Learn Hard Problems During RL with Reference Guided Fine-tuning

Anonymous ACL submission

Abstract

Reinforcement learning (RL) has become a central mechanism for improving mathematical reasoning in large language models, yet its effectiveness is often limited by reward sparsity: for challenging problems, models fail to sample correct reasoning trajectories, leading to stalled or inefficient learning. We introduce **Reference-Guided Fine-Tuning** (ReGFT), a pre-RL initialization strategy that leverages reference solutions to improve the effectiveness of subsequent RL. ReGFT uses partial reference guidance to elicit model-generated reasoning traces that remain aligned with the model's own inference distribution, enabling RL to operate in regimes where reward signals would otherwise be absent. Across three challenging benchmarks (AIME 2024, AIME 2025, and Beyond-AIME), models initialized with ReGFT consistently achieve faster RL convergence, higher final accuracy, and stronger inference-time scaling compared to both raw and ReFT baseline. These results demonstrate that ReGFT can substantially enhance the effectiveness of reinforcement learning for mathematical reasoning.

1 Introduction

Recent advances demonstrate that reinforcement learning (RL) can substantially improve the reasoning capabilities of large language models (LLMs), as exemplified by OpenAI's o1 and DeepSeek's R1 systems (OpenAI et al., 2024; DeepSeek-AI et al., 2025). A widely adopted instantiation of this paradigm is Reinforcement Learning with Verifiable Rewards (RLVR), which reinforces reasoning trajectories that lead to correct final answers as determined by an automatic rule-based verifier. By relying solely on answer correctness, RLVR avoids human annotation and directly promotes solution-consistent reasoning behaviors.

Despite its effectiveness, RLVR is fundamentally constrained by the competence of the base model. Learning signals arise only when sampled trajectories receive positive rewards, causing training to stall on hard problems where the model fails to generate any correct solutions. This reward sparsity leads to inefficient optimization and limits the extent to which RL can expand the model's reasoning boundary.

A natural solution is to leverage human-written reference chain-of-thought (CoT) solutions commonly available in mathematical datasets. However, directly fine-tuning on such references often yields limited gains due to a mismatch between reference reasoning structures and the model's intrinsic generation patterns, harming generalization (Luong et al., 2024). ReFT (Luong et al., 2024) addresses this issue by fine-tuning on verified model-generated trajectories, but it remains ineffective on problems where no correct trajectories can be sampled.

To overcome this limitation, we propose Reference-Guided Fine-Tuning (ReGFT), which explicitly leverages reference solutions to synthesize training trajectories for problems beyond the model's initial capability. Instead of directly training on reference CoTs, ReGFT uses them as guidance to elicit model-generated reasoning traces that remain solution-consistent while aligning with the model's own reasoning process. The model is fine-tuned on a mixture of these reference-guided trajectories and its self-generated solutions, increasing the pass rate on hard training problems and yielding denser reward signals for subsequent reinforcement learning. Through our experiments, we find that:

- As shown in Figure 1, 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 finetuning 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 ReGFT yields more stable and sustained improvements across benchmarks, while gains from ReFT are primarily limited to low-k regimes and do not scale as effectively.

2 Related Work

Our approach differs from prior work in *when* and *how* external reasoning signals are used. Instead of injecting additional information at inference time during RL (Li et al., 2025a; Chen et al., 2025; Zhang et al., 2025), we focus on improving the base model *before* RL. Moreover, unlike approaches that rely on external teacher models such as R1 (Cetin et al., 2025), we exploit human-written reference solutions already present in the training data. By using these references to guide the model toward generating its own solution-consistent reasoning trajectories, our method enables the model to acquire new problem-solving capabilities prior to RL, thereby increasing the availability of informative reward signals and improving the efficiency of subsequent RLVR training. We leave the detailed discussion of related work of RLVR and external reasoning guidance in Section B.

3 Method

On-policy finetuning with model-generated trajectories (ReFT). We first consider REinforced Fine-Tuning (ReFT) (Luong et al., 2024), which applies supervised fine-tuning 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 addresses the failure mode of ReFT by incorporating reference solutions to synthesize training trajectories for problems that the model cannot solve independently. Specifically, we condition the model on partial reference solutions to guide it toward producing its own reasoning trajectories that lead to correct answers. These reference-guided trajectories preserve the model's reasoning style while providing sufficient structural guidance to overcome previously unsolvable cases. We then finetune the model on a mixture of self-generated correct trajectories (as in ReFT) and reference-guided trajectories. This process substantially increases the model's pass rate on hard problems, yielding a stronger initialization for reinforcement learning.

