Solution Overview

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 optimized to run on NVIDIA DGX 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 a 175B-parameter GPT-3 model on 8 DGX systems takes 2.5 months, as a greater number of systems are necessary to accommodate the model's vast memory requirements. However, leveraging a DGX SuperPOD with 95 DGX H100 systems can reduce this training time to less than a week, showcasing a dramatic improvement in efficiency and scalability.[2] Customers can choose to build their own foundation models using the NeMo framework with DGX SuperPOD and linearly scale to trillion-parameter models.

> "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

1 Accenture Research. **Breakthrough Innovation: Is your organization equipped for breakthrough innovation?** January 2023.

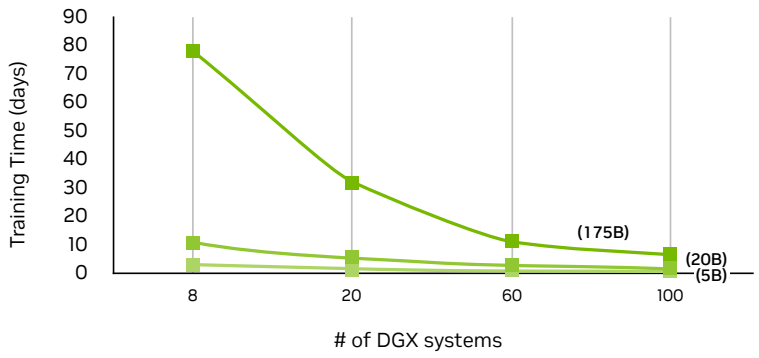
2 Based on projected time to train 300B tokens on 175B parameter GPT-3 model on DGX H100 systems, NeMo 24.05, FP8

Tools to Build Custom Multimodal Generative AI Models

For enterprises seeking to build their own foundation model, NVIDIA NeMo provides an end-to-end, 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 supervised fine-tuning (SFT) and Low-Rank Adaptation (LoRA) allow for customization, while reinforcement learning with human feedback ensures continuous improvement over time. The NeMo framework also includes support for **Mixture of Experts (MoE) based LLM architectures**, which enable model capacity to be increased without a proportional increase in both the training and inference compute requirements. This further facilitates efficient training of trillion parameter models.

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 DGX systems. For optimal deployment of NeMo models, use **NVIDIA NIM™** to speed deployment with ease of use, manageability, and security. Developers can achieve low-latency and high-throughput inference, leading to lower resource consumption.

Time to Train LLMs



Time to train ChatGPT-3 models, each with a total of 300B tokens on NVIDIA DGX H100 systems, NVIDIA NeMo 24.05, FP8

Optimized Topology for Multi-Node Training

Use the NeMo framework to train the largest models using model parallelism, with **NVIDIA® NVLink™** technology and **NVIDIA InfiniBand** networking for fast inter-node communication. A 32-node DGX SuperPOD with Hopper architecture 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. These networking technologies power record-breaking **MLPerf** benchmarks and enable dozens of supercomputers on the **TOP500** and **Green500** lists. This optimized topology, coupled with the

Key Features

DGX Platform

- DGX Systems
- DGX BasePOD
- DGX SuperPOD

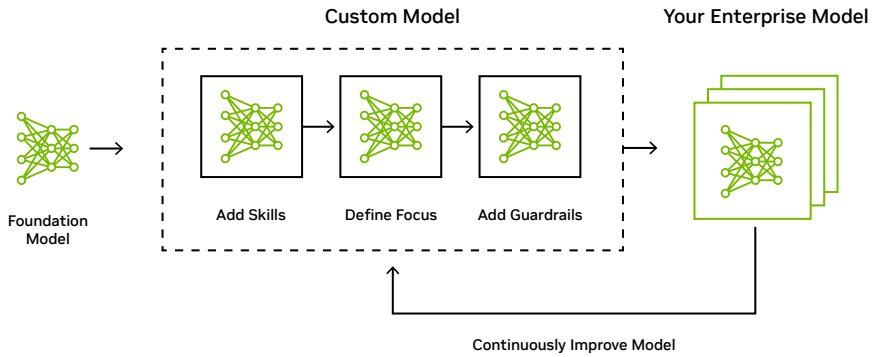
NVIDIA NeMo Framework

- Part of NVIDIA AI Enterprise, optimized to run on DGX platform
- Automatic speech recognition, text to speech, text-to-text, text-to-image, and image-to-image foundation models
- Techniques for data curation and distributed training
- Support for Mixture of Experts (MOE) based LLM architectures with expert parallelism
- State-of-the-art (SOTA) instruction tuning techniques like reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO), and customization techniques like parameter-efficient fine-tuning (PEFT) (LoRA, QLoRA)
- Accelerated multi-node and multi-GPU inferencing
- Scripts and reference examples

NVIDIA NIM

- Part of NVIDIA AI Enterprise, optimized to run on DGX platform
- Accelerated inference microservices for optimal deployment of models fine-tuned with NeMo

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**.

Get Going With NVIDIA NeMo in Three Easy Steps

1. Train specialized datasets using the NVIDIA NeMo framework on DGX systems.
2. Customize models using tasks like RLHF, p-tuning, and fine-tuning. Provide current and proprietary information to the model to get state-of-the-art accuracy.
3. Run at scale using DGX infrastructure on premises, in **colocation data centers** on in **private clouds**.

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**

Partner Logo

NVIDIA