NVIDIA Releases NIM Microservices to Safeguard Applications for Agentic AI

NVIDIA NeMo Guardrails includes new NVIDIA NIM microservices to enhance accuracy, security and control for enterprises building AI across industries.

January 16, 2025 by Kari Briski

Share

Reading Time: 5 mins

AI agents are poised to transform productivity for the world's billion knowledge workers with "knowledge robots" that can accomplish a variety of tasks. To develop AI agents, enterprises need to address critical concerns like trust, safety, security and compliance.

New NVIDIA NIM microservices for AI guardrails — part of the NVIDIA NeMo Guardrails collection of software tools — are portable, optimized inference microservices that help companies improve the safety, precision and scalability of their generative AI applications.

Central to the orchestration of the microservices is NeMo Guardrails, part of the NVIDIA NeMo platform for curating, customizing and guardrailing AI. NeMo Guardrails helps developers integrate and manage AI guardrails in large language model (LLM) applications. Industry leaders Amdocs, Cerence AI and Lowe's are among those using NeMo Guardrails to safeguard AI applications.

Developers can use the NIM microservices to build more secure, trustworthy AI agents that provide safe, appropriate responses within context-specific guidelines and are bolstered against jailbreak attempts. Deployed in customer service across industries like automotive, finance, healthcare, manufacturing and retail, the agents can boost customer satisfaction and trust.

One of the new microservices, built for moderating content safety, was trained using the Aegis Content Safety Dataset — one of the highest-quality, human-annotated data sources in its category. Curated and owned by NVIDIA, the dataset is publicly available on Hugging Face and includes over 35,000 human-annotated data samples flagged for AI safety and jailbreak attempts to bypass system restrictions.

NVIDIA NeMo Guardrails Keeps AI Agents on Track

AI is rapidly boosting productivity for a broad range of business processes. In customer service, it's helping resolve customer issues up to 40% faster. However, scaling AI for customer service and other AI agents requires secure models that prevent harmful or inappropriate outputs and ensure the AI application behaves within defined parameters.

NVIDIA has introduced three new NIM microservices for NeMo Guardrails that help AI agents operate at scale while maintaining controlled behavior:

- **Content safety NIM microservice** that safeguards AI against generating biased or harmful outputs, ensuring responses align with ethical standards.
- **Topic control NIM microservice** that keeps conversations focused on approved topics, avoiding digression or inappropriate content.
- **Jailbreak detection NIM microservice** that adds protection against jailbreak attempts, helping maintain AI integrity in adversarial scenarios.

By applying multiple lightweight, specialized models as guardrails, developers can cover gaps that may occur when only more general global policies and protections exist — as a one-size-fits-all approach doesn't properly secure and control complex agentic AI workflows.

Small language models, like those in the NeMo Guardrails collection, offer lower latency and are designed to run efficiently, even in resource-constrained or distributed environments. This makes them ideal for scaling AI applications in industries such as healthcare, automotive and manufacturing, in locations like hospitals or warehouses.

Industry Leaders and Partners Safeguard AI With NeMo Guardrails

NeMo Guardrails, available to the open-source community, helps developers orchestrate multiple AI software policies — called rails — to enhance LLM application security and control. It works with NVIDIA NIM microservices to offer a robust framework for building AI systems that can be deployed at scale without compromising on safety or performance.

Amdocs, a leading global provider of software and services to communications and media companies, is harnessing NeMo Guardrails to enhance AI-driven customer interactions by delivering safer, more accurate and contextually appropriate responses.

"Technologies like NeMo Guardrails are essential for safeguarding generative AI applications, helping make sure they operate securely and ethically," said Anthony Goonetilleke, group president of technology and head of strategy at Amdocs. "By integrating NVIDIA NeMo Guardrails into our amAIz platform, we are enhancing the platform's 'Trusted AI' capabilities to deliver agentic experiences that are safe, reliable and scalable. This empowers service providers to deploy AI solutions safely and with confidence, setting new standards for AI innovation and operational excellence."

