Plasticity vs. Rigidity: The Impact of Low-Rank Adapters on Reasoning on a Micro-Budget

Zohaib Khan and **Omer Tafveez** and **Zoha Hayat Bhatti**
University of Michigan
zohaibkh@umich.edu, omertaf@umich.edu, zohakh@umich.edu

Abstract

Recent advances in mathematical reasoning typically rely on massive scale, yet the question remains: can strong reasoning capabilities be induced in small language models (≤ 1.5B) under extreme constraints? We investigate this by training models on a single A40 GPU (48GB) for under 24 hours using Reinforcement Learning with Verifiable Rewards (RLVR) and Low-Rank Adaptation (LoRA). We find that the success of this “micro-budget" regime depends critically on the interplay between adapter capacity and model initialization. While low-rank adapters ($r = 8$) consistently fail to capture the complex optimization dynamics of reasoning, high-rank adapters ($r = 256$) unlock significant plasticity in standard instruction-tuned models. Our best result achieved an impressive 40.0% Pass@1 on AIME 24 (an 11.1% absolute improvement over baseline) and pushed Pass@16 to 70.0%, demonstrating robust exploration capabilities. However, this plasticity is not universal: while instruction-tuned models utilized the budget to elongate their chain-of-thought and maximize reward, heavily math-aligned models suffered performance collapse, suggesting that noisy, low-budget RL updates can act as destructive interference for models already residing near a task-specific optimum.

1 Introduction

Reasoning tasks—such as mathematical problem solving, logical inference, and symbolic manipulation—remain among the most challenging domains for language models (LLMs). While scaling model size has historically improved reasoning ability (Wei et al., 2023; OpenAI et al., 2024), recent work suggests that sheer parameter count is not the only path forward. Methods such as reinforcement learning with verifiable rewards (RLVR) (Shao et al., 2024; Luo et al., 2025) and supervised fine-tuning on structured reasoning traces (Muennighoff et al., 2025; Ye et al., 2025) have demonstrated that models can acquire advanced reasoning capabilities when guided by structured feedback and verifiable signals. However, the majority of these advances rely on large-scale models trained with extensive compute budgets, leaving open the question: how efficiently can small or mid-sized models be trained to reason well under tight computational constraints?

Recent studies point toward several promising directions for “reasoning on a budget”. First, compact instruction-tuned models have shown latent reasoning potential that can be unlocked with a small number of high-quality examples—the so-called LIMO hypothesis (Ye et al., 2025) that fine-tuning quality matters more than quantity. Second, Muennighoff et al. demonstrated that with as few as 1,000 curated problems and careful test-time control (methods such as “budget forcing”), a 32B model can match or exceed proprietary systems in mathematical reasoning. Third, DeepScaleR extended reinforcement learning to long-context reasoning, showing that a 1.5B model can surpass much larger baselines by progressively increasing reasoning length during RL training (Luo et al., 2025). Together, these findings highlight a growing recognition that data curation, reward structure, and inference compute may be more decisive than raw scale.

Despite this progress, the literature still lacks a systematic study of how parameter-efficient fine-tuning (PEFT) methods interact with RLVR in small-model settings. Most RL works employ full-parameter updates, assuming abundant GPU memory and stable optimization dynamics. In contrast, parameter-efficient strategies such as Low-Rank Adaptation (LoRA) (Hu et al., 2021) offer a practical means to explore the trade-off between trainable capacity and reasoning performance. Recent work has demonstrated that LoRA can be surprisingly expressive even under tight budgets: Schulman and Lab showed that, up to a certain data-to-parameter ratio, LoRA finetuning can match

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 493–501
March 26, 2026 ©2026 Association for Computational Linguistics

or exceed full-parameter finetuning, provided the adapter placement and rank are tuned appropriately. Similarly, the Tina series of models (Wang et al., 2025) achieved strong reasoning performance—reaching over 43% Pass@1 on AIME24—by applying reinforcement learning with LoRA adapters to a 1.5B base model, at a fraction of the cost of full-scale training. These results suggest that low-rank updates are not merely a compute-saving heuristic but can, under the right conditions, unlock reasoning behavior comparable to much larger or fully finetuned models.

