

NVIDIA releases Nemotron 3 Ultra with open weights, data and training recipes

Source copy kept by wheresthe.ai for the story "NVIDIA releases Nemotron 3 Ultra with open weights, data and training recipes" (<https://wheresthe.ai/d/us-nvidia-nemotron-3-ultra-open-2026-06/>).

Source: <https://developer.nvidia.com/blog/nvidia-nemotron-3-ultra-powers-faster-more-efficient-reasoning-for-long-running-agents/>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: ed43aed6651c9fe72a9b983892b5cdd60f506e35f6ba4a5cdc751ebb3e4d1bf6

Documents the source offers: <https://research.nvidia.com/labs/nemotron/files/NVIDIA-Nemotron-3-Ultra-Technical-Report.pdf>

Agentic AI / Generative AI

English 한국어

NVIDIA Nemotron 3 Ultra Powers Faster, More Efficient Reasoning for Long-Running Agents

Jun 04, 2026

By Chris Alexiuk and Chintan Patel

Like

Discuss (0)

L

T

F

R

E

AI-Generated Summary

NVIDIA Nemotron 3 Ultra is a 550B-parameter Mixture-of-Experts model with 55B active parameters built for frontier reasoning and orchestration in long-running agentic systems. The model achieves 5x higher throughput than comparable open models while delivering leading accuracy on benchmarks such as PinchBench, EnterpriseOps-Gym, Terminal-Bench 2.0, IFBench, GDPVal-AA, ProfBench, and Ruler at 1M context length. Nemotron 3 Ultra lowers the cost to task completion by up to 30% by using fewer total tokens and fewer tokens per turn on SWE-bench and Terminal-Bench 2.0. Architectural innovations include a hybrid Mamba-Transformer design, NVFP4 precision that runs on Hopper, Blackwell, and Ampere GPUs, LatentMoE for efficient expert routing, and multi-token prediction to reduce generation time. Multi-Teacher On-Policy Distillation trains the model with more than 10 specialized teacher models that score rollouts asynchronously and co-evolve with the student across iterative stages. Training data additions include 4B synthetic legal tokens, 35B synthesized Wiki-based tokens, 173B refreshed GitHub tokens through September 30, 2025, plus 10M new SFT samples, 1M new RL tasks, and 15 new RL environments. Next Steps Try Nemotron 3 Ultra on build.nvidia.com to spin up an autoresearch flow with Hermes Agent. Download the model weights from Hugging Face for local deployment or customization. Read the Nemotron 3 Ultra technical report for full architectural and benchmark details.

Powered by NVIDIA Nemotron. AI-generated content may summarize information incompletely. Verify important information. [Learn more](#)

Single-turn chatbots are evolving into long-running agents that can reason, maintain context, use tools, and run efficiently across many turns to complete complex workflows.

However, these multi-agent workflows cause token counts to grow quickly. Agents plan, call tools, invoke sub-agents, receive information, and then pass history, outputs, and reasoning steps back into the model continuously. As tasks run longer, this constant communication increases costs and the risk of goal drift.

Developers can solve this using a system of models: frontier reasoning models for orchestration and complex planning, and efficient models for high-volume execution, validation, and tool calling.

NVIDIA is releasing NVIDIA Nemotron 3 Ultra, an open model built to help long-running agents complete tasks faster while lowering cost.

Nemotron 3 Ultra for agent orchestration

Nemotron 3 Ultra is a 550B-parameter Mixture-of-Experts model with 55B active parameters, built for frontier reasoning and orchestration in agentic systems.

Within any agent workflow, most calls are routine, but a critical subset demands deeper reasoning. Nemotron 3 Ultra is built to handle these hard calls: sustaining architectural decisions across coding sessions, synthesizing contradictory evidence across hundreds of research sources, or verifying chip designs across thousands of constraints.

Nemotron 3 Ultra (550B) GLM 5.1 (744B) Kimi K2.6 (1T) Qwen3.5 (397B) Agent Productivity PinchBench 91% 84% 91% 89% Long-horizon Planning EnterpriseOps-Gym 33% 40% 29% 30% Coding Terminal-Bench 2.0 54% 64% 67% 53% Instruction Following IFBench 82% 77% 74% 78% Knowledge Work GDPVal-AA 1,448 1,594 1,508 1,192 Professional Work Tasks ProfBench (Search) 56% 46% 56% 53% Long Context Ruler @1M 95% N/A (max 256K) N/A (max 256K) 90% Table 1. Nemotron 3 Ultra delivers frontier accuracy in a smaller model

Nemotron 3 Ultra is also fast. It achieves 5x higher throughput compared to other open models in its class, enabling long-running agents to complete tasks faster and more efficiently.

Figure 2. Nemotron 3 Ultra achieves 5x faster inference while delivering leading accuracy on the Artificial Analysis Intelligence Index leaderboard, with the output speed of Nemotron 3 Ultra measured using Blackbox endpoints

Nemotron 3 Ultra is also built for efficiency. In experiments on the SWE-bench and Terminal bench 2.0, it completed benchmarks using fewer total tokens and fewer tokens per turn than comparable models. This lowers the cost for agentic tasks by up to 30%.

