NVIDIA NeMo on DGX Infrastructure

Build and Deploy Generative AI Solutions for Your Enterprise.

The Challenge of Implementing Generative AI

While traditional AI systems recognize patterns and make predictions, generative AI enables users to create new and original content, which includes text, images, sounds, animation, 3D models, and more. Building foundation models for generative AI requires mountains of data and large-scale computing infrastructure for training and inference. Additionally, deep technical expertise is required to manage the infrastructure and tap complex algorithms. Even using pretrained foundation models comes with challenges as they don't contain domain- or enterprise-specific knowledge, are captured at a point in time, and may provide undesired or biased information.

Enterprises can now tackle the most complex AI models and successfully deliver generative AI models. **NVIDIA DGX™** infrastructure provides leadership-class infrastructure with an optimized hardware architecture, advanced algorithms, and access to AI experts. By leveraging the **NVIDIA NeMo™ framework** that's part of **NVIDIA AI Enterprise**, an enterprise-ready AI software platform included with DGX H100 systems, enterprises can easily and cost-effectively deliver generative AI across their organization.

Efficiency at Extreme Scale

DGX SuperPOD delivers the supercomputing required when building large language models (LLMs). Businesses can tackle the most complex models, including large-scale GPT, shrinking time to solution from hundreds of years to weeks or even days. Training 175B-parameter GPT-3 takes 355 years on an NVIDIA V100 Tensor Core GPU and 4.8 years on one DGX A100. Using a 140-node DGX SuperPOD, the same task can be completed in about one month. Customers can choose to build their own foundation models using the NeMo framework with DGX SuperPOD and linearly scale to trillion-parameter models. Alternatively, for the fastest way to develop their own foundational models, they can leverage **DGX Cloud**, a high-performance, multi-node AI-training-as-a-service solution.

> "Enterprises that adopt next-generation AI, like large language models (LLMs) and generative AI, are **2.6X more likely to increase revenue by 10% or more**."[1]
>
> — Accenture

> "We trained our LLM models more effectively with NVIDIA DGX SuperPOD's powerful performance — as well as NeMo's optimized algorithms and 3D parallelism techniques... We considered using other platforms, but it was difficult to find an alternative that provides full-stack environments — from the hardware level to the inference level."
>
> **Hwijung Ryu,** LLM development team lead at KT

Learn More

Tools to Build Custom Multimodal Generative AI Models

For enterprises seeking to build their own foundation model, NVIDIA NeMo provides an end-to-end, cloud-native enterprise framework to build, customize, and deploy generative AI models with billions of parameters on DGX systems. The NeMo framework streamlines the development process of the largest generative AI models and provides computational efficiency and scalability to allow for cost-effective training using several state-of-the-art distributed training techniques. The new techniques include sequence parallelism and selective activation recomputation, which deliver up to **30% faster training times of LLMs**. Customization techniques for LLMs such as prompt learning and adapter techniques allow for customization while reinforcement learning with human feedback ensures continuous improvement over time.

Eliminate time wasted searching for efficient model configurations with the autoconfiguration tool, which can automatically find optimal training and inference configurations. Accelerate execution of models built on the NeMo framework with state-of-the-art optimization techniques, which can perform inference of large-scale models on multiple GPUs and multiple DGX nodes. Developers can achieve low-latency and high-throughput inference, leading to lower resource consumption.

Train Every LLM

Validated Convergence | Scalable
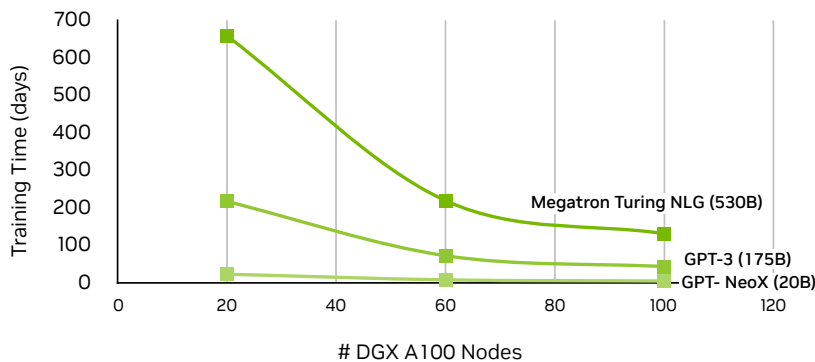
Train every large language model with NVIDIA NeMo on DGX