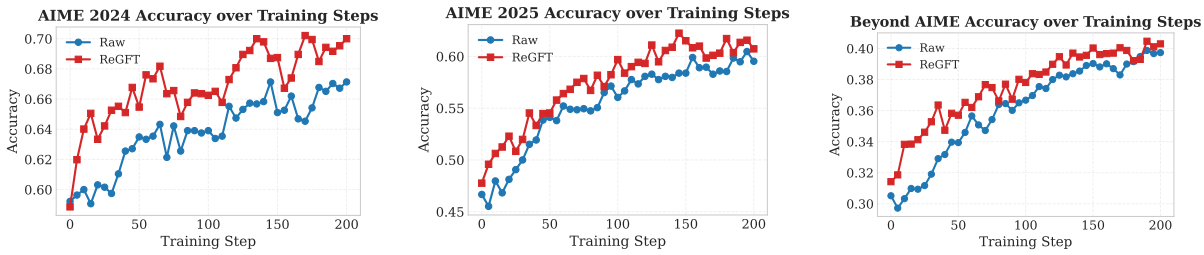
4 Experiments

We conduct our experiments on OmniMath (Gao et al., 2024) dataset and Qwen3-4B-Instruct-2507 model, which is samll, stong but not yet overtrained on math.

4.1 Reference-Guided Finetuning Enhances RL Training

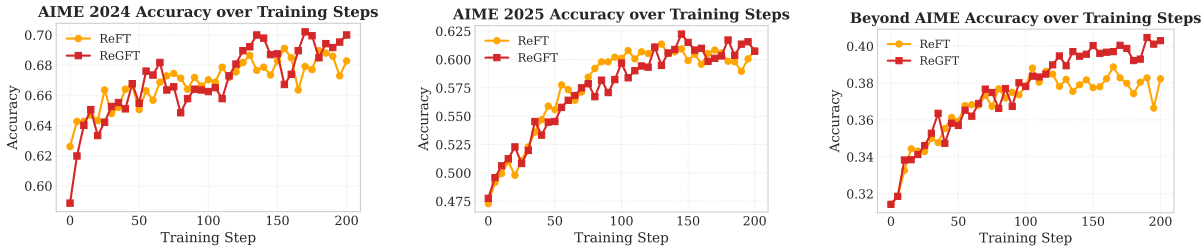
To assess the impact of ReGFT on future RL, we perform reinforcement learning using DAPO (Yu et al., 2025) starting from both the original (raw) checkpoint and the ReGFT fine-tuned checkpoint, we test the perfomance of the SFT checkpoints and put the results in Section D.2. During RL training, we sample 64 responses per prompt to encourage exploration of correct trajectories; the effect of varying the number of samples per prompt is further analyzed in Section D.3.

As shown in Figure 1, models initialized with ReGFT consistently outperform the raw checkpoint throughout RL training across all benchmarks. In addition to achieving higher final accuracy, the ReGFT-initialized model exhibits faster performance gains in early and mid training stages, indicating more efficient optimization. These results



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

Figure 1: 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 2: 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 human-guided demonstrations beyond self-generated trajectories.

suggest that reference-guided fine-tuning effectively mitigates reward sparsity by increasing the prevalence of informative trajectories, thereby accelerating RL convergence and enabling the model to reach a higher performance boundary.

4.2 Impact of Guided Demonstrations

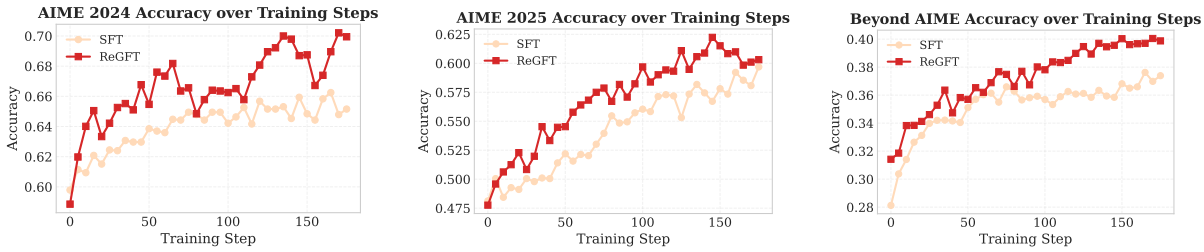
The ReGFT checkpoint is fine-tuned on a mixture of self-generated correct trajectories and reference-guided trajectories, whereas ReFT relies solely on the model's own correct rollouts. To isolate the effect of human-guided demonstrations, we perform DAPO-based reinforcement learning starting from both ReFT and ReGFT checkpoints.

