaluate the fine-tuned models obtained via ReFT and ReGFT on the OmniMath training set, as well as on challenging out-of-distribution benchmarks, including AIME 2024, AIME 2025, and Beyond-AIME. Table 1 summarizes the results. Both fine-tuning strategies yield consistent performance improvements over the base model across the training set and all evaluation benchmarks, indicating that supervised fine-tuning on self-generated trajectories is effective for strengthening mathematical reasoning.

Comparing the two approaches, the ReGFT checkpoint achieves a higher pass rate on the OmniMath training set than the ReFT checkpoint. This improvement reflects the benefit of incorporating reference-guided trajectories, which expand the set of problems for which correct solutions are available during supervised fine-tuning. In contrast, performance on the external evaluation benchmarks remains largely comparable between ReFT and ReGFT, suggesting that the additional reference guidance does not lead to overfitting to the training distribution or degradation in generalization.

Table 1. Performance comparison of models fine-tuned under different settings. We report pass@64 accuracy on the training dataset and evaluation results on AIME24, AIME25, and BEYONDAIME.

Model	Training Dataset (%)	Train pass@64 (%)	AIME24 (%)	AIME25 (%)	BeyondAIME (%)
Raw	48.4	68.6	59.2	46.7	30.5
ReFT	49.6	70.2	62.1	46.8	31.3
ReGFT	50.1	72.5	60.0	47.8	31.3

Table 2. Effect of sampling scale per problem under different training settings on AIME 2024, AIME 2025, and Beyond-AIME.

Setting	AIME'24	AIME'25	Beyond-AIME
Raw model	59.2	46.7	30.5
DAPO			
Response size 16	63.2	54.3	36.1
Response size 64	67.1	60.5	39.8
ReFT + DAPO			
Response size 16	66.2	57.8	35.0
Response size 64	68.3	60.8	38.3
ReGFT + DAPO			
Response size 16	67.0	56.7	37.4
Response size 64	**70.0**	**61.6**	**40.3**

Taken together, these results indicate that ReGFT primarily improves the model's ability to solve hard training problems by increasing the density and diversity of correct reasoning trajectories, while preserving robust generalization to unseen mathematical tasks. As a result, ReGFT provides a stronger initialization for subsequent reinforcement learning, where denser reward signals are critical for stable and effective optimization.

4.3.3. EFFECT OF SAMPLING SCALE PER PROBLEM

A common strategy for mitigating reward sparsity in reinforcement learning is to increase the number of sampled responses per problem, thereby raising the probability that at least one trajectory receives a positive reward (Hu et al., 2025). In this ablation, we examine how sampling scale interacts with different model initializations by comparing response sizes of 16 and 64 under DAPO, ReFT+DAPO, and ReGFT+DAPO.

As shown in Table 2, increasing the response size from 16 to 64 consistently improves performance across all settings, confirming that larger sampling budgets help alleviate the zero-reward regime by providing denser learning signals. However, scaling alone is not sufficient: even with 64 responses per problem, models trained without reference-guided fine-tuning lag behind those initialized with ReGFT. In particular, ReGFT+DAPO achieves the strongest results across all benchmarks, indicating that improving the model's initial competence and increasing exploration scale are complementary and jointly necessary for robust gains.

5. Conclusion

This work addresses a central bottleneck in reinforcement learning for mathematical reasoning: reward sparsity caused by the absence of correct model-generated trajectories on hard problems. When a model cannot produce any rewarding samples, reinforcement learning stalls regardless of the algorithm or sampling budget.

We introduce Reference-Guided Fine-Tuning (ReGFT), which uses reference solutions to expand the model's solvable problem set before reinforcement learning. Rather than training directly on reference chains of thought, ReGFT conditions the model on partial references and requires it to generate its own reasoning trajectories, producing solutions that are both correct and aligned with the model's inference distribution. This increases the density and diversity of verifiable trajectories, enabling reinforcement learning to receive meaningful signals on problems that were previously unsolvable.

Across OmniMath, AIME 2024, AIME 2025, and Beyond-AIME, ReGFT improves supervised pass rates, accelerates DAPO training, and consistently yields higher final performance than both raw and ReFT-based baselines. Importantly, ReGFT also improves inference-time scaling: its advantage persists as k increases in pass@k evaluation, indicating more reliable discovery of correct solutions under additional test-time compute rather than gains confined to a small number of samples.