Figure 3. Nemotron 3 Ultra lowers the cost to task completion by 30%

Breakthroughs powering Nemotron 3 Ultra

To mitigate the typical efficiency-accuracy tradeoffs for high-capacity reasoning models, the Nemotron models introduce architectural innovations:

Post-trained for agent harness Nemotron Ultra is post-trained to deliver consistent accuracy across top harnesses. The model is trained using the NVIDIA NeMo RL and Gym open libraries with one of the largest suites of long-running, task-solving, tool-using datasets in the world.

Ultra is optimized for agent-led open harnesses, not just single-turn chat, and is designed to work within workflows where agents plan, call tools, read observations, delegate to sub-agents, validate outputs, and recover from errors across many turns.

Hybrid Mamba transformer Mamba layers improve sequence efficiency for long-context workloads, while Transformer layers preserve precise recall when agents need to retrieve specific facts from large context windows.

NVFP4 precision The same NVFP4 checkpoint runs on NVIDIA Hopper, NVIDIA Blackwell, and Ampere GPUs. Developers can use one checkpoint across all NVIDIA GPU architectures thanks to specialized NVFP4 quantization kernels. NVFP4 also delivers up to 5x higher throughput per GPU at the same interactivity compared to BF16 on Blackwell.

LatentMoE LatentMoE supports more efficient expert routing, enabling the model to handle workflows spanning reasoning, code generation, tool calls, and domain-specific logic.

Multi-token prediction Multi-token prediction (MTP) helps reduce generation time by predicting multiple future tokens in a single forward pass, improving throughput for long outputs and multi-turn workflows.

Nemotron 3 Ultra adds Multi-Teacher On-Policy Distillation

Multi-Teacher On-Policy Distillation (MOPD) is a training method in which Ultra learns from multiple specialized

teacher models while generating its own attempts during training. More than 10 specialized teacher models are trained, each with its own domain-specific training pipeline. Each teacher scores the model in its area of expertise, helping Ultra improve reasoning across domains more efficiently.

Figure 4. A visual guide to MOPD and the specific flow used for Nemotron 3 Ultra

During MOPD, the student model generates rollouts across domains and receives dense reward signals from the corresponding teacher models. To maximize efficiency, MOPD runs asynchronously, with student rollout generation, teacher scoring, and student optimization fully pipelined.

MOPD is also iterative. After producing an MOPD-trained checkpoint, new rounds of teacher training are initialized from the updated student model, and the improvements are merged into the next MOPD stage.

This co-evolution between students and teachers enables continuous capability improvement and progressively stronger specialization across domains. Users can try MOPD recipes through NeMo-RL, the library that trained the Ultra model.

Training data for stronger agent reasoning

As with all Nemotron open model launches, much of the training data pipeline is released as permissively as possible. For partners in enterprise and sovereign AI development, training data transparency and provenance matter as much as capability.

Domain-specific pre-training data

Building on a 10T token pre-training foundation, Nemotron 3 Ultra adds 212B new tokens targeting three high-value domain gaps:

4B tokens of synthetic legal data, increasing the proxy LegalBench average from 64.6% to 74.7%

35B tokens of synthesized Wiki-based data, boosting proxy SimpleQA from 40.2% to 50.2%

173B refreshed GitHub tokens through Sept. 30, 2025

Post-training data and RL environments

This launch is also releasing 10M new SFT samples, 1M new RL tasks across multiple domains, and 15 net-new RL environments, bringing the cumulative Nemotron open data totals to 50M SFT samples, 2M RL tasks, and 55 RL environments.

The result is SWEBench Verified scores between 65% and 70.4% across Pi, OpenHands, Hermes, OpenCode, and Mini SWE Agent—consistent performance regardless of which framework you deploy.

Finetune for your domain

Nemotron 3 Ultra can be fine-tuned using LoRA, SFT, and reinforcement learning using the NVIDIA NeMo libraries . Developers can get started with the following recipes.

Nemotron 3 Ultra Recipes:

SFT LoRA: NeMo Automodel (H100 Recipe , GB200 Recipe)

Full SFT: NeMo Megatron Bridge Recipes

Reinforcement Learning: NeMo RL GRPO recipe , GRPO LoRA recipe , MOPD recipe

Deployment

Deploy Nemotron Ultra using Dynamo Recipes , for KV-aware routing, multi-token prediction (MTP), and disaggregated prefill/decode.

See it in action

This walkthrough shows how to spin up and run an autoresearch flow using Hermes Agent powered by Nemotron 3 Ultra on build.nvidia.com .

Video 1. A tutorial walkthrough building an autonomous assistant with Hermes Agent and Nemotron 3 Ultra

Run agents more safely with NVIDIA NemoClaw and NVIDIA OpenShell

Nemotron models integrate with leading open agent frameworks. To build a secure, always-on agentic system, it is important to understand the reference stack:

Hermes Agent and OpenClaw: These are popular agent harnesses that provide the orchestration loops, memory, and tools for multi-turn workflows. Hermes Agent is now officially available and fully supported for use with Nemotron.

