# NVIDIA AI Foundry Builds Custom Llama 3.1 Generative AI Models for the World's Enterprises

- *Enterprises and Nations Can Now Build 'Supermodels' With NVIDIA AI Foundry Using Their Own Data Paired With Llama 3.1 405B and NVIDIA Nemotron Models*
- *NVIDIA AI Foundry Offers Comprehensive Generative AI Model Service Spanning Curation, Synthetic Data Generation, Fine-Tuning, Retrieval, Guardrails and Evaluation to Deploy Custom Llama 3.1 NVIDIA NIM Microservices With New NVIDIA NeMo Retriever Microservices for Accurate Responses*
- *Accenture First to Use New Service to Build Custom Llama 3.1 Models for Clients; Aramco, AT&T, Uber and Other Industry Leaders Among First to Access New Llama NVIDIA NIM Microservices*

NVIDIA today announced a new NVIDIA AI Foundry service and NVIDIA NIM™ inference microservices to supercharge generative AI for the world's enterprises with the Llama 3.1 collection of openly available models, also introduced today.

With NVIDIA AI Foundry, enterprises and nations can now create custom "supermodels" for their domain-specific industry use cases using Llama 3.1 and NVIDIA software, computing and expertise. Enterprises can train these supermodels with proprietary data as well as synthetic data generated from Llama 3.1 405B and the NVIDIA Nemotron™ Reward model.

NVIDIA AI Foundry is powered by the NVIDIA DGX™ Cloud AI platform, which is co-engineered with the world's leading public clouds, to give enterprises significant compute resources that easily scale as AI demands change.

The new offerings come at a time when enterprises, as well as nations developing sovereign AI strategies, want to build custom large language models with domain-specific knowledge for generative AI applications that reflect their unique business or culture.

"Meta's openly available Llama 3.1 models mark a pivotal moment for the adoption of generative AI within the world's enterprises," said Jensen Huang, founder and CEO of NVIDIA. "Llama 3.1 opens the floodgates for every enterprise and industry to build state-of-the-art generative AI applications. NVIDIA AI Foundry has integrated Llama 3.1 throughout and is ready to help enterprises build and deploy custom Llama supermodels."

"The new Llama 3.1 models are a super-important step for open source AI," said Mark Zuckerberg, founder and CEO of Meta. "With NVIDIA AI Foundry, companies can easily create and customize the state-of-the-art AI services people want and deploy them with NVIDIA NIM. I'm excited to get this in people's hands."

To supercharge enterprise deployments of Llama 3.1 models for production AI, NVIDIA NIM inference microservices for Llama 3.1 models are now available for download from ai.nvidia.com. NIM microservices are the fastest way to deploy Llama 3.1 models in production and power up to 2.5x higher throughput than running inference without NIM.

Enterprises can pair Llama 3.1 NIM microservices with new NVIDIA NeMo Retriever NIM microservices to create state-of-the-art retrieval pipelines for AI copilots, assistants and digital human avatars.

**Accenture Pioneers Custom Llama Supermodels for Enterprises With AI Foundry**
Global professional services firm Accenture is first to adopt NVIDIA AI Foundry to build custom Llama 3.1 models using the Accenture AI Refinery™ framework, both for its own use as well as for clients seeking to deploy generative AI applications that reflect their culture, languages and industries.

"The world's leading enterprises see how generative AI is transforming every industry and are eager to deploy applications powered by custom models," said Julie Sweet, chair and CEO of Accenture. "Accenture has been working with NVIDIA NIM inference microservices for our internal AI applications, and now, using NVIDIA AI Foundry, we can help clients quickly create and deploy custom Llama 3.1 models to power transformative AI applications for their own business priorities."

NVIDIA AI Foundry provides an end-to-end service for quickly building custom supermodels. It combines NVIDIA software, infrastructure and expertise with open community models, technology and support from the NVIDIA AI ecosystem.

