# NVIDIA NIM

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

1. *NVIDIA Developer Program Members Gain Free Access to NIM for Research, Development and Testing*

[NVIDIA NIM](https://www.nvidia.com/en-us/ai/)™ — 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.

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](https://blog.cloudera.com/cloudera-introduces-ai-inference-service-with-nvidia-nim/), [Cohesity](https://www.cohesity.com/press/unlock-gen-ai-capabilities-via-nvidia-collaboration/), [DataStax](https://www.datastax.com/press-release/datastax-to-deliver-high-performance-rag-solution-with-20x-faster-embeddings-and-indexing-at-80-lower-cost-using-nvidia-microservices), [NetApp](https://www.netapp.com/newsroom/press-releases/news-rel-20240514-813887/), Scale AI and [Synopsys](https://news.synopsys.com/2024-03-18-Synopsys-Showcases-EDA-Performance-and-Next-Gen-Capabilities-with-NVIDIA-Accelerated-Computing,-Generative-AI-and-Omniverse) — 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](https://huggingface.co/blog/train-dgx-cloud) is now offering NIM — starting with [Meta Llama 3](https://ai.meta.com/blog/meta-llama-3/).

Healthcare

Dozens of healthcare companies are [deploying NIM](https://blogs.nvidia.com/blog/llama-3-nim-healthcare-generative-ai) to power generative AI inference across a range of applications, including surgical planning, digital assistants, drug discovery and clinical trial optimization.

Leading AI tools and MLOps partners — including Amazon SageMaker, Microsoft Azure AI, Dataiku, DataRobot, [deepset](https://haystack.deepset.ai/blog/haystack-nvidia-nim-rag-guide" \o "" \t "_blank), Domino Data Lab, [LangChain](https://blog.langchain.dev/nvidia-nim/" \o "" \t "_blank), [Llama Index](https://www.llamaindex.ai/blog/llamaindex-accelerates-enterprise-generative-ai-with-nvidia-nim), [Replicate](https://replicate.com/blog/run-nvidia-nim-models-on-replicate), Run.ai, [Saturn Cloud](https://saturncloud.io/blog/deploying-generative-ai-on-saturn-cloud-using-nvidia-inference-microservices/?utm_source=nvidia&utm_medium=press_release&utm_campaign=computexpr), Securiti AI and [Weights & Biases](https://wandb.ai/site/press-release/integration-nvidia-nim) — 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](https://www.latentview.com/press-release/latentview-analytics-accelerates-enterprise-solutions-with-nvidia-ai-enterprise/), [Quantiphi](https://quantiphi.com/media-releases/quantiphi-unveils-generative-ai-platform-baioniqtm-integrated-with-nvidia-ai-enterprise-software-empowering-enterprise-productivity/" \o "" \t "_blank), 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](https://www.nvidia.com/en-us/data-center/products/certified-systems/)™ from global infrastructure manufacturers Cisco, [Dell Technologies](https://www.dell.com/en-us/dt/corporate/newsroom/announcements/detailpage.press-releases~usa~2024~05~20240520-dell-technologies-expands-dell-ai-factory-with-nvidia-to-turbocharge-ai-adoption.htm), [Hewlett-Packard Enterprise](https://www.hpe.com/us/en/newsroom/press-release/2024/03/hewlett-packard-enterprise-debuts-end-to-end-ai-native-portfolio-for-generative-ai.html), [Lenovo](https://news.lenovo.com/netapp-and-lenovo-offer-converged-infrastructure-solution-optimized-for-genai/) and Supermicro, as well as server manufacturers [ASRock Rack](https://www.asrockrack.com/general/news.asp?id=239), [ASUS](https://servers.asus.com/NEWS/ASUS-Presents-ESC-AI-POD-with-NVIDIA-GB200-NVL72-at-Computex-2024), [GIGABYTE](https://www.gigabyte.com/Press/News/2168), [Ingrasys](https://www.foxconn.com.tw/en-us/press-center/press-releases/latest-news), [Inventec](https://ebg.inventec.com/en/news/Press%20Release/2024/85" \o "" \t "_blank), [Pegatron](https://svr.pegatroncorp.com/News/6), QCT, Wistron and Wiwynn. NIM microservices have also been integrated into [Amazon Web Services](https://aws.amazon.com/about-aws/whats-new/2024/03/amazon-sagemaker-integration-nvidia-nim-microservices/), [Google Cloud](https://nvidianews.nvidia.com/news/google-cloud-ai-development), [Azure](https://news.microsoft.com/2024/03/18/microsoft-and-nvidia-announce-major-integrations-to-accelerate-generative-ai-for-enterprises-everywhere/) and [Oracle Cloud Infrastructure](https://nvidianews.nvidia.com/news/oracle-nvidia-sovereign-ai).