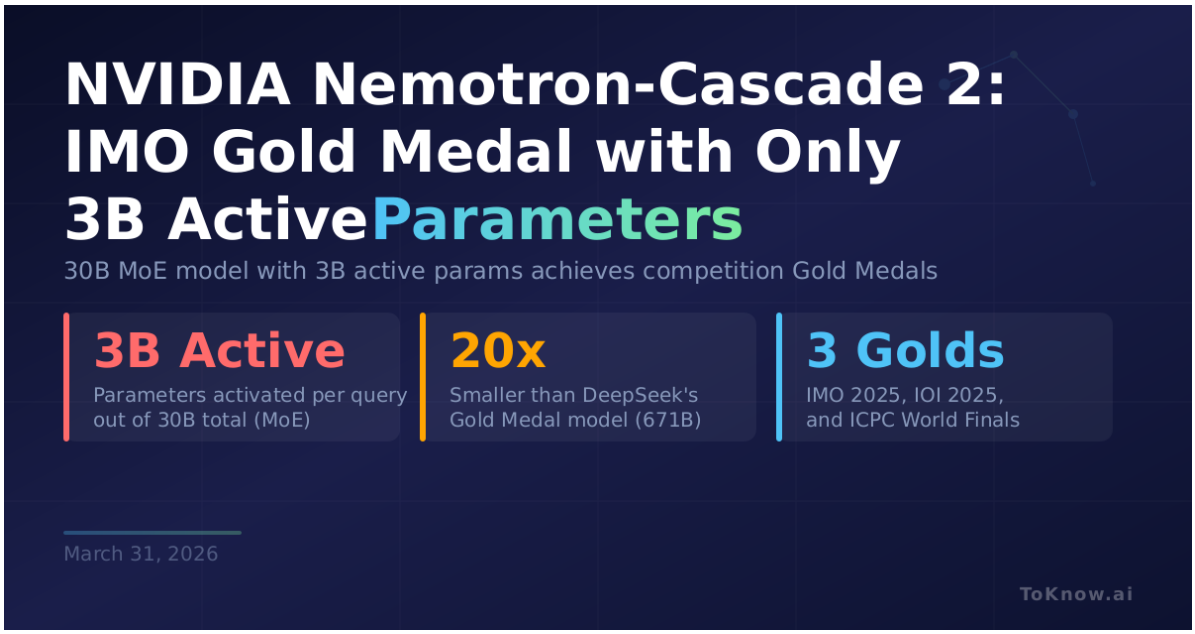
NVIDIA Nemotron-Cascade 2: IMO Gold Medal Reasoning with Only 3B Active Parameters

Kabui, Charles

2026-03-31



NVIDIA released Nemotron-Cascade 2, a 30 billion parameter mixture-of-experts model that activates only 3 billion parameters per query. It scored Gold Medal-level on the 2025 International Mathematical Olympiad (35 points), the International Olympiad in Informatics

(439.3), and the ICPC World Finals (10 of 12 problems). Only one other open-weight model has reached this tier: DeepSeek's V3.2-Speciale, which uses 671B total parameters and 37B active. That makes Nemotron-Cascade 2 roughly 20x smaller. The training recipe combines Cascade RL, where reinforcement learning is applied in stages across math, code, and agentic tasks, with multi-domain on-policy distillation, which uses the best intermediate checkpoints to prevent skill regression during training. NVIDIA released the full model, SFT dataset, and RL dataset openly.

A model solving competition-level math with 3B active parameters changes serving economics. Previous models at this reasoning level required multi-GPU setups. Nemotron-Cascade 2 runs on a single GPU via vLLM with a 262K token context window. On coding, it scores 87.2 on LiveCodeBench v6 and 50.2 on SWE-bench Verified, competitive with much larger models. For teams that need strong reasoning but can't afford to serve a 671B model, this is the new baseline.

The trend is consistent: training methods are compressing what used to require massive scale. GLM-5 showed a 744B MoE matching proprietary models. Nemotron-Cascade 2 reaches similar reasoning capability at 30B. Better RL curricula and distillation strategies are proving worth more than an order of magnitude in parameters.

Sources:

- Nemotron-Cascade 2 on HuggingFace
- Nemotron-Cascade 2 Paper (arXiv)
- Nemotron-Cascade 2 SFT Dataset
- Nemotron-Cascade 2 RL Dataset
