NVIDIA Introduces Generative AI Foundry Service on Microsoft Azure for Enterprises and Startups Worldwide

SAP, Amdocs, Getty Images Among First to Build Custom LLMs With NVIDIA AI Foundation Models, Train on NVIDIA DGX Cloud, Deploy With NVIDIA AI Enterprise Software

Microsoft Ignite—NVIDIA today introduced an AI foundry service to supercharge the development and tuning of custom generative AI applications for enterprises and startups deploying on Microsoft Azure.

The NVIDIA AI foundry service pulls together three elements — a collection of NVIDIA AI Foundation Models, NVIDIA NeMo™ framework and tools, and NVIDIA DGX™ Cloud AI supercomputing services — that give enterprises an end-to-end solution for creating custom generative AI models. Businesses can then deploy their customized models with NVIDIA AI Enterprise software to power generative AI applications, including intelligent search, summarization and content generation.

Industry leaders SAP SE, Amdocs and Getty Images are among the pioneers building custom models using the service.

“Enterprises need custom models to perform specialized skills trained on the proprietary DNA of their company — their data,” said Jensen Huang, founder and CEO of NVIDIA. “NVIDIA’s AI foundry service combines our generative AI model technologies, LLM training expertise and giant-scale AI factory. We built this in Microsoft Azure so enterprises worldwide can connect their custom model with Microsoft’s world-leading cloud services.”

“Our partnership with NVIDIA spans every layer of the Copilot stack — from silicon to software — as we innovate together for this new age of AI,” said Satya Nadella, chairman and CEO of Microsoft. “With NVIDIA’s generative AI foundry service on Microsoft Azure, we’re providing new capabilities for enterprises and startups to build and deploy AI applications on our cloud.”

Industry Leaders Building Tailored, Timely LLMs
NVIDIA’s AI foundry service can be used to customize models for generative AI-powered applications across industries, including enterprise software, telecommunications and media. Once ready to deploy, enterprises can use a technique called retrieval-augmented generation (RAG) to connect their models with their enterprise data and access new insights.

As the first customer of NVIDIA DGX Cloud on Microsoft Azure, SAP plans to use the service and optimized RAG workflow with NVIDIA DGX Cloud and NVIDIA AI Enterprise software running on Azure to help customize and deploy Joule®, its new natural language generative AI copilot.

“Joule draws on SAP’s unique position at the nexus of business and technology, and builds on our relevant, reliable and responsible approach to Business AI,” said Christian Klein, CEO and member of the Executive Board of SAP SE. “In partnership with NVIDIA, Joule can help customers unlock the potential of generative AI for their business by automating time-consuming tasks and quickly analyzing data to deliver more intelligent, personalized experiences.”

Amdocs, a leading provider of software and services to communications and media companies, is optimizing models for the Amdocs amAIz framework to speed adoption of generative AI applications and services for telcos globally.

“Generative AI technology presents an incredible opportunity for service providers to reinvent the way they engage with customers,” said Shuky Sheffer, president and CEO at Amdocs. “Leveraging NVIDIA’s and Microsoft's technology to power the Amdocs amAIz framework will bring new GenAI-powered applications to customers faster and enable them to benefit from the immense potential of generative AI, while also providing enterprise-grade security, reliability and performance.”

Curated, Optimized Models for Custom Generative AI
Customers using the NVIDIA foundry service can choose from several NVIDIA AI Foundation models, including a new family of NVIDIA Nemotron-3 8B models hosted in the Azure AI model catalog. Developers can also access the Nemotron-3 8B models on the NVIDIA NGC™ catalog, as well as community models such as Meta’s Llama 2 models optimized for NVIDIA for accelerated computing, which are also coming soon to the Azure AI model catalog.

Optimized with 8 billion parameters, the Nemotron-3 8B family includes versions tuned for different use cases and have multilingual capabilities for building custom enterprise generative AI applications.

NVIDIA DGX Cloud Now Available on Microsoft Azure Marketplace
NVIDIA DGX Cloud AI supercomputing is available today on Azure Marketplace. It features instances customers can rent, scaling to thousands of NVIDIA Tensor Core GPUs, and comes with NVIDIA AI Enterprise software, including NeMo, to speed LLM customization.

The addition of DGX Cloud on the Azure Marketplace enables Azure customers to use their existing Microsoft Azure

Consumption Commitment credits to speed model development with NVIDIA AI supercomputing and software.

NVIDIA AI Enterprise software is now integrated into Azure Machine Learning, adding NVIDIA's platform of secure, stable and supported AI and data science software. This brings NeMo and NVIDIA Triton Inference Server™ to Azure's enterprise-grade AI service.

NVIDIA AI Enterprise is also available on Azure Marketplace, providing businesses worldwide with broad options for production-ready AI development and deployment of custom generative AI applications.

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 and technologies, including the NVIDIA AI foundry service, NVIDIA NeMo, NVIDIA DGX Cloud, NVIDIA AI Enterprise, and NVIDIA AI Foundation models, including NVIDIA Nemotron-3 8B; the benefits, impact, features and timing of our collaborations or partnerships; and enterprises adopting generative AI to transform their business 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.

Shannon McPhee
NVIDIA Corporation
+1-310-920-9642
smcphee@nvidia.com