



NVIDIA NIM Revolutionizes Model Deployment, Now Available to Transform World's Millions of Developers Into Generative AI Developers

- *150+ Partners Across Every Layer of AI Ecosystem Embedding NIM Inference Microservices to Speed Enterprise AI Application Deployments From Weeks to Minutes*
- *NVIDIA Developer Program Members Gain Free Access to NIM for Research, Development and Testing*

COMPUTEX—NVIDIA today announced that the world's 28 million developers can now download [NVIDIA NIM™](#) — inference microservices that provide models as optimized containers — to deploy on clouds, data centers or workstations, giving them the ability to easily build generative AI applications for copilots, chatbots and more, in minutes rather than weeks.

These new generative AI applications are becoming increasingly complex and often utilize multiple models with different capabilities for generating text, images, video, speech and more. NVIDIA NIM dramatically increases developer productivity by providing a simple, standardized way to add generative AI to their applications.

NIM also enables enterprises to maximize their infrastructure investments. For example, running Meta Llama 3-8B in a NIM produces up to 3x more generative AI tokens on accelerated infrastructure than without NIM. This lets enterprises boost efficiency and use the same amount of compute infrastructure to generate more responses.

Nearly 200 technology partners — including Cadence, [Cloudera](#), [Cohesity](#), [DataStax](#), [NetApp](#), Scale AI and [Synopsys](#) — are integrating NIM into their platforms to speed generative AI deployments for domain-specific applications, such as copilots, code assistants and digital human avatars. [Hugging Face](#) is now offering NIM — starting with [Meta Llama 3](#).

"Every enterprise is looking to add generative AI to its operations, but not every enterprise has a dedicated team of AI researchers," said Jensen Huang, founder and CEO of NVIDIA. "Integrated into platforms everywhere, accessible to developers everywhere, running everywhere — NVIDIA NIM is helping the technology industry put generative AI in reach for every organization."

Enterprises can deploy AI applications in production with NIM through the [NVIDIA AI Enterprise](#) software platform. Starting next month, members of the [NVIDIA Developer Program](#) can access NIM for free for research, development and testing on their preferred infrastructure.

40+ NIM Microservices Power Gen AI Models Across Modalities

NIM containers are pre-built to speed model deployment for GPU-accelerated inference and can include [NVIDIA CUDA®](#) software, [NVIDIA Triton Inference Server™](#) and [NVIDIA TensorRT™-LLM](#) software.

Over 40 NVIDIA and community models are available to experience as NIM endpoints on [ai.nvidia.com](#), including [Databricks DBRX](#), Google's open model Gemma, Meta Llama 3, Microsoft Phi-3, Mistral Large, Mixtral 8x22B and Snowflake Arctic.

Developers can now access NVIDIA NIM microservices for Meta Llama 3 models from the [Hugging Face](#) AI platform. This lets developers easily access and run the Llama 3 NIM in just a few clicks using Hugging Face Inference Endpoints, powered by NVIDIA GPUs on their preferred cloud.

Enterprises can use NIM to run applications for generating text, images and video, speech and digital humans. With [NVIDIA BioNeMo™](#) NIM microservices for digital biology, researchers can build novel protein structures to accelerate drug discovery.

Dozens of healthcare companies are [deploying NIM](#) to power generative AI inference across a range of applications, including surgical planning, digital assistants, drug discovery and clinical trial optimization.

With new [NVIDIA ACE NIM microservices](#), developers can easily build and operate interactive, lifelike digital humans in applications for customer service, telehealth, education, gaming and entertainment.

Hundreds of AI Ecosystem Partners Embedding NIM

Platform providers including [Canonical](#), [Red Hat](#), [Nutanix](#) and [VMware](#) (acquired by Broadcom) are supporting NIM on [open-source KServe](#) or enterprise solutions. AI application companies [Hippocratic AI](#), [Glean](#), [Kinetica](#) and [Redis](#) are also deploying NIM to power generative AI inference.

