NVIDIA Launches Generative AI Microservices for Developers to Create and Deploy Generative AI Copilots Across NVIDIA CUDA GPU Installed Base

- *New Catalog of GPU-Accelerated NVIDIA NIM Microservices and Cloud Endpoints for Pretrained AI Models Optimized to Run on Hundreds of Millions of CUDA-Enabled GPUs Across Clouds, Data Centers, Workstations and PCs*
- *Enterprises Can Use Microservices to Accelerate Data Processing, LLM Customization, Inference, Retrieval-Augmented Generation and Guardrails*
- *Adopted by Broad AI Ecosystem, Including Leading Application Platform Providers Cadence, CrowdStrike, SAP, ServiceNow and More*

GTC—NVIDIA today launched dozens of enterprise-grade generative AI microservices that businesses can use to create and deploy custom applications on their own platforms while retaining full ownership and control of their intellectual property.

Built on top of the NVIDIA CUDA® platform, the catalog of cloud-native microservices includes NVIDIA NIM microservices for optimized inference on more than two dozen popular AI models from NVIDIA and its partner ecosystem. In addition, NVIDIA accelerated software development kits, libraries and tools can now be accessed as NVIDIA CUDA-X™ microservices for retrieval-augmented generation (RAG), guardrails, data processing, HPC and more. NVIDIA also separately announced over two dozen healthcare NIM and CUDA-X microservices.

The curated selection of microservices adds a new layer to NVIDIA's full-stack computing platform. This layer connects the AI ecosystem of model developers, platform providers and enterprises with a standardized path to run custom AI models optimized for NVIDIA's CUDA installed base of hundreds of millions of GPUs across clouds, data centers, workstations and PCs.

Among the first to access the new NVIDIA generative AI microservices available in NVIDIA AI Enterprise 5.0 are leading application, data and cybersecurity platform providers including Adobe, Cadence, CrowdStrike, Getty Images, SAP, ServiceNow, and Shutterstock.

"Established enterprise platforms are sitting on a goldmine of data that can be transformed into generative AI copilots," said Jensen Huang, founder and CEO of NVIDIA. "Created with our partner ecosystem, these containerized AI microservices are the building blocks for enterprises in every industry to become AI companies."

NIM Inference Microservices Speed Deployments From Weeks to Minutes
NIM microservices provide pre-built containers powered by NVIDIA inference software — including Triton Inference Server™ and TensorRT™-LLM — which enable developers to reduce deployment times from weeks to minutes.

They provide industry-standard APIs for domains such as language, speech and drug discovery to enable developers to quickly build AI applications using their proprietary data hosted securely in their own infrastructure. These applications can scale on demand, providing flexibility and performance for running generative AI in production on NVIDIA-accelerated computing platforms.

NIM microservices provide the fastest and highest-performing production AI container for deploying models from NVIDIA, A121, Adept, Cohere, Getty Images, and Shutterstock as well as open models from Google, Hugging Face, Meta, Microsoft, Mistral AI and Stability AI.

ServiceNow today announced that it is using NIM to develop and deploy new domain-specific copilots and other generative AI applications faster and more cost effectively.

Customers will be able to access NIM microservices from Amazon SageMaker, Google Kubernetes Engine and Microsoft Azure AI, and integrate with popular AI frameworks like Deepset, LangChain and LlamaIndex.

CUDA-X Microservices for RAG, Data Processing, Guardrails, HPC
CUDA-X microservices provide end-to-end building blocks for data preparation, customization and training to speed production AI development across industries.

To accelerate AI adoption, enterprises may use CUDA-X microservices including NVIDIA Riva for customizable speech and translation AI, NVIDIA cuOpt™ for routing optimization, as well as NVIDIA Earth-2 for high resolution climate and weather simulations.

NeMo Retriever™ microservices let developers link their AI applications to their business data — including text, images and visualizations such as bar graphs, line plots and pie charts — to generate highly accurate, contextually relevant responses.

With these RAG capabilities, enterprises can offer more data to copilots, chatbots and generative AI productivity tools to elevate accuracy and insight.

Additional NVIDIA NeMo™ microservices are coming soon for custom model development. These include NVIDIA NeMo Curator for building clean datasets for training and retrieval, NVIDIA NeMo Customizer for fine-tuning LLMs with domain-specific data, NVIDIA NeMo Evaluator for analyzing AI model performance, as well as NVIDIA NeMo Guardrails for LLMs.

Ecosystem Supercharges Enterprise Platforms With Generative AI Microservices
In addition to leading application providers, data, infrastructure and compute platform providers across the NVIDIA ecosystem are working with NVIDIA microservices to bring generative AI to enterprises.

Top data platform providers including Box, Cloudera, Cohesity, Datastax, Dropbox and NetApp are working with NVIDIA microservices to help customers optimize their RAG pipelines and integrate their proprietary data into generative AI applications. Snowflake leverages NeMo Retriever to harness enterprise data for building AI applications.

Enterprises can deploy NVIDIA microservices included with NVIDIA AI Enterprise 5.0 across the infrastructure of their choice, such as leading clouds Amazon Web Services (AWS), Google Cloud, Azure and Oracle Cloud Infrastructure.

NVIDIA microservices are also supported on over 400 NVIDIA-Certified Systems™, including servers and workstations from Cisco, Dell Technologies, Hewlett Packard Enterprise (HPE) , HP, Lenovo and Supermicro. Separately today, HPE announced availability of HPE's enterprise computing solution for generative AI, with planned integration of NIM and NVIDIA AI Foundation models into HPE's AI software.

NVIDIA AI Enterprise microservices are coming to infrastructure software platforms including VMware Private AI Foundation with NVIDIA. Red Hat OpenShift supports NVIDIA NIM microservices to help enterprises more easily integrate generative AI capabilities into their applications with optimized capabilities for security, compliance and controls. Canonical is adding Charmed Kubernetes support for NVIDIA microservices through NVIDIA AI Enterprise.

NVIDIA's ecosystem of hundreds of AI and MLOps partners, including Abridge, Anyscale, Dataiku, DataRobot, Glean, H2O.ai, Securiti AI, Scale AI, OctoAI and Weights & Biases, are adding support for NVIDIA microservices through NVIDIA AI Enterprise.

Apache Lucene, Datastax, Faiss, Kinetica, Milvus, Redis, and Weaviate are among the vector search providers working with NVIDIA NeMo Retriever microservices to power responsive RAG capabilities for enterprises.

Availability
Developers can experiment with NVIDIA microservices at ai.nvidia.com at no charge. Enterprises can deploy production-grade NIM microservices with NVIDIA AI Enterprise 5.0 running on NVIDIA-Certified Systems and leading cloud platforms.

For more information, watch the replay of Huang's GTC keynote and visit the NVIDIA booth at GTC, held at the San Jose Convention Center through March 21.

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 infrastructure 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 NVIDIA's products and technologies, including NVIDIA CUDA platform, NVIDIA NIM microservices, NVIDIA CUDA-X microservices, NVIDIA AI Enterprise 5.0, NVIDIA inference software including Triton Inference Server and TensorRT-LLM, NVIDIA Riva, NVIDIA cuOpt, NVIDIA Earth-2, NeMo Retriever, NVIDIA NeMo Curator, NVIDIA NeMo Customizer, NVIDIA NeMo Evaluator, NVIDIA NeMo Guardrails, NVIDIA AI Foundation models and NVIDIA AI Enterprise microservices; and established enterprise platforms sitting on a goldmine of data that can be transformed into generative AI copilots 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.

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