Overall, our results demonstrate that ReGFT mitigates reward sparsity by converting reference solutions into model-derived correct trajectories, yielding a stronger initialization that makes downstream RL more effective. The resulting models not only achieve higher accuracy after RL, but also exhibit stronger and more stable pass@k scaling.

Limitations Our approach assumes access to reference solutions or expert-provided signals for training problems, which may not be available in all domains. In addition,

while ReGFT improves performance on the benchmarks we study, its effectiveness may depend on the quality and coverage of the available references.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

ByteDance Seed. Beyondaime: Advancing math reasoning evaluation beyond high school olympiads. https://huggingface.co/datasets/ByteDance-Seed/BeyondAIME, 2025.

Cetin, E., Zhao, T., and Tang, Y. Reinforcement learning teachers of test time scaling, 2025. URL https://arxiv.org/abs/2506.08388.

Chen, J. C.-Y., Peng, B. X., Choubey, P. K., Huang, K.-H., Zhang, J., Bansal, M., and Wu, C.-S. Nudging the boundaries of llm reasoning, 2025. URL https://arxiv.org/abs/2509.25666.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Fu, Y., Chen, T., Chai, J., Wang, X., Tu, S., Yin, G., Lin, W., Zhang, Q., Zhu, Y., and Zhao, D. Srft: A single-stage method with supervised and reinforcement fine-tuning for reasoning, 2025. URL https://arxiv.org/abs/2506.19767.

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

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan,

P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Hu, J., Liu, M., Lu, X., Wu, F., Harchaoui, Z., Diao, S., Choi, Y., Molchanov, P., Yang, J., Kautz, J., and Dong, Y. Brorl: Scaling reinforcement learning via broadened exploration, 2025. URL https://arxiv.org/abs/2510.01180.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, J., Lin, H., Lu, H., Wen, K., Yang, Z., Gao, J., Wu, Y., and Zhang, J. Questa: Expanding reasoning capacity in llms via question augmentation, 2025a. URL https://arxiv.org/abs/2507.13266.

Li, Z., Chen, C., Yang, T., Ding, T., Sun, R., Zhang, G., Huang, W., and Luo, Z.-Q. Knapsack rl: Unlocking exploration of llms via optimizing budget allocation, 2025b. URL https://arxiv.org/abs/2509.25849.

Liu, M., Diao, S., Lu, X., Hu, J., Dong, X., Choi, Y., Kautz,

J., and Dong, Y. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models, 2025. URL https://arxiv.org/abs/2505.24864.

Luong, T. Q., Zhang, X., Jie, Z., Sun, P., Jin, X., and Li, H. Reft: Reasoning with reinforced fine-tuning, 2024. URL https://arxiv.org/abs/2401.08967.

Ma, L., Liang, H., Qiang, M., Tang, L., Ma, X., Wong, Z. H., Niu, J., Shen, C., He, R., Li, Y., Cui, B., and Zhang, W. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions, 2025. URL https://arxiv.org/abs/2506.07527.

