Llama-Nemotron: Efficient Reasoning Models

Soumye Singhal*, Jiaqi Zeng*, Alexander Bukharin*, Yian Zhang*, Gerald Shen*, Ameya Sunil Mahabaleshwarkar*, Bilal Kartal*, Yoshi Suhara*, Akhiad Bercovich*, Itay Levy*, Izik Golan*, Mohammed Dabbah*, Ran El-Yaniv*, Somshubra Majumdar*, Igor Gitman*, Evelina Bakhturina*, Jimmy J. Zhang*, Bor-Yiing Su*, Guyue Huang*, Izzy Putterman*, Mostofa Patwary*, Oluwatobi Olabiyi*, Olivier Delalleau*, Bryan Catanzaro*, Boris Ginsburg*, Tugrul Konuk*, Oleksii Kuchaiev*



* Core contributors. For full contributors list see Technical report.

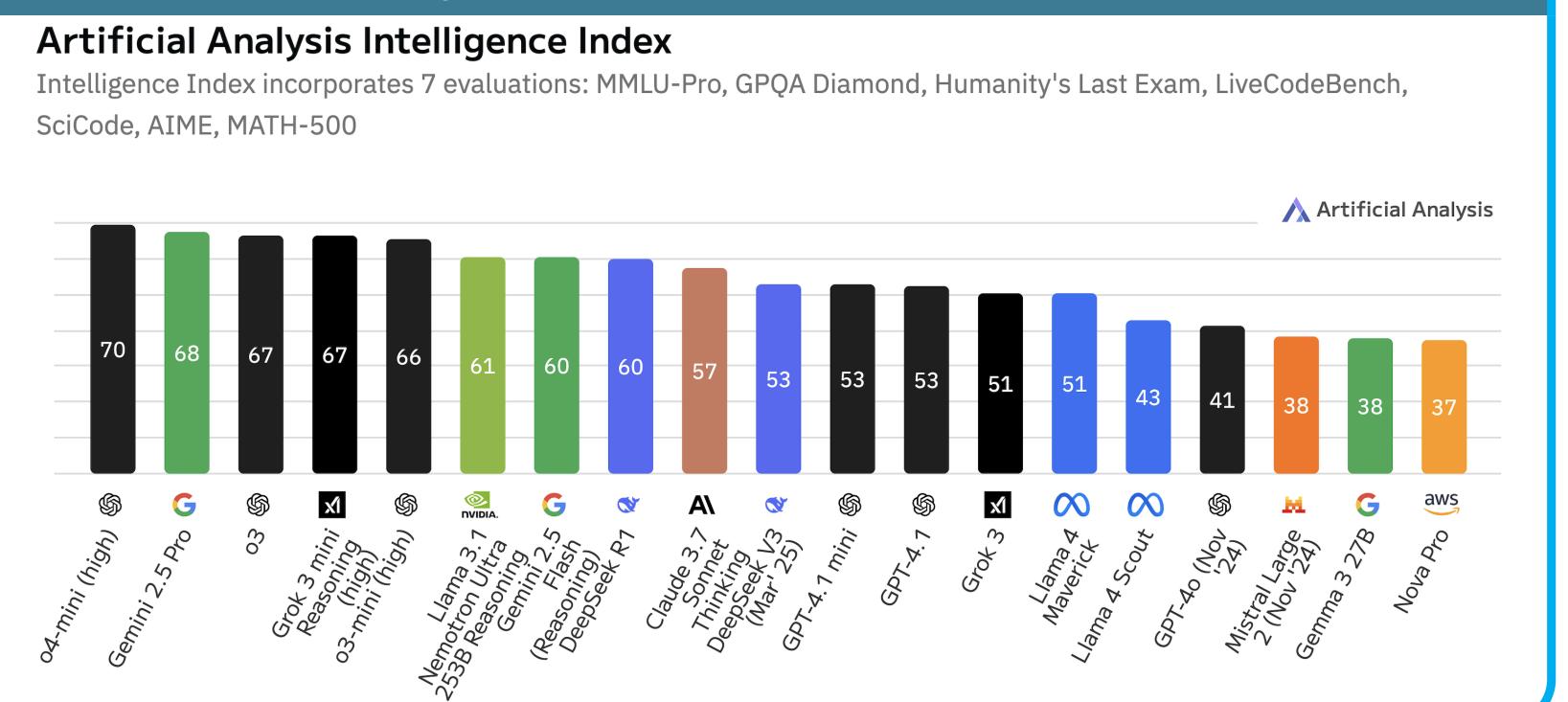
1. Llama-Nemotron-V1 family

Open-weights, open source post-training SW, open post-training & RLHF data. First open-weights with reasoning control On/Off.