However, how these dynamics extend to scenarios with *extreme* computational constraints remains an open question. Our work investigates the limits of reasoning optimization under a strict "micro-budget": **a single A40 GPU (48GB) restricted to 24 hours of training** (equating to approximately 7.2 USD[1]). Under such tight constraints, where models may undergo fewer than 300 update steps, the interaction between the base model's initialization and the LoRA adapter's capacity becomes critical. We explore this across a diverse set of small language models (≤ 1.5B), including general instruction-tuned models, including those specialized for math and intensive reasoning. By varying LoRA ranks ($r \in \{8, 64, 256\}$) within an RLVR framework using Group Relative Policy Optimization (GRPO), we test whether high-rank adapters can induce plasticity in small models even with minimal compute.

Our results reveal a stark dichotomy in how models respond to cheap post-training. We find that generalist instruction-tuned models (and even the RL-tuned DeepScaleR) exhibit high *plasticity*: when equipped with high-rank adapters ($r = 256$), they rapidly learn to elongate their reasoning chains and maximize reward, significantly boosting performance on benchmarks like MATH500 and AIME24. In contrast, heavily specialized models like Qwen2.5-Math-1.5B and Qwen3-0.6B display *rigidity*: the noisy, low-budget RL updates act as destructive interference, causing performance collapse rather than refinement. Ultimately, we propose that the most efficient path to reasoning on a budget is not to refine experts, but to catalyze generalists with high-rank adaptation.

2 Methodology

To investigate the limits of reasoning optimization under strict compute constraints, we adopted a parameter-efficient reinforcement learning framework. All experiments were conducted on a single NVIDIA A40 GPU (48GB VRAM) with a strict 24-hour training cutoff.

2.1 Models

We selected a diverse set of small language models ($\leq$ 1.5B parameters) to evaluate how different initialization strategies affect plasticity under low-budget RL. Our selection spans three categories: models like (1) Qwen2.5-1.5B-Instruct and (2) Llama-3.2-1B-Instruct possessing broad knowledge but lacking specific reasoning optimization; models like (3) Qwen2.5-Math-1.5B and (4) Qwen3-0.6B with extensive pre-training or alignment for mathematics; and an RL-optimized benchmark like (5) DeepScaleR-1.5B-Preview to test whether "cheap" RL can further refine an already optimized policy.

2.2 Datasets

We utilized the Open-RS dataset (Dang and Ngo, 2025), a collection of 7000 reasoning problems containing diverse mathematical and logical queries.

For evaluation, we tracked model validation performance during training with MATH500 and then did a final evaluation with the best rank/checkpoint on AIME24/25 and AMC23 which are competition-level math problems.

2.3 Training Procedure

We implemented our training pipeline using the `verl` framework (Sheng et al., 2025).

Fine-tuning with LoRA. Given the 48GB memory constraint, full-parameter fine-tuning was infeasible. We utilized Low-Rank Adaptation (LoRA) (Hu et al., 2021), which freezes the pre-trained weights W and injects trainable rank decomposition matrices A and B, such that $W' = W + BA$, where $A \in \mathbb{R}^{r \times d}, B \in \mathbb{R}^{d \times r}$ (see Figure 1). We swept the rank $r \in \{8, 64, 256\}$ to test the hypothesis that higher ranks are necessary to capture the complex gradient updates of RLVR.

RLVR with GRPO. We employed Group Relative Policy Optimization (GRPO) (Shao et al.,

[1] As per `vast.ai`—just a bit more than a cup of coffee

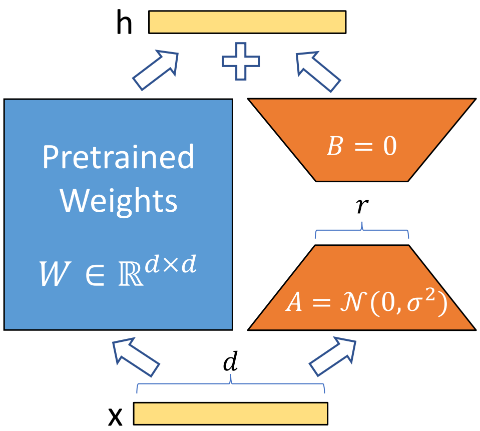

h
Pretrained
Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

Figure 1: Low-Rank Adaptation (LoRA) mechanism. By optimizing only the low-rank matrices A and B, we significantly reduce memory usage while retaining the ability to learn task-specific features.

