

Learning to Reason: Training LLMs with GPT-OSS or DeepSeek R1 Reasoning Traces

Shaltiel Shmidman^{1,†}, Asher Fredman^{2,‡}, Oleg Sudakov^{2,‡}, Meriem Bendris^{2,‡}

¹DICTA ²NVIDIA

[†]shaltiel@dicta.org.il

[‡]{afredman, osudakov, mbendris}@nvidia.com

Abstract

Test-time scaling, which leverages additional computation during inference to improve model accuracy, has enabled a new class of Large Language Models (LLMs) that are able to reason through complex problems by understanding the goal, turning this goal into a plan, working through intermediate steps, and checking their own work before answering. Frontier large language models with reasoning capabilities, such as DeepSeek-R1 and OpenAI’s gpt-oss, follow the same procedure when solving complex problems by generating intermediate reasoning traces before giving the final answer. Today, these models are being increasingly used to generate reasoning traces that serve as high-quality supervised data for post-training of small and medium-sized language models to teach reasoning capabilities without requiring expensive human curation. In this work, we compare the performance of medium-sized LLMs on Math problems after post-training on two kinds of reasoning traces. We compare the impact of reasoning traces generated by DeepSeek-R1 and gpt-oss LLMs in terms of accuracy and inference efficiency.

1 Introduction

One way to improve the reasoning capabilities in LLMs is using test-time scaling, which leverages additional compute during inference to improve accuracy. This was originally accomplished via *Chain-Of-Thought Prompting* (Wei et al., 2023), where the model was prompted to first reason through a problem before arriving at the solution. Then, reasoning models were developed such as ChatGPT-01 (OpenAI, 2024) and Gemini-2.5 (Comanici et al., 2025), which “think out loud” by generating a sequence of reasoning tokens (reasoning trace) before providing an answer without explicitly exposing these traces to the user. The release of open source large-scale reasoning models such as DeepSeek-R1 (DeepSeek-AI et al., 2025) and the OpenAI gpt-oss series

(OpenAI, 2025a) has enabled the community to access the reasoning traces of these models.

With the rise of frontier reasoning LLMs created with novel post-training techniques (DeepSeek-AI et al., 2025; Zheng et al., 2025; Lambert et al., 2025), including verifier-guided RL, rule-based format/accuracy rewards, rejection sampling, and distillation, many model builders can now leverage synthetic data produced by large-scale reasoning models. For example, to enable the generation of reasoning traces by DeepSeek R1 (DeepSeek-AI et al., 2025), a combination of human-annotated cold-start data, reasoning-oriented RL, custom loss function and specialized training techniques was used. This data serves as high-quality supervised data to fine-tune small/medium-sized models for reasoning.

However, synthetic reasoning traces produced by these models vary significantly in both style and verbosity. This motivates the central research question: which reasoning style is more effective to distill into small/medium-sized models.

In this work, we compare reasoning styles of two open source LLMs: DeepSeek-R1 and gpt-oss. We measure their impact when fine-tuning 12B-parameter-sized base models on reasoning data, in terms of accuracy as well as inference efficiency. We focus on Math problems as they require complex problem-solving skills and are heavily affected by Test-Time Scaling.

2 Math Reasoning Dataset

In our experiments, 300,000 math conversations were sampled from the Nemotron-Post-Training-Dataset-v1 (Bercovich et al., 2025; Nathawani et al., 2025). This dataset has a “math” split which provides math problems of various difficulty with their ground truth answers, as well as the corresponding generated responses from DeepSeek-R1-2508 with full reasoning traces. To compare the

reasoning styles between two models, we used gpt-oss-120b model to sample a response for each of the 300,000 questions¹, storing the full response with the reasoning traces.

Then we filtered out samples in which the answers generated by either DeepSeek-R1-2508 or gpt-oss-120b didn't match the ground truth. To automate the filtering, Qwen3-30B-A3B-Thinking-2507 (Team, 2025) was used as the judge model. The model was provided the ground truth answer and the generated answer, and prompted to determine whether the answers were the same. The full system prompt used can be found in Appendix A. Finally, only samples where both models produced the correct answers were selected. The final dataset consists of 242,000 samples (each sample consists of a math problem, ground truth answer, and full reasoning trace from both DeepSeek-R1 and gpt-oss).