As shown in Figure 2, RL initialized from the ReFT checkpoint also exhibits rapid early improvements, indicating that ReFT effectively densifies reward signals by increasing the likelihood of sampling correct trajectories. However, its final performance consistently lags behind that of ReGFT. This gap suggests that while ReFT improves optimization efficiency, it does not expand the model's underlying reasoning capability. In contrast, human-guided demonstrations in ReGFT introduce novel reasoning patterns that extend the model's competence on previously unsolvable problems, resulting in a higher asymptotic performance after RL.

4.3 Necessity of Model-Derived Reasoning

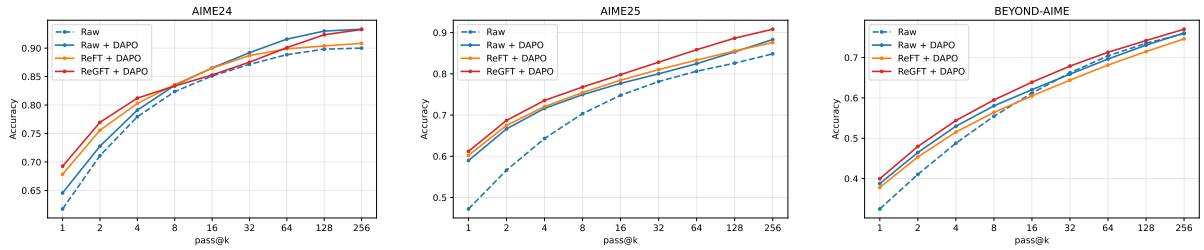
To further examine how human-written reference solutions should be incorporated, we compare ReGFT with an alternative baseline that directly fine-tunes the model on raw human-written reference solutions, without requiring the model to generate its own reasoning. This ablation tests whether exposure to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

As shown in Figure 3, directly fine-tuning on unprocessed human solutions yields substantially weaker performance in both SFT and subsequent RL. This result indicates that the model struggles to internalize reasoning patterns that are not expressed in its own latent style. In contrast, ReGFT—by prompting the model to generate its own reasoning under partial human guidance—produces trajectories that are both correct and distributionally aligned with the model's inference process. These model-derived trajectories are critical for effective learning and serve as a necessary bridge between human demonstrations and reinforcement learning.



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

Figure 3: Accuracy comparison between direct SFT on raw human-written 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 4: Inference-time scaling performance (`pass@k`) of raw and RL-trained checkpoints across benchmarks. Solid lines denote DAPO-trained models, while dashed lines indicate their corresponding pre-RL checkpoints.

4.4 Inference-time Scaling Performance

We further 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 4, reinforcement learning consistently improves inference-time scaling behavior. In contrast to prior observations (Yue et al., 2025), we find that RL yields substantial gains across a wide range of k. We attribute this difference to stronger base models, higher-quality training data, and a larger sampling budget during RL.

Across all benchmarks, **ReGFT + DAPO** achieves the strongest and most consistent scaling performance, dominating ReFT- and raw-based baselines in AIME 2025 and Beyond-AIME, and outperforming them at both early and large k regimes on AIME 2024. While ReFT improves `pass@1` over raw DAPO, these gains do not persist as k increases, which we hypothesize may be due to overfitting introduced during the SFT stage. These results highlight the effectiveness of ReGFT in improving inference-time scaling under increased test-time compute.

5 Conclusion

This work identifies reward sparsity as a bottleneck in reinforcement learning for mathematical reasoning: limited correct trajectories during RL often constrain learning on hard problems. We propose Reference-Guided Fine-Tuning (ReGFT), which transforms human-written solutions into model-aligned reasoning trajectories via guided generation, expanding the set of solvable problems while preserving generalization.