3 model sizes:

- LN-Nano (8B and 4B)
- LN-Super (49B)
- LN-Ultra (253B)

Smartest open-weights model as of April 2025. The highest ranking llama variant on lmarena.ai



2. Post-training pipeline Distillation RL for Reasoning RL for Alignment Supervised Fine-Tuning Improve Scientific Reasoning Improve model efficiency Improve Agentic Skills with Reasoning Improve Instruction Following and Chat Llama 3.1 (405B) 60B tokens of NVIDIA generated RL for IF Instruction synthetic and real-world data Following **NVIDIA NeMo NVIDIA NeMo NVIDIA NeMo** Verifier 45 6 Pruning Instruction **Neural Architecture** RL for Scientific Reasoning Reasoning ON/OFF Training Data Following 30k prompts Curriculum RL (FP8) 140k H100 Hours Llama 3.3 253B < 70B Following (IF) Calling (FC) As A Model Judge 8 RL for Chat Llama Nemotron Reward (70B) Pruned Llama 3.1 **NVIDIA** vetted responses 3M prompts Reasoning Q&A 40K Samples Llama 3.3 70B DeepSeek-R1 Qwen 2.5 Knowledge Continued HelpSteer 2 Distillation 50k prompts Reasoning Q&A **NVIDIA** curated (Llama Nemotron **NVIDIA NeMo** 200k Samples prompts Super 49B) Llama Nemotron with Reasoning (253B)ON INVIDIA.

3. RLVR for Scientific Reasoning

While SFT enables strong capabilities through teacher distillation, it limits performance to the teacher's level. Large-scale RL with verifiable rewards empowers LN-Ultra to explore beyond imitation and surpass the teacher.

Key RL Features:

50

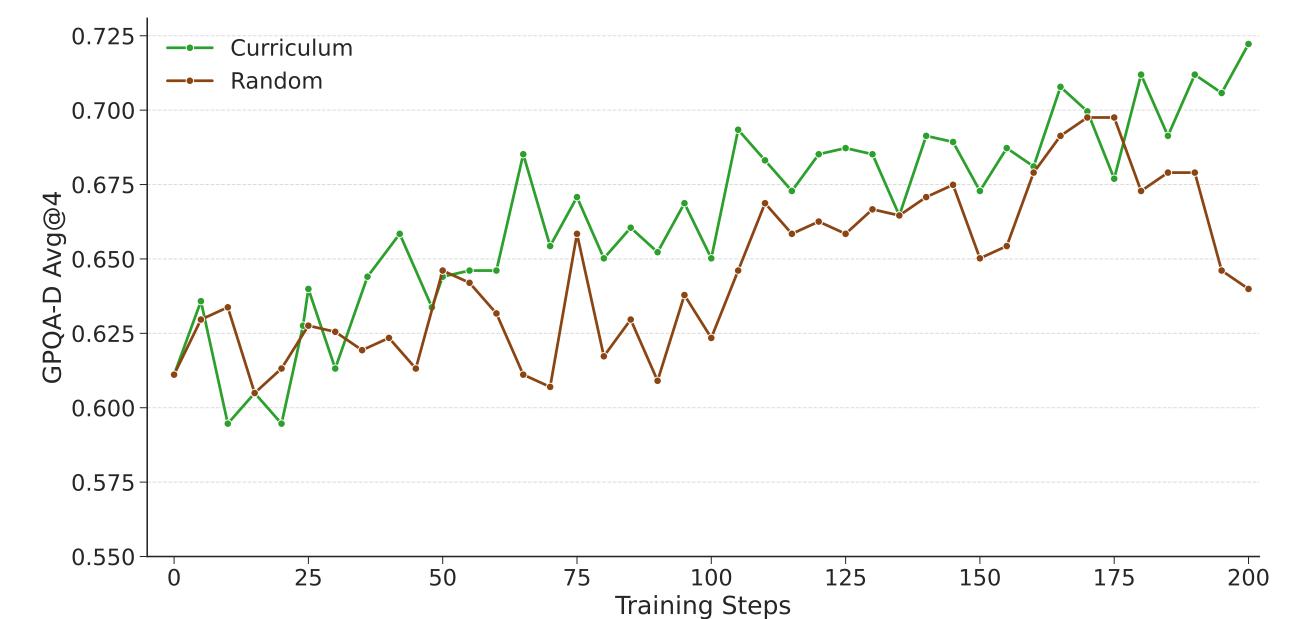
- GRPO training for 140k H100 hours with FP8 inference for rollouts
- Prompt size of 72, 16 responses per prompt with $\tau = 1$ and $top_p = 1$
- Global batch size of 576 with 2 gradient updates per rollout
- Accuracy rewards using Llama-3.3-70B-Instruct as judge
- Format rewards ensuring proper thinking tag usage

5. LN-Ultra Results 96.6 97.0 97.3 96.6 97.0 97.3 92.0 88.9 88.8 87.0 83.2

The RL stage is critical for surpassing teacher performance, particularly on GPQA where LN-Ultra achieves 76.0% vs DeepSeek-R1's 71.5%.

4. Curriculum Learning

We implement an exploration-driven progressive batching strategy to systematically challenge LN-Ultra during RL training. Data is preprocessed by generating 8 responses per question using LN-Super, calculating pass rates, and discarding easy prompts (pass rate $\geq 75\%$).



Progressive Batching:

- Gaussian distribution targeting difficulty progression across batches
- Early batches: high pass rates (easier examples)
- Later batches: low pass rates (harder examples)
- Forces exploration beyond teacher capabilities

6. Links and Resources







NeMo-RL





HF Collection

Technical Report

Post-Training
Dataset

HelpSteer3