

NVIDIA Debuts Nemotron 3 Family of Open Models

News Summary:

- The Nemotron 3 family of open models — in Nano, Super and Ultra sizes — introduces the most efficient family of open models with leading accuracy for building agentic AI applications.
- Nemotron 3 Nano delivers 4x higher throughput than Nemotron 2 Nano and delivers the most tokens per second for multi-agent systems at scale through a breakthrough hybrid mixture-of-experts architecture.
- Nemotron achieves superior accuracy from advanced reinforcement learning techniques with concurrent multi-environment post-training at scale.
- NVIDIA is the first to release a collection of state-of-the-art open models, training datasets and reinforcement learning environments and libraries for building highly accurate, efficient, specialized AI agents.

NVIDIA today announced the NVIDIA Nemotron™ 3 family of open models, data and libraries designed to power transparent, efficient and specialized agentic AI development across industries.

The Nemotron 3 models — with Nano, Super and Ultra sizes — introduce a breakthrough [hybrid latent mixture-of-experts \(MoE\)](#) architecture that helps developers build and deploy reliable multi-agent systems at scale.

As organizations shift from single-model chatbots to collaborative multi-agent AI systems, developers face mounting challenges, including communication overhead, context drift and high inference costs. In addition, developers require transparency to trust the models that will automate their complex workflows. Nemotron 3 directly addresses these challenges, delivering the performance and openness customers need to build specialized, agentic AI.

“Open innovation is the foundation of AI progress,” said Jensen Huang, founder and CEO of NVIDIA. “With Nemotron, we’re transforming advanced AI into an open platform that gives developers the transparency and efficiency they need to build agentic systems at scale.”

NVIDIA Nemotron supports NVIDIA’s broader sovereign AI efforts, with organizations from [Europe](#) to [South Korea](#) adopting open, transparent and efficient models that allow them to build AI systems aligned to their own data, regulations and values.

Early adopters, including Accenture, Cadence, CrowdStrike, Cursor, Deloitte, EY, Oracle Cloud Infrastructure, Palantir, Perplexity, ServiceNow, Siemens, Synopsys and Zoom, are integrating models from the Nemotron family to power AI workflows across manufacturing, cybersecurity, software development, media, communications and other industries.

“NVIDIA and ServiceNow have been shaping the future of AI for years, and the best is yet to come,” Bill McDermott, chairman and CEO of ServiceNow. “Today, we’re taking a major step forward in empowering leaders across all industries to fast-track their agentic AI strategy. ServiceNow’s intelligent workflow automation combined with NVIDIA Nemotron 3 will continue to define the standard with unmatched efficiency, speed and accuracy.”

As multi-agent AI systems expand, developers are increasingly relying on proprietary models for state-of-the-art reasoning while using more efficient and customizable open models to drive down costs. Routing tasks between frontier-level models and Nemotron in a single workflow gives agents the most intelligence while optimizing [tokenomics](#).

“Perplexity is built on the idea that human curiosity will be amplified by accurate AI built into exceptional tools, like AI assistants,” said Aravind Srinivas, CEO of Perplexity. “With our agent router, we can direct workloads to the best fine-tuned open models, like Nemotron 3 Ultra, or leverage leading proprietary models when tasks benefit from their unique capabilities — ensuring our AI assistants operate with exceptional speed, efficiency and scale.”

The open Nemotron 3 models enable startups to build and iterate faster on AI agents and accelerate innovation from prototype to enterprise deployment. General Catalyst, Mayfield and Sierra Ventures’ portfolio companies are exploring Nemotron 3 to build AI teammates that support human-AI collaboration.

“NVIDIA’s open model stack and the NVIDIA Inception program give early-stage companies the models, tools and a cost-effective infrastructure to experiment, differentiate and scale fast,” said Navin Chaddha, managing partner at Mayfield. “Nemotron 3 gives founders a running start on building agentic AI applications and AI teammates, and helps them tap into NVIDIA’s massive installed base.”

Nemotron 3 Reinvents Multi-Agent AI With Efficiency and Accuracy

The Nemotron 3 family of [MoE models](#) includes three sizes:

- Nemotron 3 Nano, a small, 30-billion-parameter model that activates up to 3 billion parameters at a time for targeted,

highly efficient tasks.

- Nemotron 3 Super, a high-accuracy reasoning model with approximately 100 billion parameters and up to 10 billion active per token, for multi-agent applications.
- Nemotron 3 Ultra, a large reasoning engine with about 500 billion parameters and up to 50 billion active per token, for complex AI applications.

Available today, Nemotron 3 Nano is the most compute-cost-efficient model, optimized for tasks such as software debugging, content summarization, AI assistant workflows and information retrieval at low inference costs. The model uses a unique hybrid MoE architecture to deliver gains in efficiency and scalability.

