NVIDIA and Hugging Face to Connect Millions of Developers to Generative AI Supercomputing

Integration of NVIDIA DGX Cloud in Hugging Face Platform to Speed LLM Training and Tuning, Simplifies Customizing Models for Nearly Every Industry

SIGGRAPH—NVIDIA and Hugging Face today announced a partnership that will put generative AI supercomputing at the fingertips of millions of developers building large language models (LLMs) and other advanced AI applications.

By giving developers access to [NVIDIA DGX™ Cloud](#) AI supercomputing within the Hugging Face platform to train and tune advanced AI models, the combination will help supercharge industry adoption of generative AI using LLMs that are custom-tailored with business data for industry-specific applications, including intelligent chatbots, search and summarization.

"Researchers and developers are at the heart of generative AI that is transforming every industry," said Jensen Huang, founder and CEO of NVIDIA. "Hugging Face and NVIDIA are connecting the world's largest AI community with NVIDIA's AI computing platform in the world's leading clouds. Together, NVIDIA AI computing is just a click away for the Hugging Face community."

As part of the collaboration, Hugging Face will offer a new service — called Training Cluster as a Service — to simplify the creation of new and custom generative AI models for the enterprise. Powered by NVIDIA DGX Cloud, the service will be available in the coming months.

"People around the world are making new connections and discoveries with generative AI tools, and we're still only in the early days of this technology shift," said Clément Delangue, co-founder and CEO of Hugging Face. "Our collaboration will bring NVIDIA's most advanced AI supercomputing to Hugging Face to enable companies to take their AI destiny into their own hands with open source and with speed they need to contribute to what's coming next."

Supercharging LLM Customization and Training Within Hugging Face
The Hugging Face platform lets developers build, train and deploy state-of-the-art AI models using open-source resources. Over 15,000 organizations use Hugging Face, and its community has shared over 250,000 models and 50,000 datasets.

The DGX Cloud integration with Hugging Face will bring one-click access to NVIDIA's multi-node AI supercomputing platform. With DGX Cloud, Hugging Face users will be able to connect to NVIDIA AI supercomputing, providing the software and infrastructure needed to rapidly train and tune foundation models with unique data to drive a new wave of enterprise LLM development. With Training Cluster as a Service, powered by DGX Cloud, companies will be able to leverage their unique data for Hugging Face to create uniquely efficient models in record time.

DGX Cloud Speeds Development and Customization for Massive Models
Each instance of DGX Cloud features eight [NVIDIA H100](#) or [A100](#) 80GB Tensor Core GPUs for a total of 640GB of GPU memory per node. [NVIDIA Networking](#) provides a high-performance, low-latency fabric that ensures workloads can scale across clusters of interconnected systems to meet the performance requirements of advanced AI workloads.

Support from NVIDIA experts is included with DGX Cloud to help customers optimize their models and quickly resolve development challenges.

DGX Cloud infrastructure is [hosted by leading NVIDIA cloud service provider partners](#).

Availability
The NVIDIA DGX Cloud integration with Hugging Face is expected to be available in the coming months.

Watch Huang's [SIGGRAPH keynote address](#) on demand to learn more about NVIDIA DGX Cloud.

About Hugging Face
Hugging Face is the collaboration platform for the machine learning community.

The Hugging Face Hub works as a central place where anyone can share, explore, discover, and experiment with open-source ML. HF empowers the next generation of machine learning engineers, scientists and end users to learn, collaborate and share their work to build an open and ethical AI future together.

With the fast-growing community, some of the most used open-source ML libraries and tools, and a talented science team exploring the edge of tech, Hugging Face is at the heart of the AI revolution.

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 DGX Cloud, NVIDIA AI computing, and NVIDIA Networking; NVIDIA's partnership with Hugging Face, including the benefits and impact thereof; researchers and developers being at the heart of generative AI that is transforming every industry; Hugging Face and NVIDIA connecting the world's largest AI community with NVIDIA's AI computing platform in the world's leading clouds, and NVIDIA AI computing being just a click away for the Hugging Face community; people around the world making new connections and discoveries with generative AI tools; DGX Cloud infrastructure being hosted by leading NVIDIA cloud service provider partners; and the availability of the NVIDIA DGX Cloud integration with Hugging Face 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.

Cliff Edwards
NVIDIA Corporation
+1-415-699-2755
cliffe@nvidia.com