Reasoning Elicitation is Scale Dependent

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Frontier labs employ computationally intensive reinforcement learning pipelines to enhance base models' performance on mathematical and logical reasoning tasks. However, the extent to which these improvements represent newly learned capabilities versus activation of latent abilities remains unclear. We address this question by demonstrating that both sample-efficient and parameter-efficient training methods successfully elicit reasoning capabilities, but only from sufficiently large base models. We find that finetuning with as few as 29 DeepSeek R1 reasoning traces is sufficient to recover substantial reasoning performance in the 32B parameter version of `Qwen2.5-Instruct`, while the 1.5B and 7B versions see small or negative gains despite improvements in validation loss. Furthermore, we show that a rank-1 LoRA with just 0.03% as many trainable parameters achieves substantial performance improvements on 32B models, demonstrating that these gains are not simply a consequence of larger models having greater training capacity. Our findings reveal that reasoning capabilities can be efficiently elicited from large models through minimal data and parameter updates, while smaller models fail to benefit from these same efficient methods. This stark difference in learning efficiency suggests that general reasoning capabilities may already be partially latent in larger models, with important implications for our understanding of how capabilities emerge with scale.

1 Introduction

Modern large language models are trained specifically to do chain-of-thought (CoT) reasoning [7, 11], usually through computationally expensive reinforcement learning pipelines, or through supervised finetuning using a large number of samples. These models perform substantially better on mathematical and logical reasoning benchmarks compared to models that don't have this additional training. DeepSeek has previously released a suite of distilled reasoning models based on `Qwen2.5-Instruct`, which were trained using 800k examples of reasoning traces from their 671B parameter R1 model[3].

Recent work has shown that a significant amount of this performance can be efficiently trained into base models using small, curated datasets of reasoning traces from CoT reasoning models. Sky-T1 [5] uses a dataset of 17,000 reasoning traces to achieve performance comparable to OpenAI's `o1` model, while `s1` [6] achieves even higher performance using just 1000 examples from `DeepSeek R1`. Additional work has shown that representations useful for reasoning are already present in base models, suggesting that reasoning capabilities may be elicited rather than taught [10].

In this work, we investigate the limits of sample efficient training across various model scales by finetuning 1.5B, 7B, 14B, and 32B parameter models in the `Qwen2.5-Instruct` model family [8]. We train using subsets of the s1k-1.1 dataset, from as few as 29 examples up to the full 1000-example dataset (excluding a small subset for validation). We evaluate our trained models on challenging reasoning benchmarks including AIME 2024, AIME 2025, and GPQA Diamond [1, 9].

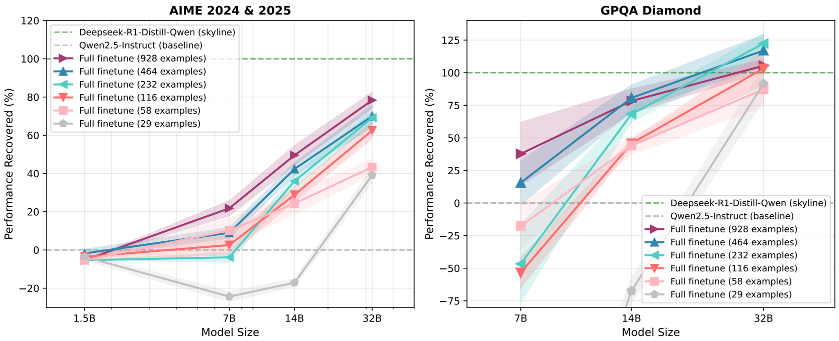

Figure 1: Performance recovery for full-parameter finetunes of models of various sizes using datasets of various sizes on AIME (left) and GPQA Diamond (right). *Recovered %* is $(\text{Score}_{\text{trained}} - \text{Score}_{\text{base}})/(\text{Score}_{\text{skyline}} - \text{Score}_{\text{base}}) \times 100\%$. No performance difference between base and DeepSeek models was observed at 1.5B for GPQA, so data for this model size is omitted.