Optimized Topology for Multi-Node Training

Use the NeMo framework to train the largest models using model parallelism, with **NVIDIA NVLink®** technology and InfiniBand networking for fast cross-node communication. A 32-node DGX SuperPOD with **DGX H100** provides 1 exaFLOPS of AI computing, a multi-rail high-performance InfiniBand network optimized with **NVIDIA Magnum IO™**, **NVIDIA Collective Communications Library (NCCL)**, and NVIDIA Scalable Hierarchical Aggregation Reduction Protocol (SHARP)™ in-network acceleration.

Enterprises can use the NeMo framework to develop their own customized models on DGX Cloud. DGX Cloud is optimized for multi-node training, and is built on the same core NVIDIA Tensor Core GPU and NVLink technologies. These networking technologies power record-breaking **MLPerf** benchmarks and enable dozens of supercomputers on the **TOP500** and **Green500** lists. This optimized topology,


Key Features

DGX Platform

> DGX Cloud

> DGX BasePOD

> DGX SuperPOD

NVIDIA NeMo Framework

> Part of NVIDIA AI Enterprise, included with the DGX platform

> Text-to-text, text-to-image, and image-to-image foundation models

> Techniques for data curation and distributed training

> Customization tools like prompt learning and reinforcement learning with human feedback (RLHF)

> Accelerated multi-node and multi-GPU inferencing

> Scripts and reference examples

> NVIDIA professional services and DGXperts

coupled with the end-to-end enterprise framework of NeMo, enables businesses to rapidly create custom-tailored generative AI solutions for their most important missions. It also understands their unique data.
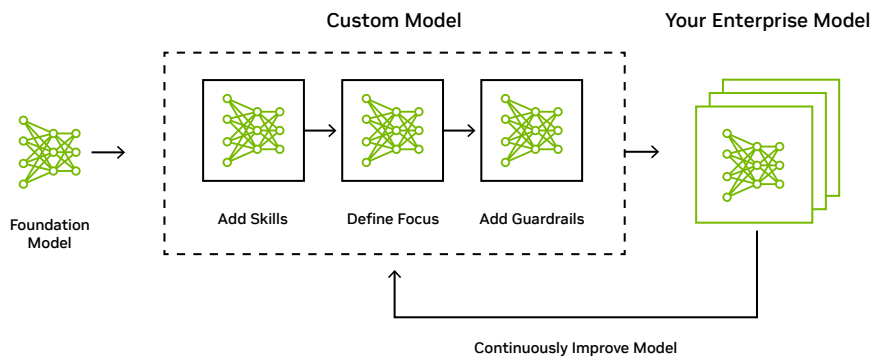
NVIDIA NeMo framework: The easy, cost-effective, and fastest way to develop generative AI models.

Direct Access to World-Class LLM Experts

NVIDIA DGX infrastructure comes with access to dedicated expertise for help with everything from installation and infrastructure management to scaling workloads and streamlining production AI. Partner with a global team of AI-fluent practitioners who have built a wealth of experience over the last decade and have successfully completed many AI infrastructure deployments, including for DGX customers on the TOP500 list of the world's fastest supercomputers.

Examples of Successful Enterprise LLM Deployments

Discover how NVIDIA DGX SuperPOD paired with NVIDIA NeMo delivers LLM applications for multiple languages and industries. **Read the Ebook.**

Ready to Get Started?

To learn more about the NVIDIA NeMo framework, visit
nvidia.com/nemo-framework

To speed up generative AI development on NVIDIA DGX, visit
nvidia.com/dgx