2024), a policy gradient method designed for efficiency. Unlike PPO, which requires a memory-intensive value network, GRPO estimates the baseline from a group of k sampled outputs for the same prompt.

- **Rollouts:** We used a group size of $k = 8$ to fit within the A40's memory.
- **Token Limit:** Following the configuration in Wang et al., we capped the maximum response length at 3584 tokens to encourage the generation of detailed chain-of-thought reasoning without exceeding context windows. Note that this is *much* smaller than what other works like Luo et al. use.

Reward Structure. We utilized a deterministic, verifiable reward function R_{total}:

$$R_{\text{total}} = 0.2 \cdot R_{\text{format}} + R_{\text{accuracy}}$$

where R_{format} provides a small shaping signal for adhering to the `<think>...</think>` structure, and R_{accuracy} is a binary reward $(+1)$ awarded solely if the final boxed answer matches the ground truth.

3 Experimental Results

We analyze the training dynamics and final performance of the five models to characterize the behavior of RLVR under strict compute constraints.

3.1 Evolution of Training Reward

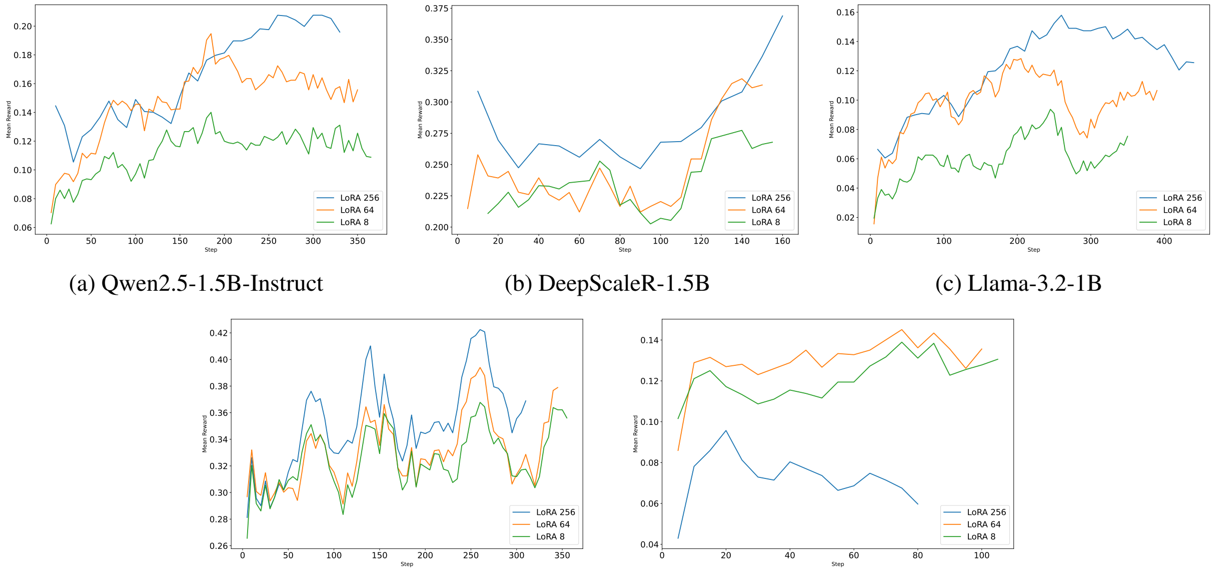
We first examine the ability of different models to optimize the verifiable reward signal (correctness + format) within the 24-hour budget. As shown in Figure 2, a clear distinction emerges based on adapter rank:

- **Generalist Plasticity:** Qwen2.5-1.5B-Instruct and DeepScaleR-1.5B exhibit (almost) monotonic reward growth at high ranks ($r = 256$) compared to lower ones. The high-rank adapters provide sufficient capacity to internalize the RL signal, aligning with what Schulman and Lab found. Even for a model like Llama that isn't suited for reasoning, it too benefits hugely from this cheap training scheme, going from near-zero to double digits in the train reward.
- **Specialist Instability:** Qwen2.5-Math-1.5B shows significant instability at high ranks. Rather than converging, the reward signal fluctuates and degrades, suggesting the updates are conflicting with the model's pre-optimized manifold. An even more concerning result is how Qwen3-0.6B borderline collapses at higher rank updates, an exaggerated case of the previous model.