This design achieves up to 4x higher token throughput compared with Nemotron 2 Nano and reduces reasoning-token generation by up to 60%, significantly lowering inference costs. With a 1-million-token context window, Nemotron 3 Nano remembers more, making it more accurate and better capable of connecting information over long, multistep tasks.

Artificial Analysis, an independent organization that benchmarks AI, [ranked the model](#) as the most open and efficient among models of the same size, with leading accuracy.

Nemotron 3 Super excels at applications that require many collaborating agents to achieve complex tasks with low latency. Nemotron 3 Ultra serves as an advanced reasoning engine for AI workflows that demand deep research and strategic planning.

Nemotron 3 Super and Ultra use NVIDIA's ultraefficient 4-bit NVFP4 training format on the NVIDIA Blackwell architecture, significantly cutting memory requirements and speeding up training. This efficiency allows larger models to be trained on existing infrastructure without compromising accuracy relative to higher-precision formats.

With the Nemotron 3 family of models, developers can choose the open model that is right-sized for their specific workloads, scaling from dozens to hundreds of agents while benefiting from faster, more accurate long-horizon reasoning for complex workflows.

New Open Tools and Data for AI Agent Customization

NVIDIA also released a collection of training datasets and state-of-the-art reinforcement learning libraries available to anyone building specialized AI agents.

Three trillion tokens of new Nemotron [pretraining](#), [post-training](#) and [reinforcement learning](#) datasets supply the rich reasoning, coding and multistep workflow examples needed to create highly capable, domain-specialized agents. The [Nemotron Agentic Safety Dataset](#) provides real-world telemetry to help teams evaluate and strengthen the safety of complex agent systems.

To accelerate development, NVIDIA released the [NeMo Gym](#) and [NeMo RL](#) open-source libraries, which provide the training environments and post-training foundation for Nemotron models, along with NeMo Evaluator to validate model safety and performance. All tools and datasets are now available on GitHub and Hugging Face.

Nemotron 3 is supported by [LM Studio](#), llama.cpp, [SGLang](#) and [vLLM](#). In addition, Prime Intellect and [Unsloth](#) are integrating NeMo Gym's ready-to-use training environments directly into their workflows, giving teams faster, easier access to powerful reinforcement learning training.

Get Started With NVIDIA Open Models

Nemotron 3 Nano is available today on [Hugging Face](#) and through inference service providers including [Baseten](#), [DeepInfra](#), [Fireworks](#), [FriendliAI](#), [OpenRouter](#) and [Together AI](#).

Nemotron is offered on enterprise AI and data infrastructure platforms, including Couchbase, DataRobot, H2O.ai, JFrog, Lambda and UiPath. For customers on public clouds, Nemotron 3 Nano will be available on AWS via Amazon Bedrock (serverless) as well as supported on Google Cloud, CoreWeave, Crusoe, Microsoft Foundry, [Nebius](#), Nscale and Yotta soon.

Nemotron 3 Nano is available as an [NVIDIA NIM™ microservice](#) for secure, scalable deployment anywhere on NVIDIA-accelerated infrastructure for maximum privacy and control.

Nemotron 3 Super and Ultra are expected to be available in the first half of 2026.

About NVIDIA

[NVIDIA](#) (NASDAQ: NVDA) is the world leader in AI and accelerated computing.

Certain statements in this press release including, but not limited to, statements as to: with Nemotron, NVIDIA transforming advanced AI into an open platform that gives developers the transparency and efficiency they need to build agentic systems at scale; the benefits, impact, performance, and availability of NVIDIA's products, services, and technologies; expectations with respect to NVIDIA's third party arrangements, including with its collaborators and partners; expectations with respect to technology developments; and other statements that are not historical facts are forward-looking statements within the meaning of Section 27A of the Securities Act of 1933, as amended, and Section 21E of the Securities Exchange Act of 1934,

as amended, which are subject to the “safe harbor” created by those sections based on management’s beliefs and assumptions and on information currently available to management and are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic and political conditions; NVIDIA’s reliance on third parties to manufacture, assemble, package and test NVIDIA’s products; the impact of technological development and competition; development of new products and technologies or enhancements to NVIDIA’s existing product and technologies; market acceptance of NVIDIA’s products or NVIDIA’s partners’ products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of NVIDIA’s products or technologies when integrated into systems; and changes in applicable laws and regulations, as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company’s website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

Many of the products and features described herein remain in various stages and will be offered on a when-and-if-available basis. The statements above are not intended to be, and should not be interpreted as a commitment, promise, or legal obligation, and the development, release, and timing of any features or functionalities described for our products is subject to change and remains at the sole discretion of NVIDIA. NVIDIA will have no liability for failure to deliver or delay in the delivery of any of the products, features or functions set forth herein.

© 2025 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, NVIDIA Nemotron and NVIDIA NIM are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated. Features, pricing, availability and specifications are subject to change without notice.

Natalie Hereth
NVIDIA Corporation
nhereth@nvidia.com