We uncover a stark pattern in learning efficiency: larger base models show substantial performance gains across all tested dataset sizes, while smaller models struggle to improve even with the maximum number of training examples. The 32B model finetuned with only 29 examples sampled from s1k-1.1 bridges as much as 80% of the GPQA performance gap between the base model and the skyline `DeepSeek-R1-Distill-Qwen-32B` model, while the same training pipeline and dataset applied to the 14B parameter model actually *worsens* reasoning performance. We additionally find that finetuning with larger datasets *can* improve reasoning performance in the 14B parameter model, but no tested dataset size improves reasoning performance in the 1.5B parameter model.

This pattern is not simply a result of smaller models overfitting to the training data. We train with early stopping based on validation loss, and show that all finetuned models achieve significant validation loss improvements. This shows that all models learn patterns which generalize to held-out data, but only larger models learn patterns that translate to improved reasoning capabilities. Smaller models instead appear to learn narrow statistical regularities, improving validation metrics without generalizing to enhanced reasoning abilities.

Finally, we demonstrate that the superior training efficiency of larger models is not merely a consequence of having a greater number of parameters which can be optimized. We show that a rank-1 LoRA [4], representing only 0.03% of the full model's trainable parameters, is enough to recover 70-80% of the performance on the fully finetuned 32B model when trained on the full s1k-1.1 dataset.

Our contributions are:

We show that the effectiveness of sample-efficient reasoning training drops off substantially as model size decreases. 32B models benefit from as few as 29 training examples while $\leq$ 7B models show small or negative improvements even when trained with $> 30\times$ as many examples.

We demonstrate that smaller models learn dataset patterns without generalizing to reasoning capabilities. Small models achieve comparable decreases in validation loss, but this doesn't translate to performance on reasoning evals.

We show that parameter-efficient finetuning methods are enough to capture the majority of reasoning performance lift in larger models. A rank-1 LoRA recovers up to 80% of the performance of the 32B model finetuned on the same dataset, demonstrating that training efficiency gains in larger models are not merely a consequence of having more trainable parameters.

2 Methodology

We extract subsets of the s1k-1.1 dataset using random sampling without replacement. We first extract a 72-example held-out evaluation set, which stays constant among all experiments. The remaining 928 training examples are subsampled into 6 different datasets with varying sizes: 928, 464, 232, 116, 58, and 29 examples with smaller datasets being strict subsets of larger ones. We perform full-parameter finetunes using each of these datasets on each of the 1.5B, 7B, 14B, and 32B

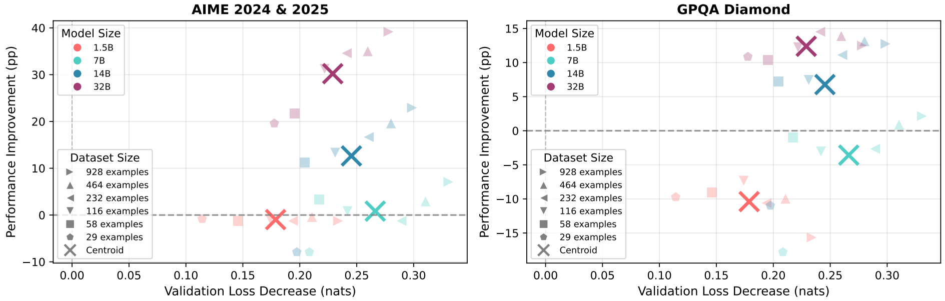

Figure 2: Scatter plot showing the relationship between validation loss decrease and reasoning performance for different model sizes. × symbols represent centroids of all training runs (29 to 928 training examples) for each model size. All models experience improvements in validation loss, but this only translates to improved reasoning performance in larger models.

versions of `Qwen2.5-Instruct`, implementing early stopping based on validation loss. Additionally, we initialize a rank-1 LoRA for each of the 4 base model sizes to adapt every layer's attention and MLP projection matrices (q, k, v, o, up, gate, down), training on the full 1000-examples of s1k-1.1 without a held-out validation set. We evaluate all models using the `aime2024`, `aime2025`, and `gpqa-diamond-cot-zeroshot` tasks using `lm-evaluation-harness` [2], computing means and standard error over four total runs for each model/task combination. Full training details are provided in Appendix A.1.