OpenAI, :, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., Iftimie, A., Karpenko, A., Passos, A. T., Neitz, A., Prokofiev, A., Wei, A., Tam, A., Bennett, A., Kumar, A., Saraiva, A., Vallone, A., Duberstein, A., Kondrich, A., Mishchenko, A., Applebaum, A., Jiang, A., Nair, A., Zoph, B., Ghorbani, B., Rossen, B., Sokolowsky, B., Barak, B., McGrew, B., Minaiev, B., Hao, B., Baker, B., Houghton, B., McKinzie, B., Eastman, B., Lugaresi, C., Bassin, C., Hudson, C., Li, C. M., de Bourcy, C., Voss, C., Shen, C., Zhang, C., Koch, C., Orsinger, C., Hesse, C., Fischer, C., Chan, C., Roberts, D., Kappler, D., Levy, D., Selsam, D., Dohan, D., Farhi, D., Mely, D., Robinson, D., Tsipras, D., Li, D., Oprica, D., Freeman, E., Zhang, E., Wong, E., Proehl, E., Cheung, E., Mitchell, E., Wallace, E., Ritter, E., Mays, E., Wang, F., Such, F. P., Raso, F., Leoni, F., Tsimpourlas, F., Song, F., von Lohmann, F., Sulit, F., Salmon, G., Parascandolo, G., Chabot, G., Zhao, G., Brockman, G., Leclerc, G., Salman, H., Bao, H., Sheng, H., Andrin, H., Bagherinezhad, H., Ren, H., Lightman, H., Chung, H. W., Kivlichan, I., O'Connell, I., Osband, I., Gilaberte, I. C., Akkaya, I., Kostrikov, I., Sutskever, I., Kofman, I., Pachocki, J., Lennon, J., Wei, J., Harb, J., Twore, J., Feng, J., Yu, J., Weng, J., Tang, J., Yu, J., Candela, J. Q., Palermo, J., Parish, J., Heidecke, J., Hallman, J., Rizzo, J., Gordon, J., Uesato, J., Ward, J., Huizinga, J., Wang, J., Chen, K., Xiao, K., Singhal, K., Nguyen, K., Cobbe, K., Shi, K., Wood, K., Rimbach, K., Gu-Lemberg, K., Liu, K., Lu, K., Stone, K., Yu, K., Ahmad, L., Yang, L., Liu, L., Maksin, L., Ho, L., Fedus, L., Weng, L., Li, L., McCallum, L., Held, L., Kuhn, L., Kondraciuk, L., Kaiser, L., Metz, L., Boyd, M., Trebacz, M., Joglekar, M., Chen, M., Tintor, M., Meyer, M., Jones, M., Kaufer, M., Schwarzer, M., Shah, M., Yatbaz, M., Guan, M. Y., Xu, M., Yan, M., Glaese, M., Chen, M., Lampe, M., Malek, M., Wang, M., Fradin, M., McClay, M., Pavlov, M., Wang, M., Wang, M., Murati, M., Bavarian, M., Rohaninejad, M., McAleese, N., Chowdhury, N., Chowdhury, N., Ryder, N., Tezak, N., Brown, N., Nachum, O., Boiko, O., Murk, O., Watkins, O., Chao, P., Ashbourne, P., Izmailov, P., Zhokhov, P., Dias, R., Arora, R., Lin, R., Lopes, R. G., Gaon, R., Miyara, R., Leike, R., Hwang, R., Garg, R., Brown, R., James, R., Shu, R., Cheu, R., Greene, R., Jain, S., Altman, S., Toizer, S., Toyer, S., Miserendino, S., Agarwal, S., Hernandez, S., Baker, S., McKinney, S., Yan, S., Zhao, S., Hu, S., Santurkar, S., Chaudhuri, S. R., Zhang, S., Fu, S., Papay, S., Lin, S., Balaji, S., Sanjeev, S., Sidor, S., Broda, T., Clark, A., Wang, T., Gordon, T., Sanders, T., Patwardhan, T., Sottiaux, T., Degry, T., Dimson, T., Zheng, T., Garipov, T., Stasi, T., Bansal, T., Creech, T., Peterson, T., Eloundou, T., Qi, V., Kosaraju, V., Monaco, V., Pong, V., Fomenko, V., Zheng, W., Zhou, W., McCabe, W., Zaremba, W., Dubois, Y., Lu, Y., Chen, Y., Cha, Y., Bai, Y., He, Y., Zhang, Y., Wang, Y., Shao, Z., and Li, Z. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Shao, R., Li, S. S., Xin, R., Geng, S., Wang, Y., Oh, S., Du, S. S., Lambert, N., Min, S., Krishna, R., Tsvetkov, Y., Hajishirzi, H., Koh, P. W., and Zettlemoyer, L. Spurious rewards: Rethinking training signals in rlvr, 2025. URL https://arxiv.org/abs/2506.10947.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., and Wu, C. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Wu, F., Xuan, W., Lu, X., Liu, M., Dong, Y., Harchaoui, Z., and Choi, Y. The invisible leash: Why rlvr may or may not escape its origin, 2026. URL https://arxiv.org/abs/2507.14843.