Analyzing the reasoning trace styles, we observed that DeepSeek-R1 generated on average $4.4 \times$ as many tokens compared to gpt-oss-120b:

- **DeepSeek-R1** response length was on average $\approx 15,500$ tokens.
- **gpt-oss-120b** response length was on average $\approx 3,500$ tokens.

3 Experimental Setup

In our experiment, we trained two LLMs on the two distinct reasoning styles, and evaluated their performance on various Math benchmarks.

3.1 Base Models

We conducted out experiments with two 12B parameter models:

- **NVIDIA-Nemotron-Nano-12B-v2-Base** (NVIDIA, 2025) - a high-performing 12B base model pretrained from scratch by NVIDIA, which was infused with reasoning traces from DeepSeek-R1 during the mid-training.
- **Mistral-Nemo-Base-2407** (Team, 2024) - a high-performing 12B model. It was chosen for comparison due to its release prior to reasoning models, and therefore not having any reasoning traces in the pretraining dataset.

¹We did not use any system prompt when sampling, so the model uses the default reasoning effort.

3.2 Infrastructure

We ran our experiments on a cluster of H200 141GB GPUs on NVIDIA DGX Cloud Leptron (Anand and Soman, 2025). All training was done using NVIDIA NeMo Framework (Harper et al.), a scalable generative AI framework built for researchers and developers working on large language models, optimized for large-scale model training on NVIDIA hardware.

Models evaluation was performed using NVIDIA NeMo-Skills², following the same settings described in NVIDIA (2025).

3.3 Training Details

Each training session ran for 3,000 steps, with a global batch size of 64 ($\approx 4M$ tokens / step, $\approx 11.5B$ tokens total). We used a learning rate of $5e-6$, with a warmup ratio of 0.03 from $5e-7$. We used the AdamW (Loshchilov and Hutter, 2019) optimizer with a cosine annealing schedule (Loshchilov and Hutter, 2017). Each sample was compiled using the NVIDIA-Nemotron-Nano-12B-v2 chat template³. All samples were then packed into training samples of 60k tokens, using the *first fit decreasing* algorithm. Following the community standard, we only computed the loss on the completions during training. The full example code for training is available⁴.

4 Results

We evaluated the trained models on three popular math benchmarks: GSM8k (Cobbe et al., 2021), AIME 2025 (aim, 2025), and MATH-500 (OpenAI, 2025b). All models were evaluated under the same conditions - temperature = 0.6, top_p = 0.95, tokens_to_generate = 32768, and number_of_repeats= 8.

Table 1 shows the results of the trained models using both DeepSeek-R1 and gpt-oss reasoning styles, evaluated on Math benchmarks. We observed that both tracing styles deliver similar downstream accuracy, with the gpt-oss style generating $4 \times$ fewer tokens per response on average.

Figure 1 shows the training loss when fine-tuning Nemotron-Nano-12B-V2 on the two datasets.

²<https://github.com/NVIDIA-NeMo/Skills/>

³https://huggingface.co/nvidia/NVIDIA-Nemotron-Nano-12B-v2/blob/main/tokenizer_config.json#L8008

⁴<https://gist.github.com/shaltielshmid/af27fc1ac24fcb85592bbf12dadcd10f>

Benchmark		Pass@8		Avg Tokens Generated	
		gpt-oss	DeepSeek-R1	gpt-oss	DeepSeek-R1
Mistral-Nemo-Base-2407	AIME25	30.0	23.3	12,951	29,328
	GSM8K	96.4	97.2	609	3,883
	MATH-500	88.8	88.0	3,456	14,293
Nemotron-Nano-12B-v2-Base	AIME25	83.3	83.3	7,980	20,179
	GSM8K	97.0	97.7	375	2,010
	MATH-500	98.0	99.0	1,350	5,931

Table 1: Pass@8 accuracy and average token usage for training Mistral-Nemo 12B and Nano-V2-12B on the two datasets - gpt-oss vs. DeepSeek-R1. Both datasets produce similar accuracy, while the model trained using DeepSeek-R1 traces produces 4× as many tokens on average.

