NVIDIA AI Workbench Speeds Adoption of Custom Generative AI for World’s Enterprises

New Developer Toolkit Introduces Simplified Model Tuning and Deployment on NVIDIA AI Platforms — From PCs and Workstations to Enterprise Data Centers, Public Clouds and NVIDIA DGX Cloud

SIGGRAPH—NVIDIA today announced NVIDIA AI Workbench, a unified, easy-to-use toolkit that allows developers to quickly create, test and customize pretrained generative AI models on a PC or workstation — then scale them to virtually any data center, public cloud or NVIDIA DGX™ Cloud.

AI Workbench removes the complexity of getting started with an enterprise AI project. Accessed through a simplified interface running on a local system, it allows developers to customize models from popular repositories like Hugging Face, GitHub and NVIDIA NGC™ using custom data. The models can then be shared easily across multiple platforms.

“Enterprises around the world are racing to find the right infrastructure and build generative AI models and applications,” said Manuvir Das, vice president of enterprise computing at NVIDIA. “NVIDIA AI Workbench provides a simplified path for cross-organizational teams to create the AI-based applications that are increasingly becoming essential in modern business.”

A New Era for AI Developers

While hundreds of thousands of pretrained models are now available, customizing them with the many open-source tools can require hunting through multiple online repositories for the right framework, tools and containers, and employing the right skills to customize a model for a specific use case.

With NVIDIA AI Workbench, developers can customize and run generative AI in just a few clicks. It allows them to pull together all necessary enterprise-grade models, frameworks, software development kits and libraries from open-source repositories and the NVIDIA AI platform into a unified developer toolkit.

Leading AI infrastructure providers — including Dell Technologies, Hewlett Packard Enterprise, HP Inc., Lambda, Lenovo and Supermicro — are embracing AI Workbench for its ability to augment their latest generation of multi-GPU-capable desktop workstations, high-end mobile workstations and virtual workstations.

Developers with a Windows or Linux-based NVIDIA RTX™ PC or workstation will also be able to initiate, test and fine-tune enterprise-grade generative AI projects on their local RTX systems, and easily access data center and cloud computing resources to scale as needed.

New NVIDIA AI Enterprise 4.0 Software Advances AI Deployment

To further accelerate the adoption of generative AI, NVIDIA announced the latest version of its enterprise software platform, NVIDIA AI Enterprise 4.0. It gives businesses the tools needed to adopt generative AI, while also offering the security and API stability required for reliable production deployments.

Newly supported software and tools in NVIDIA AI Enterprise that help streamline generative AI deployment include:

- NVIDIA NeMo™, a cloud-native framework to build, customize and deploy large language models. With NeMo, NVIDIA AI Enterprise provides end-to-end support for creating and customizing LLM applications.
- NVIDIA Triton™ Management Service, which helps automate and optimize production deployments. It allows enterprises to automatically deploy multiple NVIDIA Triton Inference Server instances in Kubernetes with model orchestration for efficient operation of scalable AI.
- NVIDIA Base Command Manager Essentials cluster management software, which helps enterprises maximize performance and utilization of AI servers across data center, multi-cloud and hybrid-cloud environments.

NVIDIA AI Enterprise software — which lets users build and run NVIDIA AI-enabled solutions across the cloud, data center and edge — is certified to run on mainstream NVIDIA-Certified Systems™, NVIDIA DGX systems, all major cloud platforms and newly announced NVIDIA RTX workstations.

Leading software companies ServiceNow and Snowflake, as well as infrastructure provider Dell Technologies, which offers Dell Generative AI Solutions, recently announced they are collaborating with NVIDIA to enable new generative AI solutions and services on their platforms. The integration of NVIDIA AI Enterprise 4.0 and NVIDIA NeMo provides a foundation for production-ready generative AI for customers.

NVIDIA AI Enterprise 4.0 will be integrated into partner marketplaces, including AWS Marketplace, Google Cloud and Microsoft Azure, as well as through NVIDIA cloud partner Oracle Cloud Infrastructure.

Additionally, MLOps providers, including Azure Machine Learning, ClearML, Domino Data Lab, Run:AI, and Weights & Biases, are adding seamless integration with the NVIDIA AI platform to simplify production-grade generative AI model development.

Broad Partner Support
"Dell Technologies and NVIDIA are committed to helping enterprises build purpose-built AI models to access the immense opportunity of generative AI. With NVIDIA AI Workbench, developers can take advantage of the full Dell Generative AI Solutions portfolio to customize models on PCs, workstations and data center infrastructure." — *Meghana Patwardhan, vice president of commercial client products at Dell Technologies*

"Most enterprises do not have the expertise, budget and data center resources to manage the high complexity of AI software and systems. We look forward to NVIDIA AI Workbench's potential to simplify generative AI project creation with one-click training and deployment on the HPE GreenLake edge-to-cloud platform." — *Evan Sparks, chief product officer for AI at HPE*

"As a workstation market leader offering the performance and efficiency needed for the most demanding data science and AI models, we have a long history collaborating with NVIDIA. HP is embracing the next generation of high-performance systems, coupled with NVIDIA RTX Ada Generation GPUs and NVIDIA AI Workbench, and bringing the power of generative AI to our enterprise customers and helping move AI workloads between the cloud and locally." — *Jim Nottingham, senior vice president of advanced computing solutions at HP Inc.*

"Lenovo and NVIDIA are helping customers overcome deployment complexities and more easily implement generative AI to deliver transformative services and products to the market. NVIDIA AI Workbench and the Lenovo AI-ready portfolio enable developers to leverage the power of their smart devices and scale across edge-to-cloud infrastructure." — *Rob Herman, vice president and general manager of Lenovo Workstation & Client AI*

"The longstanding VMware and NVIDIA partnership has helped unlock the power of AI for every business by delivering an end-to-end enterprise platform optimized for AI workloads. Together, we are making generative AI more accessible and easier to implement in the enterprise. With AI Workbench, NVIDIA is giving developers a set of powerful tools to help enterprises accelerate gen AI adoption. With the new NVIDIA AI Workbench, development teams can seamlessly move AI workloads from the desktop to production." — *Chris Wolf, vice president of VMware AI Labs*

Watch NVIDIA founder and CEO Jensen Huang's SIGGRAPH keynote address on demand to learn more about NVIDIA AI Workbench and NVIDIA AI Enterprise 4.0.

AI Workbench is coming soon in early access. Sign up to get notified when it is available.

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, features and availability of our products, services and technologies, including NVIDIA AI Workbench, NVIDIA AI Enterprise, NVIDIA AI Enterprise 4.0, NVIDIA NeMo, Triton Management Service and NVIDIA Base Command; enterprises around the world racing to create collaborative, cross-organizational teams and find the right infrastructure to meet surging demand for generative AI; AI-based applications increasingly becoming essential in modern business; major system providers embracing AI Workbench; the adoption of generative AI; third parties partnering with NVIDIA to enable generative AI; NVIDIA AI Enterprise 4.0 being integrated into partner marketplaces; and MLOps providers adding seamless integration with the NVIDIA AI platform to simplify production-grade generative AI model development 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.

Anna Kiachian
Senior PR Manager
NVIDIA Corporation
+1-650-224-9820
akiachian@nvidia.com