Cerence AI, a company specializing in AI solutions for the automotive industry, is using NVIDIA NeMo Guardrails to help ensure its in-car assistants deliver contextually appropriate, safe interactions powered by its CaLLM family of large and small language models.

"Cerence AI relies on high-performing, secure solutions from NVIDIA to power our in-car assistant technologies," said Nils Schanz, executive vice president of product and technology at Cerence AI. "Using NeMo Guardrails helps us deliver trusted, context-aware solutions to our automaker customers and provide sensible, mindful and hallucination-free responses. In addition, NeMo Guardrails is customizable for our automaker customers and helps us filter harmful or unpleasant requests, securing our CaLLM family of language models from unintended or inappropriate content delivery to end users."

Lowe's, a leading home improvement retailer, is leveraging generative AI to build on the deep expertise of its store associates. By providing enhanced access to comprehensive product knowledge, these tools empower associates to answer customer questions, helping them find the right products to complete their projects and setting a new standard for retail innovation and customer satisfaction.

"We're always looking for ways to help associates to above and beyond for our customers," said Chandhu Nair, senior vice president of data, AI and innovation at Lowe's. "With our recent deployments of NVIDIA NeMo Guardrails, we ensure AI-generated responses are safe, secure and reliable, enforcing conversational boundaries to deliver only relevant and appropriate content."

To further accelerate AI safeguards adoption in AI application development and deployment in retail, NVIDIA recently announced at the NRF show that its NVIDIA AI Blueprint for retail shopping assistants incorporates NeMo Guardrails microservices for creating more reliable and controlled customer interactions during digital shopping experiences.

Consulting leaders Taskus, Tech Mahindra and Wipro are also integrating NeMo Guardrails into their solutions to provide their enterprise clients safer, more reliable and controlled generative AI applications.

NeMo Guardrails is open and extensible, offering integration with a robust ecosystem of leading AI safety model and guardrail providers, as well as AI observability and development tools. It supports integration with ActiveFence's ActiveScore, which filters harmful or inappropriate content in conversational AI applications, and provides visibility, analytics and monitoring.

Hive, which provides its AI-generated content detection models for images, video and audio content as NIM microservices, can be easily integrated and orchestrated in AI applications using NeMo Guardrails.

The Fiddler AI Observability platform easily integrates with NeMo Guardrails to enhance AI guardrail monitoring capabilities. And Weights & Biases, an end-to-end AI developer platform, is expanding the capabilities of W&B Weave by adding integrations with NeMo Guardrails microservices. This enhancement builds on Weights & Biases' existing portfolio of NIM integrations for optimized AI inferencing in production.

NeMo Guardrails Offers Open-Source Tools for AI Safety Testing

Developers ready to test the effectiveness of applying safeguard models and other rails can use NVIDIA Garak — an open-source toolkit for LLM and application vulnerability scanning developed by the NVIDIA Research team.

With Garak, developers can identify vulnerabilities in systems using LLMs by assessing them for issues such as data leaks, prompt injections, code hallucination and jailbreak scenarios. By generating test cases involving inappropriate or incorrect outputs, Garak helps developers detect and address potential weaknesses in AI models to enhance their robustness and safety.

Availability

NVIDIA NeMo Guardrails microservices, as well as NeMo Guardrails for rail orchestration and the NVIDIA Garak toolkit, are now available for developers and enterprises. Developers can get started building AI safeguards into AI agents for customer service using NeMo Guardrails with this tutorial.

See notice regarding software product information.