3.2 Validation Performance (MATH500)

To ensure the reward optimization translates to actual reasoning capability, we tracked Zero-Shot Pass@1 on the MATH500 benchmark throughout training. Figure 3 confirms the "damage vs. help" trade-off:

- **The Learners:** DeepScaleR-1.5B ($r = 256$) and Qwen2.5-1.5B-Instruct ($r = 256$) show strong, consistent gains in validation accuracy. The gains in the former model are much higher than that of the latter, and we attribute this to the former adjusting moreso to the reward function as compared to learning new



(a) Qwen2.5-1.5B-Instruct

(b) DeepScaleR-1.5B

(c) Llama-3.2-1B

(d) Qwen2.5-Math-1.5B

(e) Qwen3-0.6B

Figure 2: **Evolution of Mean Reward.** High-rank adapters ($r = 256$, blue lines) drive consistent learning for generalist models (Top Row), whereas models that underwent less conventional training (Bottom Row) struggle to optimize the reward signal.

reasoning abilities, which is likely happening in the latter.

- **The Collapse:** Qwen2.5-Math-1.5B suffers a sharp performance crash at Rank 256. The RL updates actively harmed its ability to solve math problems compared to its initialization. Another similar pattern is observed with Qwen3-0.6B. It is important to note here that the higher rank updates may lead to a collapse, but the lower rank updates mostly allow the model to stay stagnant.

3.3 Evolution of Response Length

We analyzed the average response length (number of generated tokens) to understand the mechanism behind the performance gains. As seen in Figure 4, plasticity relates well with actual test-time-compute albeit not having a consistent pattern:

- **Expansion:** Llama-3.2-1B and Qwen2.5-1.5B-Instruct demonstrated active exploration by elongating their reasoning chains. Notably, Llama-3.2-1B nearly doubled its response length from ∼700 to over 1,200 tokens. DeepScaleR, while already starting with a long context (∼3,150 tokens) may have learned conciseness in reasoning owing to the limited token budget compared to its previous post-training runs.

- **Contraction:** In contrast, the failing or saturated models (Qwen3-0.6B and Qwen2.5-Math-1.5B) reverted to shorter or unstable responses. Qwen3-0.6B, for instance, saw its response length contract, correlating with its inability to improve validation performance.

