NVIDIA Launches AI Foundation Models for RTX AI PCs

NVIDIA NIM Microservices and AI Blueprints Help Developers and Enthusiasts Build AI Agents and Creative Workflows on PC

CES—NVIDIA today announced foundation models running locally on NVIDIA RTX™ AI PCs that supercharge digital humans, content creation, productivity and development.

These models — offered as NVIDIA NIM™ microservices — are accelerated by new GeForce RTX™ 50 Series GPUs, which feature up to 3,352 trillion operations per second of AI performance and 32GB of VRAM. Built on the NVIDIA Blackwell architecture, RTX 50 Series are the first consumer GPUs to add support for FP4 compute, boosting AI inference performance by 2x and enabling generative AI models to run locally in a smaller memory footprint, compared with previous-generation hardware.

GeForce™ has long been a vital platform for AI developers. The first GPU-accelerated deep learning network, AlexNet, was trained on the GeForce GTX™ 580 in 2012 — and last year, over 30% of published AI research papers cited the use of GeForce RTX.

Now, with generative AI and RTX AI PCs, anyone can be a developer. A new wave of low-code and no-code tools, such as AnythingLLM, ComfyUI, Langflow and LM Studio, enable enthusiasts to use AI models in complex workflows via simple graphical user interfaces.

NIM microservices connected to these GUIs will make it effortless to access and deploy the latest generative AI models. NVIDIA AI Blueprints, built on NIM microservices, provide easy-to-use, preconfigured reference workflows for digital humans, content creation and more.

To meet the growing demand from AI developers and enthusiasts, every top PC manufacturer and system builder is launching NIM-ready RTX AI PCs with GeForce RTX 50 Series GPUs.

“AI is advancing at light speed, from perception AI to generative AI and now agentic AI,” said Jensen Huang, founder and CEO of NVIDIA. “NIM microservices and AI Blueprints give PC developers and enthusiasts the building blocks to explore the magic of AI.”

Making AI NIMble
Foundation models — neural networks trained on immense amounts of raw data — are the building blocks for generative AI.

NVIDIA will release a pipeline of NIM microservices for RTX AI PCs from top model developers such as Black Forest Labs, Meta, Mistral and Stability AI. Use cases span large language models (LLMs), vision language models, image generation, speech, embedding models for retrieval-augmented generation (RAG), PDF extraction and computer vision.

“GeForce RTX 50 Series GPUs with FP4 compute will unlock a massive range of models that can run on PC, which were previously limited to large data centers,” said Robin Rombach, CEO of Black Forest Labs. “Making FLUX an NVIDIA NIM microservice increases the rate at which AI can be deployed and experienced by more users, while delivering incredible performance.”

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

NIM microservices include the key components for running AI on PCs and are optimized for deployment across NVIDIA GPUs — whether in RTX PCs and workstations or in the cloud.

Developers and enthusiasts will be able to quickly download, set up and run these NIM microservices on Windows 11 PCs with Windows Subsystem for Linux (WSL).

“AI is driving Windows 11 PC innovation at a rapid rate, and Windows Subsystem for Linux (WSL) offers a great cross-platform environment for AI development on Windows 11 alongside Windows Copilot Runtime,” said Pavan Davuluri, corporate vice president of Windows at Microsoft. “NVIDIA NIM microservices, optimized for Windows PCs, give developers and enthusiasts ready-to-integrate AI models for their Windows apps, further accelerating deployment of AI capabilities to Windows users.”

The NIM microservices, running on RTX AI PCs, will be compatible with top AI development and agent frameworks, including AI Toolkit for VSCode, AnythingLLM, ComfyUI, CrewAI, Flowise AI, LangChain, Langflow and LM Studio. Developers can connect applications and workflows built on these frameworks to AI models running NIM microservices

through industry-standard endpoints, enabling them to use the latest technology with a unified interface across the cloud, data centers, workstations and PCs.

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

Putting a Face on Agentic AI
To demonstrate how enthusiasts and developers can use NIM to build AI agents and assistants, NVIDIA today 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.

The avatar is rendered using NVIDIA RTX Neural Faces, a new generative AI algorithm that augments traditional rasterization with entirely generated pixels. The face is then animated by a new diffusion-based NVIDIA Audio2Face™-3D model that improves lip and tongue movement. R2X can be connected to cloud AI services such as OpenAI's GPT4o and xAI's Grok, and NIM microservices and AI Blueprints, such as PDF retrievers or alternative LLMs, via developer frameworks such as CrewAI, Flowise AI and Langflow. Sign up for Project R2X updates.

AI Blueprints Coming to PC
NIM microservices are also available to PC users through AI Blueprints — reference AI workflows that can run locally on RTX PCs. With these blueprints, developers can create podcasts from PDF documents, generate stunning images guided by 3D scenes and more.

The blueprint for PDF to podcast extracts text, images and tables from a PDF to create a podcast script that can be edited by users. It can also generate a full audio recording from the script using voices available in the blueprint or based on a user's voice sample. In addition, users can have a real-time conversation with the AI podcast host to learn more about specific topics.

The blueprint uses NIM microservices like Mistral-Nemo-12B-Instruct for language, NVIDIA Riva for text-to-speech and automatic speech recognition, and the NeMo Retriever collection of microservices for PDF extraction.

The AI Blueprint for 3D-guided generative AI gives artists finer control over image generation. While AI can generate amazing images from simple text prompts, controlling image composition using only words can be challenging. With this blueprint, creators can use simple 3D objects laid out in a 3D renderer like Blender to guide AI image generation. The artist can create 3D assets by hand or generate them using AI, place them in the scene and set the 3D viewport camera. Then, a prepackaged workflow powered by the FLUX NIM microservice will use the current composition to generate high-quality images that match the 3D scene.

NVIDIA NIM microservices and AI Blueprints will be available starting in February 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 will 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.

Learn more about how NIM microservices, AI Blueprints and NIM-ready RTX AI PCs are accelerating generative AI by joining NVIDIA at CES.

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, and availability of our products, services, and technologies, including NVIDIA RTX AI PCs, GeForce RTX 50 Series GPUs, NVIDIA Blackwell architecture, GeForce GTX 580, Project R2X, NVIDIA ACE and NIM microservices, NVIDIA AI Blueprints, NVIDIA Grace Blackwell platform, Llama Nemotron, NVIDIA ChatRTX, NVIDIA RTX Neural Faces, NVIDIA Audio2Face-3D model, Mistral-Nemo-12B-Instruct for language, NVIDIA Riva, NeMo Retriever, FLUX NIM microservice, GeForce RTX 4090 and 4080, and NVIDIA RTX 6000 and 5000 professional GPUs third parties using or adopting NVIDIA's products and technologies, and the benefits and impact thereof; and AI advancing at light speed, from perception AI to generative AI and now agentic AI 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.

Jordan Dodge
SHIELD, GeForce NOW
NVIDIA Corp.
+1-408-506-6849
jdodge@nvidia.com