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

Yang, Z., Guo, Z., Huang, Y., Wang, Y., Xie, D., Wang, Y., Liang, X., and Tang, J. Depth-breadth synergy in rlvr: Unlocking llm reasoning gains with adaptive exploration, 2025b. URL https://arxiv.org/abs/2508.13755.

Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., Liu, X., Lin, H., Lin, Z., Ma, B., Sheng, G., Tong, Y., Zhang, C., Zhang, M., Zhang, W., Zhu, H., Zhu, J., Chen, J., Chen, J., Wang, C., Yu, H., Song, Y., Wei, X., Zhou, H., Liu, J., Ma, W.-Y., Zhang, Y.-Q., Yan, L., Qiao, M., Wu, Y., and Wang, M. Dapo: An open-source llm reinforcement learning system at scale, 2025a. URL https://arxiv.org/abs/2503.14476.

Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., YuYue, Dai, W., Fan, T., Liu, G., Liu, J., Liu, L., Liu, X., Lin, H., Lin, Z., Ma, B., Sheng, G., Tong, Y., Zhang, C., Zhang, M., Zhang, R., Zhang, W., Zhu, H., Zhu, J., Chen, J., Chen, J., Wang, C., Yu, H., Song, Y., Wei, X., Zhou, H., Liu, J., Ma, W.-Y., Zhang, Y.-Q., Yan, L., Wu, Y., and Wang, M. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=2a36EMSSTp.

Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Yue, Y., Song, S., and Huang, G. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?, 2025. URL https://arxiv.org/abs/2504.13837.

Zhang, C., Neubig, G., and Yue, X. On the interplay of pre-training, mid-training, and rl on reasoning language models, 2025a. URL https://arxiv.org/abs/2512.07783.

Zhang, X., Huang, Z., Li, Y., Ni, C., Chen, J., and Oymak, S. Bread: Branched rollouts from expert anchors bridge sft & rl for reasoning, 2025b. URL https://arxiv.org/abs/2506.17211.

Zheng, C., Liu, S., Li, M., Chen, X.-H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A., Zhou, J., and Lin, J. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Zheng, L., Yin, L., Xie, Z., Sun, C., Huang, J., Yu, C. H., Cao, S., Kozyrakis, C., Stoica, I., Gonzalez, J. E., Barrett, C., and Sheng, Y. Sglang: Efficient execution of structured language model programs, 2024. URL https://arxiv.org/abs/2312.07104.

Zhou, R., Li, S., Zhang, A., and Leqi, L. Expo: Unlocking hard reasoning with self-explanation-guided reinforcement learning, 2026. URL https://arxiv.org/abs/2507.02834.

A. RL Training Implementation Details.

Our reinforcement learning setup follows the DAPO(Yu et al., 2025b) framework, and we adopt its recommended hyperparameter configuration unless otherwise stated. We use the AdamW optimizer with a constant learning rate of 1×10^{-6}, together with a linear warm-up over the first 20 rollout steps. During rollout, we adjust the prompt batch size based on the number of sampled responses per prompt: when sampling 16 responses per prompt, we use a batch size of 512 prompts, and when sampling 64 responses per prompt, we use a batch size of 128 prompts. The mini-batch size is set to 2048, corresponding to 4 gradient updates per rollout step. For the Clip-Higher mechanism, we set the lower and upper clipping thresholds to $\varepsilon_{\text{low}} = 0.2$ and $\varepsilon_{\text{high}} = 0.28$, same as in the DAPO paper.

B. Prompts.

Prompt for Solution Generation

```
Question: {question}
Please reason step by step, and put your final answer within $boxed{}$.
```

Prompt for Reference Guided Sampling

```
Question: {question}
Hint: {partial reference solution}
Given the partial reference solution known to be correct as hints, derive
your solution to the question. You must solve it by yourself, and may follow
the ideas in the hint. Please reason step by step, and put your final answer
within $boxed{}$.
```

C. Problem Examples.

We present representative training examples from OmniMath along with their corresponding truncated reference solutions used as guidance. The reference solutions are concise and omit detailed reasoning process, highlighting the need for reference-guided sampling to encourage the model to construct its own reasoning trajectories.