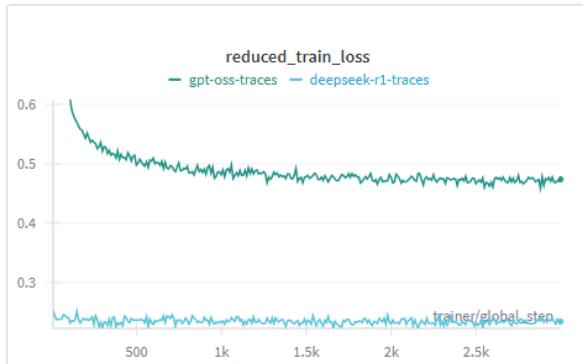


Figure 1: Graph comparing the training loss when fine-tuning Nemotron-Nano-12B-V2 on the two datasets. The DeepSeek-R1 dataset doesn't seem to significantly affect the loss, which is most likely because of the introduction of the reasoning traces in the mid-training.

We observed that the fine-tuning loss on traces generated by DeepSeek-R1 starts very low and stays more or less constant throughout the training. In contrast, the fine-tuning loss on reasoning traces generated by gpt-oss, starts much higher and decreases gradually. This is explained by the fact that the Nemotron-Nano-12B-v2 mid-training dataset already included reasoning traces generated from DeepSeek-R1.

5 Discussion

As we saw in Table 1, the gpt-oss reasoning style achieves accuracy comparable to the DeepSeek-R1 style, while requiring significantly fewer tokens. This preliminary comparative analysis indicates that reasoning with more tokens does not systematically lead to better performance and that a trade-off can be found between the

number of tokens and model accuracy when solving Math problems.

Moreover, our experiments show that even when a reasoning model, such as Nemotron-Nano-12B-V2, has been trained on the more verbose reasoning style, it is possible to teach it to produce correct answers with significantly fewer tokens. This indicates that models are likely not permanently locked into a single reasoning format and can be effectively trained to adopt a new, more efficient one.

The most significant implication of this finding is related to inference efficiency. Models trained on the gpt-oss traces generated, on average, 4× fewer tokens during evaluation. In real-world applications, this translates directly to a $\approx 4\times$ reduction in latency and cost for generating responses. For applications built on test-time scaling, this efficiency gain is non-negligible, enabling faster user experiences and substantially lower operational costs.

For future work, expanding this study to other domains and larger model scales would be highly valuable. Indeed, while math is an excellent proxy for complex, multi-step reasoning, our observations may not generalize to other domains like coding, creative writing, or general instruction-following, where verbosity might play a different role. In addition, we would like to explore a hybrid training approach, mixing both reasoning styles during training and investigate whether models can learn to apply the optimal level of verbosity for a given problem's difficulty.

6 Conclusion

In this work, we compare the impact of training medium-sized language models to utilize test-time scaling with two different reasoning styles gener-

ated from DeepSeek-R1 and gpt-oss. Our experiments demonstrate that both styles achieve comparable accuracy on math benchmarks, while exhibiting different degrees of verbosity. The findings here suggest that verbose reasoning on a higher number of tokens does not necessarily translate to better performance, and that models can be trained to adopt more efficient reasoning patterns. The reduction in the number of generated tokens directly affects deployment costs and latency in production systems. We hope this work encourages further investigation into the relationship between reasoning trace characteristics and both model performance and inference efficiency. Dicta is making the dataset from this research available on HuggingFace⁵ to support further research and serve as a resource for the community.

7 Acknowledgments

We would like to thank the NVIDIA DGX Cloud Lepton team for early access to the platform, and for providing us with the necessary compute to train this model.