Empirically, ReGFT improves training pass rates, accelerates DAPO convergence, and achieves stronger inference-time scaling than both raw and ReFT-based baselines, with ablations confirming the necessity of both human guidance and model-derived reasoning. Overall, ReGFT mitigates reward sparsity by increasing the availability of verifiable trajectories and complements inference-time scaling for more effective RL. While evaluated under specific settings, these results suggest reference-guided data synthesis as a promising direction for strengthening RL-based reasoning systems.

References

Edoardo Cetin, Tianyu Zhao, and Yujin Tang. 2025. Reinforcement learning teachers of test time scaling. *Preprint*, arXiv:2506.08388.

Justin Chih-Yao Chen, Becky Xiangyu Peng, Prafulla Kumar Choubey, Kung-Hsiang Huang, Jiaxin Zhang, Mohit Bansal, and Chien-Sheng Wu. 2025. Nudging the boundaries of llm reasoning. *Preprint*, arXiv:2509.25666.

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, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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. 2024. Omni-math: A universal olympiad level mathematic benchmark for large language models. *Preprint*, arXiv:2410.07985.

Jian Hu, Mingjie Liu, Ximing Lu, Fang Wu, Zaid Harchaoui, Shizhe Diao, Yejin Choi, Pavlo Molchanov, Jun Yang, Jan Kautz, and Yi Dong. 2025. Brorl: Scaling reinforcement learning via broadened exploration. *Preprint*, arXiv:2510.01180.

Jiazheng Li, Hongzhou Lin, Hong Lu, Kaiyue Wen, Zaiwen Yang, Jiaxuan Gao, Yi Wu, and Jingzhao Zhang. 2025a. Questa: Expanding reasoning capacity in llms via question augmentation. *Preprint*, arXiv:2507.13266.

Ziniu Li, Congliang Chen, Tianyun Yang, Tian Ding, Ruoyu Sun, Ge Zhang, Wenhao Huang, and Zhi-Quan Luo. 2025b. Knapsack rl: Unlocking exploration of llms via optimizing budget allocation. *Preprint*, arXiv:2509.25849.

Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. Reft: Reasoning with reinforced fine-tuning. *Preprint*, arXiv:2401.08967.

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, and 244 others. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Yiyou Sun, Yuhan Cao, Pohao Huang, Haoyue Bai, Hannaneh Hajishirzi, Nouha Dziri, and Dawn Song. 2025. Rl grokking recipe: How does rl unlock and transfer new algorithms in llms? *Preprint*, arXiv:2509.21016.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Zhicheng Yang, Zhijiang Guo, Yinya Huang, Yongxin Wang, Dongchun Xie, Yiwei Wang, Xiaodan Liang, and Jing Tang. 2025b. Depth-breadth synergy in rlvr: Unlocking llm reasoning gains with adaptive exploration. *Preprint*, arXiv:2508.13755.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2025. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Preprint*, arXiv:2504.13837.

Xuechen Zhang, Zijian Huang, Yingcong Li, Chenshun Ni, Jiasi Chen, and Samet Oymak. 2025. Bread: Branched rollouts from expert anchors bridge sft & rl for reasoning. *Preprint*, arXiv:2506.17211.

A Limitations

Our evaluation focuses primarily on mathematical reasoning benchmarks. We selected this domain strategically because it provides a rigorous testbed for studying reward sparsity under verifiable rewards. While we demonstrate efficacy in this challenging setting, it would be interesting to further study how reference-guided fine-tuning transfers to non-mathematical domains.

Additionally, we utilize DAPO (Yu et al., 2025) as our representative RL algorithm to validate our method. It remains to be seen how well our method works with other RL algorithms. That being said, our ReGFT functions as a pre-RL initialization strategy rather than a modification to the RL objective itself, and we anticipate that it is theoretically orthogonal to the choice of RL optimizer.

B Related Work

Reinforcement Learning with Verifiable Rewards. Reinforcement Learning with Verifiable Rewards (RLVR) has become a standard framework for improving large language models on mathematical and symbolic reasoning tasks, due to its

reliance on automatic, rule-based verifiers rather than learned reward models or human annotations. However, multiple studies have shown that the effectiveness of RLVR is fundamentally constrained by the competence of the base model (Yue et al., 2025). When the model fails to sample any correct trajectories for hard problems, reward signals vanish and training can stagnate, resulting in limited or no performance gains (Sun et al., 2025).