With NVIDIA AI Foundry, enterprises can create custom models using Llama 3.1 models and the NVIDIA NeMo platform — including the NVIDIA Nemotron-4 340B Reward model, ranked first on the Hugging Face RewardBench.

Once custom models are created, enterprises can create NVIDIA NIM inference microservices to run them in production using their preferred MLOps and AIOps platforms on their preferred cloud platforms and NVIDIA-Certified Systems™ from global server manufacturers.

NVIDIA AI Enterprise experts and global system integrator partners work with AI Foundry customers to accelerate the entire process, from development to deployment.

**NVIDIA Nemotron Powers Advanced Model Customization**
Enterprises that need additional training data for creating a domain-specific model can use Llama 3.1 405B and Nemotron-4 340B together to generate synthetic data to boost model accuracy when creating custom Llama supermodels.

Customers that have their own training data can customize Llama 3.1 models with NVIDIA NeMo for domain-adaptive pretraining, or DAPT, to further increase model accuracy.

NVIDIA and Meta have also teamed to provide a distillation recipe for Llama 3.1 that developers can use to build smaller custom Llama 3.1 models for generative AI applications. This enables enterprises to run Llama-powered AI applications on a broader range of accelerated infrastructure, such as AI workstations and laptops.

**Industry-Leading Enterprises Supercharge AI With NVIDIA and Llama**
Companies across healthcare, energy, financial services, retail, transportation and telecommunications are already working with NVIDIA NIM microservices for Llama. Among the first to access the new NIM microservices for Llama 3.1 are Aramco, AT&T and Uber.

Trained on over 16,000 NVIDIA H100 Tensor Core GPUs and optimized for NVIDIA accelerated computing and software — in the data center, in the cloud and locally on workstations with NVIDIA RTX™ GPUs or PCs with GeForce RTX GPUs — the Llama 3.1 collection of multilingual LLMs is a collection of generative AI models in 8B-, 70B- and 405B-parameter sizes.

**New NeMo Retriever RAG Microservices Boost Accuracy and Performance**
Using new NVIDIA NeMo Retriever NIM inference microservices for retrieval-augmented generation (RAG), organizations can enhance response accuracy when deploying customized Llama supermodels and Llama NIM microservices in production.

Combined with NVIDIA NIM inference microservices for Llama 3.1 405B, NeMo Retriever NIM microservices deliver the highest open and commercial text Q&A retrieval accuracy for RAG pipelines.

**Enterprise Ecosystem Ready to Power Llama 3.1 and NeMo Retriever NIM Deployments**
Hundreds of NVIDIA NIM partners providing enterprise, data and infrastructure platforms can now integrate the new microservices in their AI solutions to supercharge generative AI for the NVIDIA community of more than 5 million developers and 19,000 startups.

Production support for Llama 3.1 NIM and NeMo Retriever NIM microservices is available through NVIDIA AI Enterprise. Members of the NVIDIA Developer Program will soon be able to access NIM microservices for free for research, development and testing on their preferred infrastructure.

**About NVIDIA**
NVIDIA (NASDAQ: NVDA) is the world leader in accelerated computing.

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 AI Foundry, NVIDIA Nemotron models, NVIDIA Nemotron-4 models, NVIDIA DGX Cloud, NVIDIA NeMo Retriever NIM microservices, NVIDIA NeMo platform, NVIDIA-Certified Systems, NVIDIA Tensor Core GPUs, NVIDIA RTX GPUs and GeForce RTX GPUs; third parties' use or adoption of NVIDIA products, technologies and platforms, and the benefits and impacts thereof; our collaboration with third parties and the benefits and impacts thereof; Llama 3.1 opening the floodgates for every enterprise and industry to build state-of-the-art generative AI applications; and NVIDIA AI Foundry being ready to help enterprises build and deploy custom Llama supermodels 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 hereto 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.

Natalie Hereth
NVIDIA Corporation
nhereth@nvidia.com