References

2025. Aime 2025: A mathematical reasoning benchmark for llms. <https://huggingface.co/datasets/math-ai/aime25>. Dataset of 30 problems from the American Invitational Mathematics Examination (AIME) 2025 used to evaluate large language models.
- Janisha Anand and Sowmyan Soman. 2025. [Introducing nvidia dgx cloud lepton: A unified ai platform built for developers](#). Blog post, NVIDIA Developer Blog.
- Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shafaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, Gerald Shen, Ameya Sunil Mahabaleshwarkar, Bilal Kartal, and Yoshi Suhara et al. 2025. [Llama-nemotron: Efficient reasoning models](#).
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#).
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, and Nikunj Saunshi et al. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#).
- 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, and Zhi-hong Shao et al. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#).
- Eric Harper, Somshubra Majumdar, Oleksii Kuchaiiev, Li Jason, Yang Zhang, Evelina Bakhturina, Vahid Noroozi, Sandeep Subramanian, Koluguri Nithin, Huang Jocelyn, Fei Jia, Jagadeesh Balam, Xuesong Yang, Micha Livne, Yi Dong, Sean Naren, and Boris Ginsburg. [NeMo: a toolkit for Conversational AI and Large Language Models](#).
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. [Tulu 3: Pushing frontiers in open language model post-training](#).
- Ilya Loshchilov and Frank Hutter. 2017. [Sgdr: Stochastic gradient descent with warm restarts](#).
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#).
- Dhruv Nathawani, Igor Gitman, Somshubra Majumdar, Evelina Bakhturina, Ameya Sunil Mahabaleshwarkar, Jian Zhang, and Jane Polak Scowcroft. 2025. [Nemotron-Post-Training-Dataset-v1](#).
- NVIDIA. 2025. [Nvidia nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model](#).
- OpenAI. 2024. [Openai o1 system card](#). *arXiv*, 2412.16720.
- OpenAI. 2025a. [Introducing gpt-oss-120b & gpt-oss-20b: Open-weight language models](#). Web page.
- Hugging Face H4 & OpenAI. 2025b. [Math-500: A 500-problem subset of the math benchmark](#). <https://huggingface.co/datasets/HuggingFaceH4/MATH-500>. Dataset of 500 competition-style mathematics problems sampled from the MATH dataset.

⁵<https://huggingface.co/datasets/dicta-il/MathCOT-oss-vs-DeepSeek>

Mistral AI Team. 2024. Mistral-nemo-base-2407.
[https://huggingface.co/mistralai/
Mistral-Nemo-Base-2407](https://huggingface.co/mistralai/Mistral-Nemo-Base-2407). Pretrained 12B-parameter multilingual/coding LLM under Apache 2 license.

Qwen Team. 2025. [Qwen3 technical report](#).

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](#).

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. [Group sequence policy optimization](#).

A Appendix: System Prompt used for verifying correctly generated answer

You are a meticulous and intelligent judge of mathematical solutions. Your sole purpose is to compare a <GENERATED_ANSWER>, which provides a full step-by-step trace to a problem, with a <CORRECT_ANSWER>, which contains only the final, verified solution.

Your task is to determine if the final conclusion of the <GENERATED_ANSWER> is mathematically equivalent to the <CORRECT_ANSWER>.

Instructions:

1. **Identify the Final Answer:** Carefully parse the <GENERATED_ANSWER> to locate its final conclusion. This may be at the end of the text, often after a phrase like "Therefore, the answer is" or enclosed in a LaTeX \boxed{}.
2. **Compare for Mathematical Equivalence:** Compare the extracted final answer from <GENERATED_ANSWER> with the content of <CORRECT_ANSWER>.
3. **Handle Discrepancies:**
 - * **LaTeX Formatting:** Do not be concerned with differences in LaTeX syntax if the rendered mathematical expression is identical. For example, $x=5$ is the same as $x = 5$.
 - * **Multiple Solutions:** The <GENERATED_ANSWER> may offer more than one possible solution. The check should pass if the <CORRECT_ANSWER> is present as one of these solutions.
4. **Provide a Verdict:**
 - * If the final answers are mathematically equivalent, respond with **MATCH**.
 - * If they are not equivalent, respond with **MISMATCH**.

Think very hard and make sure that you never produce a **MATCH** when it's not correct. Analyze the inputs and provide your verdict.