Categories: Generative AI

Tags: Artificial Intelligence | Cybersecurity | NVIDIA Blueprints | NVIDIA NeMo | NVIDIA NIM

All NVIDIA News

How This NVIDIA Kaggle Grandmaster Merges Innovation and Play

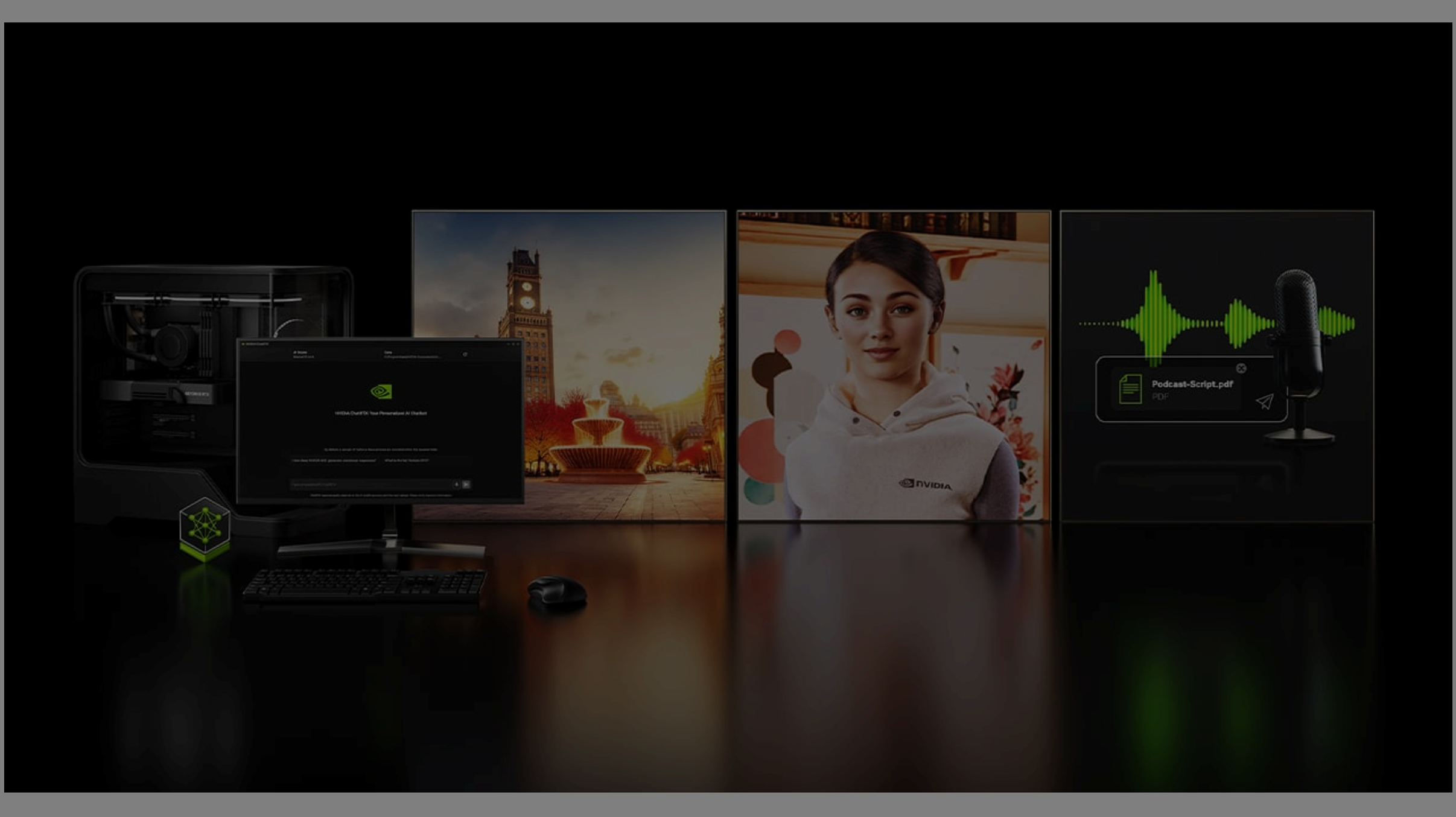
Fantastic Four-ce Awakens: Season One of 'Marvel Rivals' Joins GeForce NOW

How AI Is Enhancing Surgical Safety and Education

Unveiling a New Era of Local AI With NVIDIA NIM Microservices and AI Blueprints

New NIM microservices and AI Blueprints unlock generative AI on RTX AI PCs and workstation — plus, more announcements from CES recapped in this first installment of the RTX AI Garage series.