3 Discussion

Larger models gain more performance from smaller datasets Our data shows a clear relationship between model scale and performance recovery from sample-efficient training. Our finetuned 32B model achieves significant performance improvements on both AIME and GPQA when trained with only 29 reasoning examples, while the 14B model sees performance degradation on both benchmarks when trained with the same dataset (Figure 1). While a larger dataset with 928 examples is effective at extracting performance from the 14B model, the recovered performance is less than that of the 32B model, and the 1.5B model sees no gain from any of the tested datasets. Exact performance values can be found in Table 1.

This pattern suggests a scale-dependent threshold for sample-efficient training. As model size increases, the minimum dataset size required for positive performance gains appears to decrease: larger models need fewer training samples to learn reasoning.

Smaller models learn narrow patterns We investigate whether the failure of smaller models to gain reasoning performance is merely a consequence of overfitting to training data. In Figure 2, we plot validation loss improvement versus reasoning performance gain for each model size. We find that the 14B models actually experience greater validation loss improvements compared to 32B models, and 7B models see the greatest validation loss improvements out of all 4 model sizes, on each dataset. The 1.5B models see the smallest improvement in validation loss, but this improvement is still substantially positive. This data indicates that all model sizes are learning patterns which lead to improved prediction of held-out validation data.

Counterintuitively, we find that the greater relative improvement in validation loss for 14B and 7B models does not translate to greater relative improvement on reasoning benchmarks. As noted previously, the 32B models see significantly greater performance improvements compared to 14B models, which themselves see greater performance improvements compared to 7B models. Despite experiencing the greatest improvements in validation loss, 7B models see little to no improvement on reasoning tasks.

These trends suggest that the nature of the patterns learned may differ fundamentally between large and small models: smaller models learn statistical regularities which improve performance on validation data, while larger models learn more general patterns which also improve performance on generative reasoning tasks. This interpretation aligns with the hypothesis that reasoning capabilities

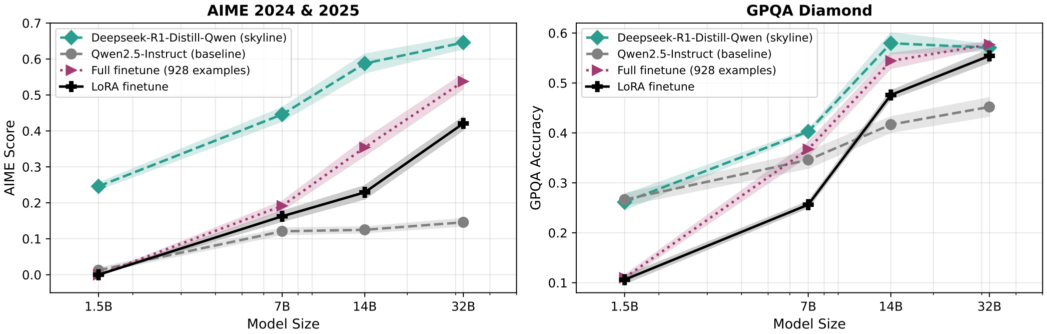

Figure 3: Performance scores for baseline model, skyline model, full-parameter finetune, and LoRA.

may be partially latent in larger models, only requiring light training to elicit, while smaller models must learn these capabilities from scratch.

LoRA performance recovery We show that the reasoning performance gains observed in large models can be mostly achieved through minimal parameter modifications. A rank-1 LoRA trained on all layers and projection matrices of the 32B parameter model encodes only 0.03% as many trainable parameters, but still achieves 70-80% of the performance improvement which full-parameter finetuning achieves using the largest dataset (Figure 3).

This finding has two important implications. First, it shows that the superior performance gain exhibited by larger models is not merely a result of having more trainable parameters: they see similar gains when we restrict parameter updates significantly. Second, the fact that rank-1 updates are sufficient to elicit reasoning performance indicates that the new representations required to encode reasoning capabilities are themselves minimal, further suggesting that important representations or mechanisms required for reasoning capabilities are already latent in larger base models.