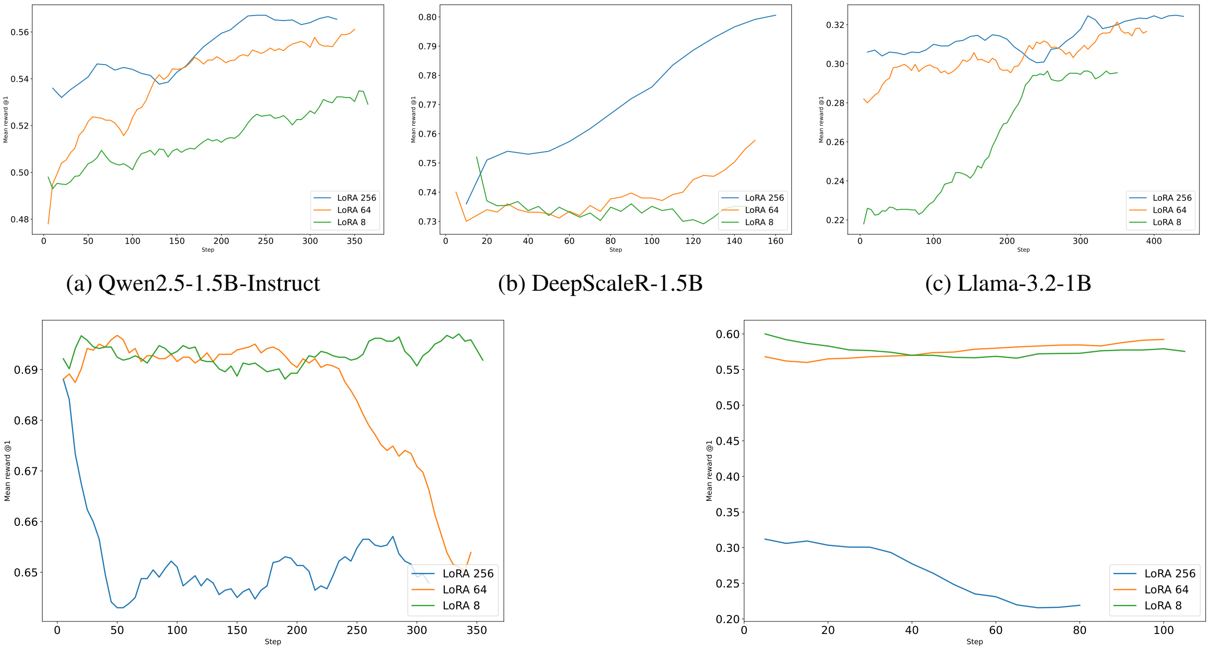
3.4 Evaluation

To assess whether the training gains observed on MATH500 translate to robust generalization, we evaluated the final checkpoints on three held-out competition benchmarks: **AMC 23**, **AIME 24**, and **AIME 25**. We report Zero-Shot Pass@1, as well as Pass@8 and Pass@16 to gauge the models' consistency.

Benchmark Performance. As detailed in Table 1, the impact of low-budget RLVR varies dramatically across model families:

- **DeepScaleR-1.5B** shows the greatest improvement over all benchmarks in all Pass@k estimates. It showed significant gains over its baseline and serves as evidence that LoRA finetuning for a capable instruction-finetuned base can be very effective.

- **Qwen2.5-Math-1.5B and Qwen3-0.6B** show improvements in some benchmarks, albeit there is no consistent pattern. Interestingly we can note how the Pass@8 and Pass@16 are



(a) Qwen2.5-1.5B-Instruct

(b) DeepScaleR-1.5B

(c) Llama-3.2-1B

(d) Qwen2.5-Math-1.5B

(e) Qwen3-0.6B

Figure 3: **Validation Accuracy on MATH500.** Successful models (Top Row) show correlation between training reward and validation score. Qwen-Math (Bottom Center) exhibits "specialist collapse" at high ranks.

more likely to improve than Pass@1, pointing to how the latent reasoning abilities are still improving. We can also recall that the validation scores on MATH500 collapsed for higher ranks, but mostly stayed stagnant for lower ranks which is reflected in these measures. This backs up one of our hypotheses surrounding cheap noisy RL updates disrupting the fragile manifold of heavily pre-optimized models (be it for solving math problems or reasoning, respectively). They would likely require many more training steps.

- **Qwen2.5-1.5B-Instruct and Llama-3.2-1B** did not really show any changes from the baseline when it came to these much harder problems *even though* we saw decent gains in MATH500 scores. Unlike the aforementioned collapse, these models showed minor fluctuations or stagnancy, suggesting that while they possess plasticity, they may require a longer "warm-up" period or more data than the 24-hour budget allowed to bridge the gap to expert reasoning. This may stem from how a benchmark like MATH500 is much easier than soemthing like AIME24 and hence reflects marginal improvements in reasoning ability better.