January 8, 2025 by Jesse Clayton



Share

Reading Time: 5 mins

Over the past year, generative AI has transformed the way people live, work and play, enhancing everything from writing and content creation to gaming, learning and productivity. PC enthusiasts and developers are leading the charge in pushing the boundaries of this groundbreaking technology.

Countless times, industry-defining technological breakthroughs have been invented in one place — a garage. This week marks the start of the *RTX AI Garage* series, which will offer routine content for developers and enthusiasts looking to learn more about NVIDIA NIM microservices and AI Blueprints, and how to build AI agents, creative workflow, digital human, productivity apps and more on AI PCs. Welcome to the *RTX AI Garage*.

This first installment spotlights announcements made earlier this week at CES, including new AI foundation models available on NVIDIA RTX AI PCs that take digital humans, content creation, productivity and development to the next level.

These models — offered as NVIDIA NIM microservices — are powered by new GeForce RTX 50 Series GPUs. Built on the NVIDIA Blackwell architecture, RTX 50 Series GPUs deliver up to 3,352 trillion AI operations per second of performance, 32GB of VRAM and feature FP4 compute, doubling AI inference performance and enabling generative AI to run locally with a smaller memory footprint.

NVIDIA also introduced NVIDIA AI Blueprints — ready-to-use, preconfigured workflows, built on NIM microservices, for applications like digital humans and content creation.

NIM microservices and AI Blueprints empower enthusiasts and developers to build, iterate and deliver AI-powered experiences to the PC faster than ever. The result is a new wave of compelling, practical capabilities for PC users.

Fast-Track AI With NVIDIA NIM

There are two key challenges to bringing AI advancements to PCs. First, the pace of AI research is breakneck, with new models appearing daily on platforms like Hugging Face, which now hosts over a million models. As a result, breakthroughs quickly become outdated.

Second, adapting these models for PC use is a complex, resource-intensive process. Optimizing them for PC hardware, integrating them with AI software and connecting them to applications requires significant engineering effort.

NVIDIA NIM helps address these challenges by offering prepackaged, state-of-the-art AI models optimized for PCs. These NIM microservices span model domains, can be installed with a single click, feature application programming interfaces (APIs) for easy integration, and harness NVIDIA AI software and RTX GPUs for accelerated performance.

At CES, NVIDIA announced a pipeline of NIM microservices for RTX AI PCs, supporting use cases spanning large language models (LLMs), vision-language models, image generation, speech, retrieval-augmented generation (RAG), PDF extraction and computer vision.

The new Llama Nemotron family of open models provide high accuracy on a wide range of agentic tasks. The Llama Nemotron Nano model, which will be offered as a NIM microservice for RTX AI PCs and workstations, excels at agentic AI tasks like instruction following, function calling, chat, coding and math.

Soon, developers will be able to quickly download and run these microservices on Windows 11 PCs using Windows Subsystem for Linux (WSL).

Introducing Project R2X | A Preview of a RTX-Powered Digital Human Interface

To demonstrate how enthusiasts and developers can use NIM to build AI agents and assistants, NVIDIA previewed Project R2X, a vision-enabled PC avatar that can put information at a user's fingertips, assist with desktop apps and video conference calls, read and summarize documents, and more. Sign up for Project R2X updates.

By using NIM microservices, AI enthusiasts can skip the complexities of model curation, optimization and backend integration and focus on creating and innovating with cutting-edge AI models.

What's in an API?

An API is the way in which an application communicates with a software library. An API defines a set of "calls" that the application can make to the library and what the application can expect in return. Traditional AI APIs require a lot of setup and configuration, making AI capabilities harder to use and hampering innovation.

NIM microservices expose easy-to-use, intuitive APIs that an application can simply send requests to and get a response. In addition, they're designed around the input and output media for different model types. For example, LLMs take text as input and produce text as output, image generators convert text to image, speech recognizers turn speech to text and so on.

