

NVIDIA Launches Family of Open Reasoning AI Models for Developers and Enterprises to Build Agentic AI Platforms

- ***Post-Trained by NVIDIA, New Llama Nemotron Reasoning Models Provide Business-Ready Foundation for Agentic AI***
- ***Accenture, Amdocs, Atlassian, Box, Cadence, CrowdStrike, Deloitte, IQVIA, Microsoft, SAP and ServiceNow Pioneering Reasoning AI Agents With NVIDIA to Transform Work***

GTC—NVIDIA today announced the open Llama Nemotron family of models with reasoning capabilities, designed to provide developers and enterprises a business-ready foundation for creating advanced AI agents that can work independently or as connected teams to solve complex tasks.

Built on Llama models, the NVIDIA Llama Nemotron reasoning family delivers on-demand AI reasoning capabilities. NVIDIA enhanced the new reasoning model family during post-training to improve multistep math, coding, reasoning and complex decision-making.

This refinement process boosts accuracy of the models by up to 20% compared with the base model and optimizes inference speed by 5x compared with other leading open reasoning models. The improvements in inference performance mean the models can handle more complex reasoning tasks, enhance decision-making capabilities and reduce operational costs for enterprises.

Leading agent AI platform pioneers — including [Accenture](#), [Amdocs](#), Atlassian, [Box](#), [Cadence](#), [CrowdStrike](#), Deloitte, [IQVIA](#), Microsoft, [SAP](#) and [ServiceNow](#) — are collaborating with NVIDIA on its new reasoning models and software.

“Reasoning and agentic AI adoption is incredible,” said Jensen Huang, founder and CEO of NVIDIA. “NVIDIA’s open reasoning models, software and tools give developers and enterprises everywhere the building blocks to create an accelerated agentic AI workforce.”

NVIDIA Post-Training Boosts Accuracy and Reliability for Enterprise Reasoning

Built to deliver production-ready AI reasoning, the Llama Nemotron model family is available as [NVIDIA NIM™ microservices](#) in Nano, Super and Ultra sizes — each optimized for different deployment needs.

The Nano model delivers the highest accuracy on PCs and edge devices, the Super model offers the best accuracy and highest throughput on a single GPU, and the Ultra model will provide maximum agentic accuracy on multi-GPU servers.

NVIDIA conducted extensive post-training on [NVIDIA DGX™ Cloud](#) using high-quality curated synthetic data [generated](#) by NVIDIA Nemotron™ and other open models, as well as additional curated datasets cocreated by NVIDIA.

The tools, datasets and post-training optimization techniques used to develop the models will be openly available, giving enterprises the flexibility to build their own custom reasoning models.

Agentic Platforms Team With NVIDIA to Enhance Reasoning for Industries

Agentic AI platform industry leaders are working with the Llama Nemotron reasoning models to deliver advanced reasoning to enterprises.

Microsoft is integrating Llama Nemotron reasoning models and NIM microservices into Microsoft Azure AI Foundry. This expands the Azure AI Foundry model catalog with options for customers to enhance services like Azure AI Agent Service for Microsoft 365.

SAP is tapping Llama Nemotron models to advance SAP Business AI solutions and Joule, the AI copilot from SAP. Additionally, it is using NVIDIA NIM and NVIDIA NeMo™ microservices to promote increased code completion accuracy for SAP ABAP programming language models.

“We are collaborating with NVIDIA to integrate Llama Nemotron reasoning models into Joule to enhance our AI agents, making them more intuitive, accurate and cost effective,” said Walter Sun, global head of AI at SAP. “These advanced reasoning models will refine and rewrite user queries, enabling our AI to better understand inquiries and deliver smarter, more efficient AI-powered experiences that drive business innovation.”

ServiceNow is harnessing Llama Nemotron models to build AI agents that offer greater performance and accuracy to enhance enterprise productivity across industries.

Accenture has made NVIDIA Llama Nemotron reasoning models available on its AI Refinery platform — including new industry agent solutions [announced today](#) — to enable clients to rapidly develop and deploy custom AI agents tailored to industry-specific challenges, accelerating business transformation.

Deloitte is planning to incorporate Llama Nemotron reasoning models into its recently announced Zora AI agentic AI platform designed to support and emulate human decision-making and action with agents that include deep functional- and industry-specific business knowledge and built-in transparency.

NVIDIA AI Enterprise Delivers Essential Tools for Agentic AI

Developers can deploy NVIDIA Llama Nemotron reasoning models with new NVIDIA agentic AI tools and software to streamline the adoption of advanced reasoning in collaborative AI systems.

All part of the [NVIDIA AI Enterprise](#) software platform, the latest agentic AI building blocks include:

- The [NVIDIA AI-Q Blueprint](#), which enables enterprises to connect knowledge to AI agents that can autonomously perceive, reason and act. Built with NVIDIA NIM microservices, the blueprint integrates NVIDIA NeMo Retriever™ for multimodal information retrieval and enables agent and data connections, optimization and transparency using the open-source [NVIDIA Agent Intelligence toolkit](#).
- The [NVIDIA AI Data Platform](#), a customizable reference design for a new class of enterprise infrastructure with AI query agents built with the AI-Q Blueprint.
- [New NVIDIA NIM microservices](#), which optimize inference for complex agentic AI applications and enable continuous learning and real-time adaptation across any environment. The microservices ensure reliable deployment of the latest models from leading model builders including Meta, Microsoft and Mistral AI.
- NVIDIA NeMo microservices, which provide an efficient, enterprise-grade solution to quickly establish and maintain a robust [data flywheel](#) that enables AI agents to continuously learn from human- and AI-generated feedback. The NVIDIA AI Blueprint for building a data flywheel will offer a reference architecture for developers to easily build and optimize data flywheels using NVIDIA microservices.

Availability

The NVIDIA Llama Nemotron Nano and Super models and NIM microservices are available as a hosted application programming interface from [build.nvidia.com](#) and Hugging Face. [Access](#) for development, testing and research is free for members of the NVIDIA Developer Program.

Enterprises can run Llama Nemotron NIM microservices in production with NVIDIA AI Enterprise on accelerated data center and cloud infrastructure. Developers can sign up to be notified when NVIDIA NeMo microservices are publicly available.

The [NVIDIA AI-Q Blueprint](#) is expected to be available in April. The NVIDIA Agent Intelligence toolkit is available now on [GitHub](#).

About NVIDIA

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

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, availability, and performance of NVIDIA's products, services, and technologies; third parties adopting NVIDIA's products and technologies and the benefits and impact thereof; NVIDIA's open reasoning models, software and tools giving developers and enterprises everywhere the building blocks to create an accelerated agentic AI workforce are forward-looking statements that 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 conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems; 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, DGX, NVIDIA NeMo, NVIDIA Nemotron, NVIDIA NeMo Retriever 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.

Anna Kiachian
Senior PR Manager
NVIDIA Corporation
+1-650-224-9820
akiachian@nvidia.com