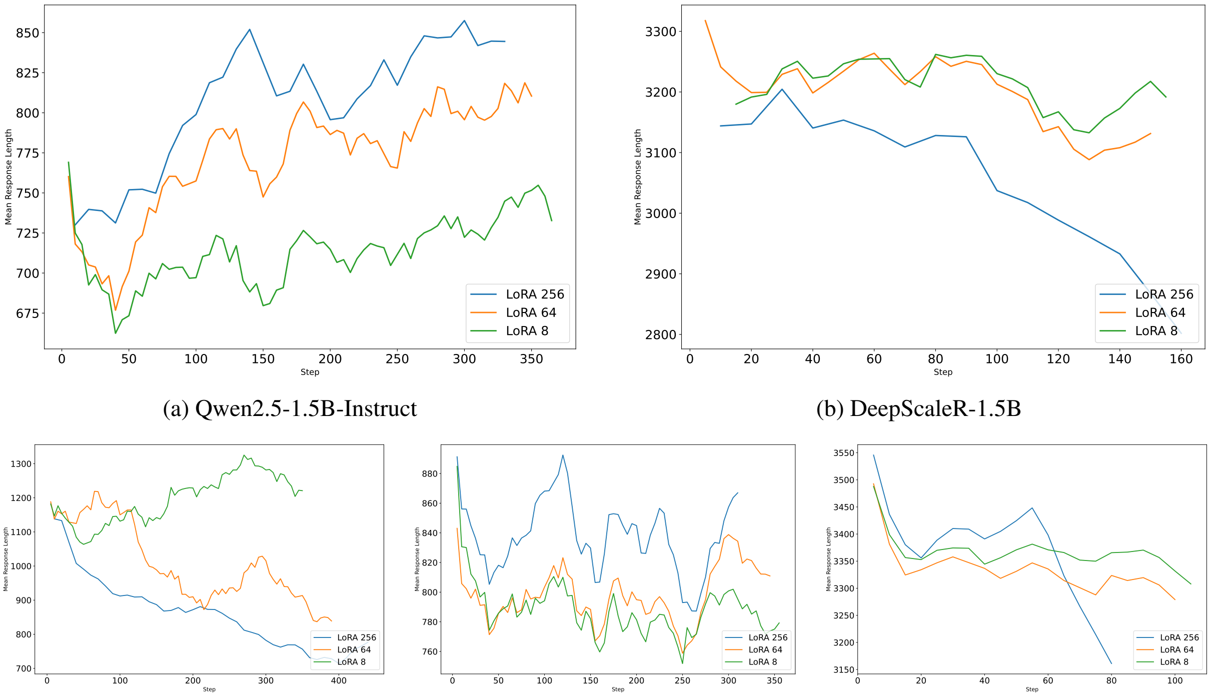
3.5 Entropy Dynamics and Policy Divergence

To understand the mechanism of adaptation under strict compute constraints, we additionally analyze the evolution of the model's policy entropy throughout training. We define the mean token-level entropy $H(\pi)$ for a response sequence y given prompt x as:

$$H(\pi) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{v\in V}\pi(v|x, y_{<t})\log\pi(v|x, y_{<t}) \tag{1}$$

Recent theoretical work suggests that reinforcement learning acts as an entropy regulation mechanism, where the model trades policy entropy (uncertainty) for higher expected reward (Cui et al., 2025). We track the relative change in this metric to quantify how far the fine-tuned policy diverges from its initialization, as show in Table 2.

Rank-Dependent Capacity. We observe that the magnitude of policy divergence is heavily influenced by adapter rank. Across all architectures, models trained with rank $r = 256$ exhibited relative entropy shifts up to 3x larger than those with $r = 8$. This confirms that low-rank constraints mechanically limit the policy's ability to deviate from the pre-trained manifold, effectively anchoring the model to its initialization regardless of the gradient signal.



(a) Qwen2.5-1.5B-Instruct

(b) DeepScaleR-1.5B

(c) Llama-3.2-1B

(d) Qwen2.5-Math-1.5B

(e) Qwen3-0.6B

Figure 4: **Evolution of Response Length.** Plastic models (Top Row) dynamically increased their context usage ("thinking") to maximize reward. Rigid models (Bottom Row) failed to adapt or suffered length collapse.

Divergence vs. Directed Exploration. However, high policy divergence is a necessary but insufficient condition for performance capability. Both DeepScaleR-1.5B and Llama-3.2-1B exhibited significant entropy volatility at high ranks, yet their outcomes diverged. DeepScaleR-1.5B utilized this capacity to explore valid reasoning paths (increasing Pass@16), whereas Llama-3.2-1B, lacking strong reasoning priors, drifted stochastically without converging on high-reward regions.

Optimization Collapse in Aligned Models. For models that are already heavily optimized for the target task (e.g., Qwen2.5-Math-1.5B), high-rank updates acted as destructive interference. Instead of refining the policy, the noisy RL gradients caused a sharp reduction in entropy (mode collapse). The model effectively retreated to low-entropy, safety-seeking behaviors (such as short responses) rather than exploring the solution space, leading to performance degradation.