The microservices are designed to integrate seamlessly with leading AI development and agent frameworks such as AI Toolkit for VSCode, AnythingLLM, ComfyUI, Flowise AI, LangChain, Langflow and LM Studio. Developers can easily download and deploy them from build.nvidia.com.

By bringing these APIs to RTX, NVIDIA NIM will accelerate AI innovation on PCs.

Enthusiasts are expected to be able to experience a range of NIM microservices using an upcoming release of the NVIDIA ChatRTX tech demo.

A Blueprint for Innovation

By using state-of-the-art models, prepackaged and optimized for PCs, developers and enthusiasts can quickly create AI-powered projects. Taking things a step further, they can combine multiple AI models and other functionality to build complex applications like digital humans, podcast generators and application assistants.

NVIDIA AI Blueprints, built on NIM microservices, are reference implementations for complex AI workflows. They help developers connect several components, including libraries, software development kits and AI models, together in a single application.

NVIDIA NIM Microservices for RTX AI PCs

AI Blueprints include everything that a developer needs to build, run, customize and extend the reference workflow, which includes the reference application and source code, sample data, and documentation for customization and orchestration of the different components.

At CES, NVIDIA announced two AI Blueprints for RTX: one for PDF to podcast, which lets users generate a podcast from any PDF, and another for 3D-guided generative AI, which is based on FLUX.1 [dev] and expected be offered as a NIM microservice, offers artists greater control over text-based image generation.

With AI Blueprints, developers can quickly go from AI experimentation to AI development for cutting-edge workflows on RTX PCs and workstations.

Built for Generative AI

The new GeForce RTX 50 Series GPUs are purpose-built to tackle complex generative AI challenges, featuring fifth-generation Tensor Cores with FP4 support, faster G7 memory and an AI-management processor for efficient multitasking between AI and creative workflows.

The GeForce RTX 50 Series adds FP4 support to help bring better performance and more models to PCs. FP4 is a lower quantization method, similar to file compression, that decreases model sizes. Compared with FP16 — the default method that most models feature — FP4 uses less than half of the memory, and 50 Series GPUs provide over 2x performance compared with the previous generation. This can be done with virtually no loss in quality with advanced quantization methods offered by NVIDIA TensorRT Model Optimizer.

For example, Black Forest Labs' FLUX.1 [dev] model at FP16 requires over 23GB of VRAM, meaning it can only be supported by the GeForce RTX 4090 and professional GPUs. With FP4, FLUX.1 [dev] requires less than 10GB, so it can run locally on more GeForce RTX GPUs.

With a GeForce RTX 4090 with FP16, the FLUX.1 [dev] model can generate images in 15 seconds with 30 steps. With a GeForce RTX 5090 with FP4, images can be generated in just over five seconds.

Get Started With the New AI APIs for PCs

NVIDIA NIM microservices and AI Blueprints are expected to be available starting next month, with initial hardware support for GeForce RTX 50 Series, GeForce RTX 4090 and 4080, and NVIDIA RTX 6000 and 5000 professional GPUs. Additional GPUs will be supported in the future.

NIM-ready RTX AI PCs are expected to be available from Acer, ASUS, Dell, GIGABYTE, HP, Lenovo, MSI, Razer and Samsung, and from local system builders Corsair, Falcon Northwest, LDLC, Maingear, Mifcon, Origin PC, PCS and Scan.

GeForce RTX 50 Series GPUs and laptops deliver game-changing performance, power transformative AI experiences, and enable creators to complete workflows in record time. Rewatch NVIDIA CEO Jensen Huang's keynote to learn more about NVIDIA's AI news unveiled at CES.

See notice regarding software product information.

Categories: Generative AI

Tags: AI Decoded | Artificial Intelligence | GeForce | NVIDIA RTX

All NVIDIA News

How This NVIDIA Kaggle Grandmaster Merges Innovation and Play

NVIDIA Releases NIM Microservices to Safeguard Applications for Agentic AI

Fantastic Four-ce Awakens: Season One of 'Marvel Rivals' Joins GeForce NOW

How AI Is Enhancing Surgical Safety and Education