NVIDIA Brings Generative AI to Millions, With Tensor Core GPUs, LLMs, Tools for RTX PCs and Workstations

Leading AI Platform Gets RTX-Accelerated Boost From New GeForce RTX SUPER GPUs, AI Laptops From Every Top Manufacturer

CES — NVIDIA today announced GeForce RTX™ SUPER desktop GPUs for supercharged generative AI performance, new AI laptops from every top manufacturer, and new NVIDIA RTX™-accelerated AI software and tools for both developers and consumers.

Building on decades of PC leadership, with over 100 million of its RTX GPUs driving the AI PC era, NVIDIA is now offering these tools to enhance PC experiences with generative AI: NVIDIA TensorRT™ acceleration of the popular Stable Diffusion XL model for text-to-image workflows, NVIDIA RTX Remix with generative AI texture tools, NVIDIA ACE microservices and more games that use DLSS 3 technology with Frame Generation.

AI Workbench, a unified, easy-to-use toolkit for AI developers, will be available in beta later this month. In addition, NVIDIA TensorRT-LLM (TRT-LLM), an open-source library that accelerates and optimizes inference performance of the latest large language models (LLMs), now supports more pre-optimized models for PCs. Accelerated by TRT-LLM, Chat with RTX, an NVIDIA tech demo also releasing this month, allows AI enthusiasts to interact with their notes, documents and other content.

“Generative AI is the single most significant platform transition in computing history and will transform every industry, including gaming,” said Jensen Huang, founder and CEO of NVIDIA. “With over 100 million RTX AI PCs and workstations, NVIDIA is a massive installed base for developers and gamers to enjoy the magic of generative AI.”

Running generative AI locally on a PC is critical for privacy, latency and cost-sensitive applications. It requires a large installed base of AI-ready systems, as well as the right developer tools to tune and optimize AI models for the PC platform.

To meet these needs, NVIDIA is delivering innovations across its full technology stack, driving new experiences and building on the 500+ AI-enabled PC applications and games already accelerated by NVIDIA RTX technology.

RTX AI PCs and Workstations
NVIDIA RTX GPUs — capable of running a broad range of applications at the highest performance — unlock the full potential of generative AI on PCs. Tensor Cores in these GPUs dramatically speed AI performance across the most demanding applications for work and play.

The new GeForce RTX 40 SUPER Series graphics cards, also announced today at CES, include the GeForce RTX 4080 SUPER, 4070 Ti SUPER and 4070 SUPER for top AI performance. The GeForce RTX 4080 SUPER generates AI video 1.5x faster — and images 1.7x faster — than the GeForce RTX 3080 Ti GPU. The Tensor Cores in SUPER GPUs deliver up to 836 trillion operations per second, bringing transformative AI capabilities to gaming, creating and everyday productivity.

Leading manufacturers — including Acer, ASUS, Dell, HP, Lenovo, MSI, Razer and Samsung — are releasing a new wave of RTX AI laptops, bringing a full set of generative AI capabilities to users right out of the box. The new systems, which deliver a performance increase ranging from 20x-60x compared with using neural processing units, will start shipping this month.

Mobile workstations with RTX GPUs can run NVIDIA AI Enterprise software, including TensorRT and NVIDIA RAPIDS™ for simplified, secure generative AI and data science development. A three-year license for NVIDIA AI Enterprise is included with every NVIDIA A800 40GB Active GPU, providing an ideal workstation development platform for AI and data science.

New PC Developer Tools for Building AI Models
To help developers quickly create, test and customize pretrained generative AI models and LLMs using PC-class performance and memory footprint, NVIDIA recently announced NVIDIA AI Workbench.

AI Workbench, which will be available in beta later this month, offers streamlined access to popular repositories like Hugging Face, GitHub and NVIDIA NGC™, along with a simplified user interface that enables developers to easily reproduce, collaborate on and migrate projects.

Projects can be scaled out to virtually anywhere — whether the data center, a public cloud or NVIDIA DGX™ Cloud — and then brought back to local RTX systems on a PC or workstation for inference and light customization.

In collaboration with HP, NVIDIA is also simplifying AI model development by integrating NVIDIA AI Foundation Models and Endpoints, which include RTX-accelerated AI models and software development kits, into the HP AI Studio, a centralized platform for data science. This will allow users to easily search, import and deploy optimized models across PCs and the

cloud.

After building AI models for PC use cases, developers can optimize them using NVIDIA TensorRT to take full advantage of RTX GPUs' Tensor Cores.

NVIDIA recently extended TensorRT to text-based applications with TensorRT-LLM for Windows, an open-source library for accelerating LLMs. The latest update to TensorRT-LLM, available now, adds Phi-2 to the growing list of pre-optimized models for PC, which run up to 5x faster compared to other inference backends.

RTX-Accelerated Generative AI Powers New PC Experiences
At CES, NVIDIA and its developer partners are releasing new generative AI-powered applications and services for PCs, including:

- NVIDIA RTX Remix, a platform for creating stunning RTX remasters of classic games. Releasing in beta later this month, it delivers generative AI tools that can transform basic textures from classic games into modern, 4K-resolution, physically based rendering materials.
- NVIDIA ACE microservices, including generative AI-powered speech and animation models, which enable developers to add intelligent, dynamic digital avatars to games.
- TensorRT acceleration for Stable Diffusion XL (SDXL) Turbo and latent consistency models, two of the most popular Stable Diffusion acceleration methods. TensorRT improves performance for both by up to 60% compared with the previous fastest implementation. An updated version of the Stable Diffusion WebUI TensorRT extension is also now available, including acceleration for SDXL, SDXL Turbo, LCM - Low-Rank Adaptation (LoRA) and improved LoRA support.
- NVIDIA DLSS 3 with Frame Generation, which uses AI to increase frame rates up to 4x compared with native rendering, will be featured in a dozen of the 14 new RTX games announced, including *Horizon Forbidden West*, *Pax Dei* and *Dragon's Dogma 2*.
- Chat with RTX, an NVIDIA tech demo available later this month, allows AI enthusiasts to easily connect PC LLMs to their own data using a popular technique known as retrieval-augmented generation (RAG). The demo, accelerated by TensorRT-LLM, enables users to quickly interact with their notes, documents and other content. It will also be available as an open-source reference project, so developers can easily implement the same capabilities in their own applications.

Learn more about the latest generative AI breakthroughs by joining NVIDIA at CES.

About NVIDIA
Since its founding in 1993, NVIDIA (NASDAQ: NVDA) has been a pioneer in accelerated computing. The company's invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, ignited the era of modern AI and is fueling industrial digitalization across markets. NVIDIA is now a full-stack computing company with data-center-scale offerings that are reshaping industry. More information at https://nvidianews.nvidia.com/.

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 GeForce RTX SUPER desktop GPUs, NVIDIA RTX-accelerated AI software and tools, NVIDIA TensorRT, NVIDIA RTX Remix, NVIDIA ACE, NVIDIA DLSS 3, Chat with RTX, GeForce RTX 40 SUPER Series graphics cards, NVIDIA RAPIDS, NVIDIA AI Workbench; our products enhancing PC experience with generative AI; NVIDIA delivering innovations across its full technology stack, driving new experiences and building on the 500+ AI-enabled PC applications and games already accelerated by NVIDIA RTX technology; leading manufacturers releasing a new wave of RTX AI laptops and bringing a full set of generative AI capabilities to users; the ability of mobile workstations with RTX GPUs to run NVIDIA AI Enterprise software; our collaborations with third parties; and developers optimizing AI models by using our products 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