4 Discussion & Future Work

The Latent Reasoning Gap & Entropy Dissociation. We find that the delta between Pass@1 and Pass@16 acts as a critical feasibility signal for RLVR. A large gap (e.g., DeepScaleR-1.5B-Preview's 40% vs. 70%) indicates "latent" capability that GRPO can effectively bootstrap. This observation complicates recent findings on the "Entropy Mechanism" (Cui et al., 2025), which posit that performance improvements strictly trade off with policy entropy. While valid for capable models, our results with Llama-3.2-1B-Instruct challenge this universality: the model exhibited significant entropy reduction (collapse) without corresponding performance gains. This suggests that for models with weak reasoning priors, entropy reduction may signal a regression into simple convergent behaviors rather than optimization, dissociating the link between certainty and correctness as observed in contemporary works on stronger reasoning models.

The "Warm-Start" Hypothesis. Our results reinforce the "LIMA hypothesis" (Zhou et al., 2023) within an RL context: RLVR acts primarily as an alignment mechanism to expose latent knowledge, rather than a pedagogical tool to teach new theorems. Llama-3.2-1B-Instruct's failure suggests a "Cold Start" problem where random exploration cannot bridge the gap to the first non-zero reward. We posit that RLVR is most cost-effective when applied to "warm" models—those already seeded with reasoning behaviors via pre-training or SFT—allowing the optimizer to focus on uti-

Table 1: Comparison of Baseline vs. Final (LoRA) performance. **Bold** values indicate improvement over the baseline.

Model	**Config**	**AIME 24 (%)**			**AIME 25 (%)**			**AMC 23 (%)**		
		@1	@8	@16	@1	@8	@16	@1	@8	@16
DeepScaleR-1.5B-Preview	Baseline	28.9	53.6	62.5	17.7	35.8	45.2	58.8	87.5	92.2
	Final	**40.0**	**68.0**	**70.0**	**28.1**	**50.2**	**56.7**	**79.2**	**96.0**	**97.5**
Qwen2.5-Instruct-1.5B	Baseline	2.5	10.0	16.7	0.6	4.2	6.7	24.2	57.7	65.0
	Final	2.5	10.0	16.6	0.6	4.2	6.7	24.2	57.6	65.0
Llama-3.2-1B	Baseline	1.0	7.6	13.3	0.2	1.7	3.3	10.6	34.8	50.0
	Final	1.0	7.6	13.3	0.2	1.7	3.3	10.6	34.7	50.0
Qwen2.5-Math-1.5B	Baseline	3.8	10.6	13.3	2.5	14.0	20.0	22.2	59.7	72.5
	Final	**4.8**	**15.1**	**20.0**	**2.9**	**14.4**	20.0	18.9	**62.1**	**77.5**
Qwen3-0.6B	Baseline	9.4	28.4	36.7	14.0	31.3	36.7	46.9	78.2	85.0
	Final	8.5	**28.5**	36.7	**14.6**	**31.9**	36.7	**48.3**	77.0	82.5

Table 2: Relative Change in Policy Entropy (%) by Rank. High-rank adapters ($r = 256$) drive massive entropy shifts compared to $r = 8$.

Model	**Relative Entropy Change (ΔH_{rel})**		
	Rank 8	**Rank 64**	**Rank 256**
DeepScaleR-1.5B	-6.7	-2.7	**-27.0**
Llama-3.2-1B	-13.7	-67.9	**-92.5**
Qwen2.5-1.5B-Instruct	-5.5	-62.1	**-60.1**
Qwen2.5-Math-1.5B	+37.7	-3.8	**-31.7**
Qwen3-0.6B	-2.0	-11.0	**-61.5**

lizing the latent space rather than constructing it. Future work should investigate brief SFT phases as a warm-up for reasoning“Reasoning Warm-up" to prime off-the-shelf models before RLVR.

Algorithmic Constraints & Policy Deviations. The "rigidity" observed in Qwen2.5-Math-1.5B suggests that the conservative trust-region constraints of standard PPO/GRPO may be counterproductive when fine-tuning specialists with noisy, micro-budget updates. We hypothesize that algorithms which relax the aggressive clipping of the policy gradient—such as Dr. GRPO (Liu et al., 2025) or DAPO/CLIP-Higher (Yu et al., 2025)—could allow the policy to deviate sufficiently from its local optimum to discover more robust reasoning paths. By permitting larger distinct updates, these methods might prevent the mode collapse we observed in specialists, provided the reward signal remains verifiable.