Problem in Omnimath 1

Suppose $a_i, b_i, c_i, i = 1, 2, \cdots, n,$ are $3n$ real numbers in the interval $[0, 1]$. Define

$$S = \{(i, j, k) \,|\, a_i + b_j + c_k < 1\}, \quad T = \{(i, j, k) \,|\, a_i + b_j + c_k > 2\}.$$

Now we know that $|S| \geq 2018$, $|T| \geq 2018$. Try to find the minimal possible value of n.

Provided Partial Reference Solution 1

Suppose a_i, b_i, c_i for $i = 1, 2, \ldots, n$ are $3n$ real numbers in the interval $[0, 1]$. Define the sets

$$S = \{(i, j, k) \mid a_i + b_j + c_k < 1\}$$

and

$$T = \{(i, j, k) \mid a_i + b_j + c_k > 2\}.$$

We are given that $|S| \geq 2018$ and $|T| \geq 2018$. We aim to find the minimal possible value of n.
To establish a lower bound for n, consider the projections of the sets S and T onto the coordinate planes. Note that $S_{xy} \cap T_{xy} = \emptyset$, meaning that no pair (a_i, b_j) can simultaneously satisfy $a_i + b_j + c_k < 1$ and $a_i + b_j + c_k > 2$ for any c_k.
Thus, we have the inequalities:

$$|S_{xy}| + |T_{xy}| \leq n^2, \quad |S_{yz}| + |T_{yz}| \leq n^2, \quad |S_{zx}| + |T_{zx}| \leq n^2.$$

Applying the Projection Inequality and Hölder's Inequality, we obtain:

$$\begin{aligned} 2 \cdot 2018^{2/3} &\leq |S|^{2/3} + |T|^{2/3} \leq |S_{xy}|^{1/3} \cdot |S_{yz}|^{1/3} \cdot |S_{zx}|^{1/3} + |T_{xy}|^{1/3} \cdot |T_{yz}|^{1/3} \cdot |T_{zx}|^{1/3} \\ &\leq (|S_{xy}| + |T_{xy}|)^{1/3} (|S_{yz}| + |T_{yz}|)^{1/3} (|S_{zx}| + |T_{zx}|)^{1/3} \leq n^2. \end{aligned}$$

Problem in OmniMath 2

For a given integer $n \geq 2$, let $a_0, a_1, \ldots, a_n$ be integers satisfying $0 = a_0 < a_1 < \ldots < a_n = 2n - 1$. Find the smallest possible number of elements in the set $\{a_i + a_j \mid 0 \leq i \leq j \leq n\}$.

Provided Partial Reference Solution 2

For a given integer $n \geq 2$, let $a_0, a_1, \ldots, a_n$ be integers satisfying $0 = a_0 < a_1 < \ldots < a_n = 2n - 1$. We aim to find the smallest possible number of elements in the set $\{a_i + a_j \mid 0 \leq i \leq j \leq n\}$.
First, we prove that the set $\{a_i + a_j \mid 1 \leq i \leq j \leq n - 1\}$ takes all residues modulo $2n - 1$. Consider the $2n$ numbers:

$$a_0 < a_1 < \cdots < a_{n-1} < a_n$$

and

$$r - a_0 > r - a_1 > \cdots > r - a_{n-1} > r - a_n$$

for any integer $0 \leq r \leq 2n - 2$. By the Pigeonhole Principle, there must be two numbers that are congruent modulo $2n - 1$. Since $a_i \not\equiv a_j \pmod{2n - 1}$ for $1 \leq i < j \leq n - 1$, there exist $1 \leq i, j \leq n - 1$ such that $a_i \equiv r - a_j \pmod{2n - 1}$, meaning $a_i + a_j \equiv r \pmod{2n - 1}$.
Thus, the set $\{a_i + a_j \mid 1 \leq i \leq j \leq n - 1\}$ takes all residues modulo $2n - 1$.
Returning to the original problem, we note that there are $2n + 1$ distinct numbers:

$$a_0 + a_0 < a_0 + a_1 < \cdots < a_0 + a_{n-1} < a_n + a_0 < a_n + a_1 < \cdots < a_n + a_n,$$

which, modulo $2n - 1$, take only n different residues. Combining this with the fact that $\{a_i + a_j \mid 1 \leq i \leq j \leq n - 1\}$ takes all residues modulo $2n - 1$, there are at least $n - 1$ additional distinct numbers.