Leading AI tools and MLOps partners — including Amazon SageMaker, Microsoft Azure AI, Dataiku, DataRobot, [deepset](#), Domino Data Lab, [LangChain](#), [Llama Index](#), [Replicate](#), Run.ai, [Saturn Cloud](#), Securiti AI and [Weights & Biases](#) — have also embedded NIM into their platforms to enable developers to build and deploy domain-specific generative AI applications with optimized inference.

Global system integrators and service delivery partners Accenture, Deloitte, Infosys, [Latentview](#), [Quantiphi](#), SoftServe, Tata Consultancy Services (TCS) and Wipro have created NIM competencies to help the world's enterprises quickly develop and deploy production AI strategies.

Enterprises can run NIM-enabled applications virtually anywhere, including on [NVIDIA-Certified Systems](#)™ from global infrastructure manufacturers Cisco, [Dell Technologies](#), [Hewlett-Packard Enterprise](#), [Lenovo](#) and Supermicro, as well as server manufacturers [ASRock Rack](#), [ASUS](#), [GIGABYTE](#), [Ingrasys](#), [Inventec](#), [Pegatron](#), QCT, Wistron and Wiwynn. NIM microservices have also been integrated into [Amazon Web Services](#), [Google Cloud](#), [Azure](#) and [Oracle Cloud Infrastructure](#).

Titans of Industry Amp Up Generative AI With NIM

Industry leaders Foxconn, Pegatron, [Amdocs](#), Lowe's, [ServiceNow](#) and Siemens are among the businesses using NIM for generative AI applications in manufacturing, healthcare, financial services, retail, customer service and more:

- **Foxconn** — the world's largest electronics manufacturer — is using NIM in the development of domain-specific LLMs embedded into a variety of internal systems and processes in its AI factories for smart manufacturing, smart cities and smart electric vehicles.
- **Pegatron** — a Taiwanese electronics manufacturing company — is leveraging NIM for Project TaME, a Taiwan Mixtral of Experts model designed to advance the development of local LLMs for industries.
- **Amdocs** — a leading global provider of software and services to communications and media companies — is [using NIM](#) to run a customer billing LLM that significantly lowers the cost of tokens, improves accuracy by up to 30% and reduces latency by 80%, driving near real-time responses.
- **Lowe's** — a FORTUNE® 50 home improvement company — is using generative AI for a variety of use cases. For example, the retailer is leveraging NVIDIA NIM inference microservices to elevate experiences for associates and customers.
- **ServiceNow** — the AI platform for business transformation — announced earlier this year that it was one of the first platform providers to access NIM to enable fast, scalable and more cost-effective LLM development and deployment for its customers. NIM microservices are integrated within the Now AI multimodal model and are available to customers that have ServiceNow's generative AI experience, Now Assist, installed.
- **Siemens** — a global technology company focused on industry, infrastructure, transport and healthcare — is integrating its operational technology with NIM microservices for shop floor AI workloads. It is also building an on-premises version of its Industrial Copilot for Machine Operators using NIM.

Availability

Developers can experiment with NVIDIA microservices at [ai.nvidia.com](#) at no charge. Enterprises can deploy production-grade NIM microservices with [NVIDIA AI Enterprise](#) running on NVIDIA-Certified Systems and leading cloud platforms. Starting next month, members of the [NVIDIA Developer Program](#) will gain free access to NIM for research and testing.

Watch [Huang's COMPUTEX keynote](#) to learn more about NVIDIA NIM.

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, performance, features, and availability of NVIDIA's products and technologies, including NVIDIA NIM, NVIDIA CUDA, NVIDIA Triton Inference Server, NVIDIA TensorRT-LLM software, NVIDIA Developer program, NVIDIA BioNeMo, NVIDIA-Certified Systems, and NVIDIA AI Enterprise; our collaborations and partnerships with third parties and the benefits and impacts thereof; third parties using or adopting our products or technologies; every enterprise looking to add generative AI to its operations; and NVIDIA NIM helping the technology industry put generative AI in reach for every organization 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.

© 2024 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, BioNeMo, CUDA, NVIDIA NIM, NVIDIA Triton Inference Server and TensorRT 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