4 Conclusion

We investigate the scale-dependent nature of sample-efficient reasoning elicitation across the Qwen2.5-Instruct model family. We find that reasoning performance can be elicited from a 32B parameter model using as few as 29 training examples, while smaller models require more data and gain less performance. This difference cannot be trivially explained by overfitting: all trained models achieve improved validation loss, with 7B and 14B models experiencing greater improvements in validation loss compared to the 32B model. This disconnect is consistent with larger models possessing latent structures useful for reasoning, while smaller models may lack these structures and instead learn to express narrow, less general patterns. We additionally show that a rank-1 LoRA is enough to capture most of the performance gain in our finetuned 32B model, highlighting how reasoning capabilities require only minimal parameter changes to be expressed in larger models. Our findings underscore how reasoning behavior elicitation becomes easier as a function of model scale, with important implications for understanding how large base models may have substantial latent capabilities.

Limitations Several limitations should be considered when interpreting our results. First, we only evaluate reasoning performance on two different benchmarks. While we think these benchmarks provide a large amount of signal into the general reasoning capabilities of our studied models, additional benchmarks may reveal differences in how reasoning capabilities in other domains are elicited. We also only evaluate one model family, Qwen2.5-Instruct, using traces from only the s1k-1.1 dataset. Other models or datasets may yield different patterns. Additionally, we were not able to explore the effects of training with a larger number of reasoning examples due to computational costs–this analysis would likely provide a better understanding of how small models can learn reasoning capabilities. Finally, while our results suggest that reasoning capabilities may be latent in large models, this study lacks mechanistic evidence to prove this hypothesis. Future work should address these limitations to better understand the nature of reasoning capability elicitation.

References

[1] Art of Problem Solving. Aime problems and solutions (index). https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions, 2025.

[2] Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. Lessons from the trenches on reproducible evaluation of language models. 2024. URL https://arxiv.org/abs/2405.14782.

[3] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. 2025. doi: 10.48550/arXiv.2501.12948.

[4] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. 2021. doi: 10.48550/arXiv.2106.09685.

[5] Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhamaneshi, Shishir G. Patil, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Llms can easily learn to reason from demonstrations: Structure, not content, is what matters! 2025. doi: 10.48550/arXiv.2502.07374. Sky-T1.

[6] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Fei-Fei Li, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. 2025. doi: 10.48550/arXiv.2501.19393.

[7] OpenAI. Openai o1 system card. 2024. URL https://arxiv.org/abs/2412.16720.

[8] Qwen Team. Qwen2.5 technical report. 2024. doi: 10.48550/arXiv.2412.15115.

[9] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. 2023. doi: 10.48550/arXiv.2311.12022.

[10] Jake Ward, Chuqiao Lin, Constantin Venhoff, and Neel Nanda. Reasoning-finetuning repurposes latent representations in base models. 2025. doi: 10.48550/arXiv.2507.12638.

[11] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. 2022. doi: 10.48550/arXiv.2201.11903. NeurIPS 2022.

A Appendix

A.1 Training and Evaluation

Full-parameter Finetuning Details For all experiments, we use the AdamW optimizer with a learning rate of 1e-5 and a cosine schedule, and weight decay of 1e-4. We choose the last checkpoint which is within 0.1 nats of the best validation loss for each run. We find experimentally that this 0.1 nat allowance results in improved performance on reasoning benchmarks when smaller datasets are used. We allow all experiments to train for up to 300 steps with a batch size of 16.

LoRA Finetuning Details The LoRA is trained using the same pipeline and hyperparameters as the full-parameter finetunes, but with a learning rate of 1e-3.

Evaluation Details For all experiments, we use a maximum completion length of 29000 tokens and sample with temperature = 0.6.

Reference models We define baseline as the same-size Qwen2.5-Instruct model and skyline as the same-size DeepSeek-R1-Distill-Qwen model used in our plots; all Recovered % values use these references.