Scaling Laws of High-Rank Adaptation. Finally, our success with high-rank LoRA ($r = 256$) on DeepScaleR-1.5B-Preview suggests a scalable paradigm for reasoning alignment: treating high-rank adapters as a cost-effective alternative to full-parameter fine-tuning. If RLVR is largely about surface-level alignment of latent reasoning (as seen in DeepSeek-R1), then massive full-parameter updates may be redundant. Future work should extend this study to larger scales, comparing high-rank LoRA against full-finetuning over extended epochs to determine if the “plasticity" provided by $r = 256$ is sufficient to replicate the gains of full-scale training at a fraction of the GPU-hour cost.

5 Conclusion

Our investigation into reasoning alignment under strict compute constraints reveals that high-performance mathematical reasoning is attainable on a “micro-budget", provided the alignment strategy matches the model’s initialization. We demonstrate that plasticity is the governing resource: generalist models like Qwen2.5-1.5B-Instruct and DeepScaleR-1.5B possess the latent capacity to actively explore and internalize reasoning behaviors when empowered by high-rank adapters ($r = 256$), enabling DeepScaleR to achieve a state-of-the-art 40.0% Pass@1 on AIME 24. Conversely, the rigidity of heavily optimized models like Qwen2.5-Math renders them vulnerable to the noisy updates of low-budget RL, leading to performance collapse. Ultimately, we propose that the most efficient path to democratizing reasoning is not to incrementally refine experts, but to catalyze generalists—using high-rank adaptation to unlock the latent reasoning capabilities already present in their pre-trained manifolds.

Limitations

Our study was designed to probe the feasibility of reasoning alignment under extreme constraints, and as such, several limitations apply to our findings:

- **Model Scale:** Due to the single-GPU memory constraint (48GB), our investigation was restricted to small language models ($\leq$ 1.5B parameters). It remains verifying whether the “plasticity vs. rigidity" trade-off we observed holds for larger architectures (e.g. 7B,8B,32B models), which often possess more robust internal representations and might be more resilient to noisy LoRA updates.

- **Hyperparameter Scope:** The strict 24-hour compute budget precluded a comprehensive grid search. We utilized fixed values for critical hyperparameters such as learning rate and LoRA alpha across all runs. It is possible that the “collapse" observed in some models could be mitigated with a more conservative learning rate or a tuned alpha/rank ratio, rather than being an intrinsic failure of the method itself.

- **Training Duration:** We limited training to a maximum of 24 hours ($\sim$300 update steps). While sufficient to observe divergence in plasticity, this window may be too short for "slow-learning" generalist models to fully converge. Longer training horizons might reveal that models like Llama-3.2-1B eventually overcome the "cold start" problem given enough exploration time.

- **Single-Seed Stochasticity:** Finally, due to resource limitations, each experimental configuration was conducted with a single random seed. Given the inherent high variance of reinforcement learning—particularly with the GRPO estimator—our results may be influenced by initialization noise. Future work with greater resources should employ multi-seed averaging to report confidence intervals and ensure statistical significance.

References

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *Preprint*, arXiv:2505.22617.

Quy-Anh Dang and Chris Ngo. 2025. Reinforcement learning for reasoning in small llms: What works and what doesn’t. *Preprint*, arXiv:2503.16219.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. *Preprint*, arXiv:2503.20783.

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

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *Preprint*, arXiv:2501.19393.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

John Schulman and Thinking Machines Lab. 2025. Lora without regret. *Thinking Machines Lab: Connectionism*. Https://thinkingmachines.ai/blog/lora/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys ’25, page 1279–1297. ACM.

Shangshang Wang, Julian Asilis, Ömer Faruk Akgül, Enes Burak Bilgin, Ollie Liu, and Willie Neiswanger. 2025. Tina: Tiny reasoning models via lora. *Preprint*, arXiv:2504.15777.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *Preprint*, arXiv:2502.03387.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.