A line of recent work addresses this issue by increasing exploration during RL. Methods such as scaling the number of sampled trajectories per problem or dynamically allocating sampling budgets aim to increase the likelihood of obtaining positive rewards on difficult instances (Hu et al., 2025; Yang et al., 2025b; Li et al., 2025b). While these approaches can partially alleviate sparse-reward regimes, they remain ineffective when problem difficulty exceeds the base model's intrinsic reasoning capability.

External Guidance and Reasoning Augmentation. To handle problems that are unsolvable by the base model, several works incorporate external guidance or auxiliary reasoning signals into RL. For instance, Li et al. (2025a) leverage partial solutions from DeepSeek R1 (DeepSeek-AI et al., 2025) as hints to expand the set of solvable problems during RL. Other approaches modify or inject intermediate reasoning traces into prompts to nudge models toward correct solutions (Chen et al., 2025). Relatedly, Cetin et al. (2025) distill reasoning knowledge by training models to rewrite R1-generated solutions via RL, improving downstream RLVR performance.

C Experiments Setup Details

We conduct our experiments on OmniMath (Gao et al., 2024), a benchmark consisting of 4,428 Olympiad-level mathematics problems, each paired with a reference solution. OmniMath is particularly well suited to our study for two reasons. First, the difficulty of the problems induces severe reward sparsity: for many instances, standard model rollouts fail to produce any correct solutions, even with extensive sampling. This property makes OmniMath a natural testbed for evaluating methods designed to improve reinforcement learning under sparse or zero-reward conditions. Second, the availability of verified reference solutions enables controlled investigation of reference-guided strategies, allowing us to assess how auxiliary supervision can be leveraged to improve learning without relying on additional external data.

All experiments are conducted using the Qwen3-4B-Instruct-2507 model (Yang et al., 2025a). We select this model because it strikes a favorable balance between reasoning capability and computational efficiency, making it suitable for large-scale trajectory sampling and reinforcement learning experiments. During generation, we use a temperature of 0.7 and top-p sampling with $p = 0.9$, following standard practice for reasoning tasks. To support long chain-of-thought generation required by Olympiad-level problems, we set both the training and evaluation sequence lengths to 16,384 tokens, which allows the majority of model generations to complete without truncation.

D Additional Experiments

D.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, corresponding to the ReFT setting, and reference-guided sampling, corresponding to ReGFT. For each problem, we sample 64 independent reasoning trajectories under each strategy. In the reference-guided condition, the model is provided with the first 90% of the human-written reference solution as contextual guidance, while the remaining 10%—which typically contains the final answer—is withheld. This design prevents direct answer copying and ensures that correctness is determined by the rule-based verifier based on the model's generated reasoning.

Across the 4,428 problems in OmniMath, normal sampling solves 68.58% of problems, whereas reference-guided sampling solves 70.82%. Importantly, reference-guided sampling enables the model to solve an additional 5.85% of problems that are never solved under normal 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 normal sampling. Qualitative inspection suggests that these cases often arise when the human reference solution follows a reasoning strategy that differs substantially from the model's preferred approach, and the imposed guid-

Table 1: Performance comparison of models fine-tuned under different settings. We report pass@16 accuracy on the training dataset and evaluation results on AIME24, AIME25, and BEYONDAIME.

Model	Training Dataset (%)	Train pass@16 (%)	AIME24 (%)	AIME25 (%)	BeyondAIME (%)
Raw	48.4	63.7	59.2	46.7	30.5
ReFT	49.6	65.9	62.1	46.8	31.3
ReGFT	49.8	67.1	60.0	47.8	31.3

ance can bias the model toward less effective reasoning paths. Overall, reference-guided sampling increases the total proportion of solvable problems, demonstrating its effectiveness in eliciting correct reasoning on previously unsolvable instances.

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.

D.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 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.

D.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 study 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 effectively alleviate the zero-reward regime by densifying reward signals. How-

ever, sampling scale alone is insufficient: even with 64 responses per problem, models trained without reference-guided fine-tuning underperform those initialized with ReGFT. In particular, ReGFT+DAPO achieves the strongest results across all benchmarks, indicating that improving the model's initial competence and scaling exploration are complementary. Together, these results suggest that reference-guided fine-tuning reduces the reliance on excessively large sampling budgets while enabling reinforcement learning to reach a higher performance boundary.