A.2 Additional Data

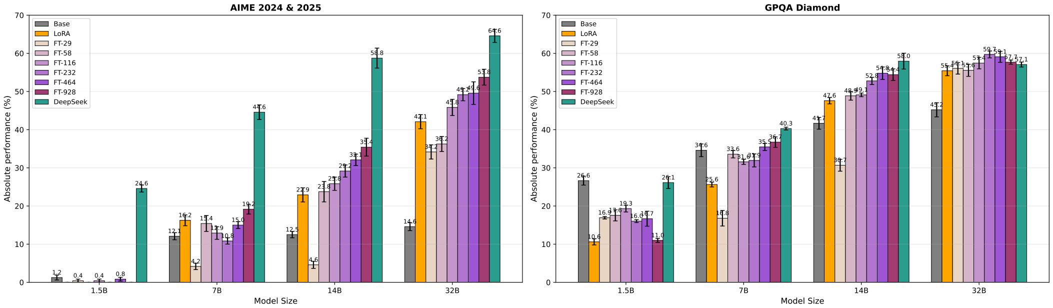

Figure 4: All measured absolute performance values.

Table 1: Performance on AIME (2024 & 2025) and GPQA Diamond across model sizes and training configurations, including mean +/- std err over 4 seeds.

Model Size	Configuration	AIME (%)	GPQA (%)
1.5B	Qwen2.5-Instruct (base)	1.25 ± 0.64	26.64 ± 1.17
	DeepSeek-R1-Distill	24.58 ± 0.93	26.14 ± 1.55
	LoRA (1000 examples)	0.00 ± 0.00	10.61 ± 0.80
	Full finetune (29 examples)	0.42 ± 0.42	16.92 ± 0.33
	Full finetune (58 examples)	0.00 ± 0.00	17.55 ± 1.45
	Full finetune (116 examples)	0.42 ± 0.42	19.32 ± 0.88
	Full finetune (232 examples)	0.00 ± 0.00	16.04 ± 0.38
	Full finetune (464 examples)	0.83 ± 0.48	16.67 ± 1.97
	Full finetune (928 examples)	0.00 ± 0.00	10.98 ± 0.52
7B	Qwen2.5-Instruct (base)	12.08 ± 0.93	34.60 ± 1.67
	DeepSeek-R1-Distill	44.58 ± 1.91	40.28 ± 0.38
	LoRA (1000 examples)	16.25 ± 1.42	25.63 ± 0.63
	Full finetune (29 examples)	4.17 ± 0.83	16.79 ± 2.03
	Full finetune (58 examples)	15.42 ± 2.08	33.59 ± 0.98
	Full finetune (116 examples)	12.92 ± 1.65	31.57 ± 0.67
	Full finetune (232 examples)	10.83 ± 0.83	31.94 ± 1.70
	Full finetune (464 examples)	15.00 ± 0.90	35.48 ± 0.95
	Full finetune (928 examples)	19.17 ± 1.27	36.74 ± 1.37
14B	Qwen2.5-Instruct (base)	12.50 ± 0.83	41.67 ± 1.55
	DeepSeek-R1-Distill	58.75 ± 2.62	57.95 ± 2.08
	LoRA (1000 examples)	22.92 ± 1.85	47.60 ± 0.83
	Full finetune (29 examples)	4.58 ± 0.93	30.68 ± 1.56
	Full finetune (58 examples)	23.75 ± 2.67	48.86 ± 1.14
	Full finetune (116 examples)	25.83 ± 1.73	49.12 ± 0.48
	Full finetune (232 examples)	29.17 ± 1.60	52.78 ± 0.96
	Full finetune (464 examples)	32.08 ± 1.50	54.80 ± 1.63
	Full finetune (928 examples)	35.42 ± 2.34	54.42 ± 1.51
32B	Qwen2.5-Instruct (base)	14.58 ± 1.05	45.20 ± 1.85
	DeepSeek-R1-Distill	64.58 ± 1.72	57.07 ± 0.65
	LoRA (1000 examples)	42.08 ± 1.85	55.43 ± 1.29
	Full finetune (29 examples)	34.17 ± 1.86	56.06 ± 1.50
	Full finetune (58 examples)	36.25 ± 1.97	55.56 ± 1.58
	Full finetune (116 examples)	45.83 ± 2.12	57.45 ± 1.51
	Full finetune (232 examples)	49.17 ± 1.60	59.72 ± 0.91
	Full finetune (464 examples)	49.58 ± 2.96	59.09 ± 1.44
	Full finetune (928 examples)	53.75 ± 2.08	57.70 ± 0.56