NVIDIA OpenShell: Available now in early preview, OpenShell is the secure runtime environment (part of the NVIDIA Agent Toolkit) where autonomous agents and their generated code execute.

NVIDIA NemoClaw: This is the open-source blueprint that ties the environment together. With a single command, NemoClaw installs the OpenShell runtime—providing a secure environment for running autonomous agents like Hermes Agent more safely alongside open-source models like Nemotron.

Build safer and voice-enabled agents

Two new Nemotron models are also launching:

Nemotron 3.5 Content Safety For teams building safer enterprise AI, Nemotron 3.5 Content Safety is an open, efficient 4B guardrail model for classifying unsafe, disallowed, or policy-violating content across text, images, and combined inputs.

Covering 23 safety categories and 12 languages, it can be used as an inference-time guardrail, as a judge for LLM safety testing and evaluation, or with the accompanying training dataset to post-train models for safer behavior. Custom policy support and reasoning trails help enterprises adapt safety decisions to domain-specific rules, audit classifications, and deploy safety controls across global AI workflows. Read the Hugging Face post to learn more.

Nemotron 3.5 ASR For voice-native agents, Nemotron 3.5 ASR uses the same cache-aware streaming architecture as its English predecessor, Nemotron 3 ASR, to process audio deltas instantly. Eliminating redundant buffered compute ensures sub-100 ms latency for natural, real-time voice orchestration for your agentic swarms.

The English model has seen strong developer adoption, including powering the voice input feature in Microsoft GitHub Copilot CLI , used by more than 20M developers. An independent benchmark of 50+ on-device ASR configurations identified Nemotron 3 ASR as the strongest candidate for real-time English streaming on resource-constrained hardware. Now, that same architecture goes multilingual with support for 40+ languages in a single checkpoint.

Updated open licensing for broader adoption

Nemotron model releases are moving to OpenMDW-1.1, the Linux Foundation's permissive license purpose-built for open AI model distributions. OpenMDW is designed to cover the full set of model materials, including architecture, parameters, documentation, software, and other related artifacts, under a single framework.

This gives developers and enterprises clearer terms for using, modifying, redistributing, and deploying Nemotron models, while reducing the licensing ambiguity that can slow evaluation and adoption of open models.

Start building today

Nemotron 3 Ultra is fully open—including weights, data, and recipes—so developers can adapt the models to domain-specific workflows and deploy them anywhere. It is available across leading inference platforms, and packaged as an NVIDIA NIM microservice, can run anywhere.

Try it on Perplexity with a Pro subscription or through API, OpenRouter , Anaconda , or build.nvidia.com .

Download the weights from Hugging Face , launch an optimized instance through NVIDIA NIM, or start with the cookbooks to get running in minutes.

Nemotron 3 Ultra is available through an ecosystem of partners:

Model customization services: Applied Compute , Prime Intellect , Unsloth

Inference software: SGLang , TRT-LLM , vLLM

Cloud service providers: Amazon SageMaker JumpStart , Google Cloud, Microsoft Foundry , Oracle Cloud

Inference service providers: Baseten , DeepInfra, Eigen AI , fal (ASR), Fireworks AI, FriendlyAI, Modal , ModelScope , Ollama cloud , Simplismart

AI cloud and services: Bitdeer AI , CoreWeave , Dell Enterprise Hub , Crusoe , DigitalOcean , GMI Cloud , Lightning AI , Nebius Token Factory , Together AI , Vultr

Check out the GitHub repository for getting-started instructions for agent harness, including BlackBox AI , Cline , CrewAI , Factory AI , Hermes Agent, Kilo Code , LangChain Deep Agents , OpenClaw , OpenCode , OpenHands , and Pi .

For the full technical details, read the Nemotron 3 Ultra technical report .

Stay up to date on NVIDIA Nemotron by subscribing to NVIDIA news and following NVIDIA AI on LinkedIn , X , Discord , and YouTube .

Visit the Nemotron developer page for resources to get started. Explore open Nemotron models and datasets on Hugging Face and Blueprints on build.nvidia.com .

Engage with Nemotron livestreams , tutorials , and the developer community on the NVIDIA forum and Discord .

Discuss (0)

Like

Tags

Agentic AI / Generative AI | Developer Tools & Techniques | General | Agent toolkit | Nemotron | Intermediate Technical | Deep dive | AI Agent | featured | LLMs | NemoClaw | OpenShell

About the Authors

About Chris Alexiuk

Chris Alexiuk is a deep learning developer advocate at NVIDIA, working on creating technical assets that help developers use the incredible suite of AI tools available at NVIDIA. Chris comes from a machine learning and data science background, and he is obsessed with everything and anything about large language models.

[View all posts by Chris Alexiuk](#)

About Chintan Patel

Chintan Patel is a senior product manager at NVIDIA focused on bringing GPU-accelerated solutions to the HPC community. He leads the management and offering of the HPC application containers on the NVIDIA GPU Cloud registry. Prior to NVIDIA, he held product management, marketing and engineering positions at Micrel, Inc. He holds an MBA from Santa Clara University and a bachelor's degree in electrical engineering and computer science from UC Berkeley.

[View all posts by Chintan Patel](#)

Comments