



NVIDIA NeMo Megatron on NVIDIA DGX Systems

Simplify development and deployment of enterprise-grade large language models.



Large Language Models are on the Rise

- > The global natural language processing (NLP) market is expected to **grow to \$49.4 billion by 2027**.¹ - Markets and Markets
- > 80% of organizations are waking up to the fact that 80% of **their content is unstructured**.² - Accenture
- > 58% of global respondents have adopted NLP, **94% are using or planning to use it in the next year**.³ - Deloitte Insights

The Challenge for Creating World-Class NLP

Businesses around the world are tapping into the power of NLP to deliver enhanced customer experiences that feel more natural and intuitive. These demands are driving the creation of large language models (LLMs) with exponentially increasing model complexity and size. LLMs require computing on a scale beyond that required by commonplace NLP apps. Instead, state-of-the-art LLM applications must be customized for each enterprise and their unique vocabulary, their customer relationships, and the datasets on which their business runs, and delivered at the speed of business.

Enterprises can now tackle the most complex AI models and deliver superhuman language understanding. DGX infrastructure provides leadership-class infrastructure with an optimized hardware architecture, advanced algorithms, and access to AI experts so you can stand up your own world-class NLP architecture in no time.

Efficiency at Extreme Scale

DGX SuperPOD delivers the supercomputing power and scale to deliver LLMs for enterprises. Businesses can tackle the most complex models, including GPT-3, 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 GPU and 14.8 years on one DGX A100. Using a 140-node DGX SuperPOD, this can be done in about one month. The Full-Stack Solution for NeMo Megatron on DGX SuperPOD enables you to linearly scale to trillion-parameter models.

Key Features

- > NVIDIA DGX SuperPOD™ or DGX BasePOD™ for AI training and inference
- > Full-Stack Solution for NVIDIA NeMo Megatron
 - NVIDIA Nemo Megatron for data curation and distributed training
 - NVIDIA Triton™ Inference Server with FasterTransformer backend
 - Scripts and reference models
- > NVIDIA Professional Services and DGXperts

Large-Scale AI with Chinese Language Models

JD Explore Academy, the research and development division of JD.com, a leading supply chain-based technology and service provider and China's largest online retailer, is using NVIDIA DGX SuperPOD to develop NLP for the application of smart customer service, smart retail, smart logistics, IoT, healthcare, and more. The company is training GPT-3-like large-scale models to investigate how to efficiently and trustworthily transfer knowledge from large-scale data to the parameters of the pretraining model, and thus comprehensively improve various downstream NLP tasks, such as sentiment analysis, dialogue, and translation. [Learn more >](#)

1. Markets and Markets. [Natural Language Processing \(NLP\) Market by Component \(Solutions & Services\), Application \(Sentiment Analysis, Social Media Monitoring\), Technology \(IVR, OCR, Auto Coding\), Vertical \(BFSI, Retail & eCommerce, IT & ITES\) & Region - Global Forecast to 2027](#).

2. Paul Nelson. [Search and unstructured data analytics: 5 trends to watch in 2020](#). Accenture. January 2020.

3. Beena Ammanath, David Jarvis, and Susanne Hupfer. [Thriving in the era of pervasive AI: Deloitte's State of AI in the Enterprise, 3rd Edition](#). July 2020.

Tools to Build Your Own Custom Language Models

The Full-Stack Solution for NeMo Megatron provides a comprehensive framework, alongside scripts and recipes on a single codebase, delivered as a containerized framework. NVIDIA **NeMo Megatron** streamlines the development process of the largest language 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**. This builds on the **NVIDIA Megatron research project** that was used to train the world's largest transformer-based language models.

Eliminate time wasted searching for efficient model configurations with NeMo Megatron's hyperparameter tool, which can automatically find optimal training and inference configurations. Accelerate execution of models with optimization techniques from **NVIDIA Triton Inference Server** with **FasterTransformer** backend, which can perform inference with huge GPT-3 models on multiple GPUs and multiple DGX nodes. With NVIDIA Triton Inference Server, you can achieve low-latency and high-throughput inference.

Optimized Topology for Multi-Node Training

Train the largest models using model parallelism, with NVLink and InfiniBand for fast cross-node communication. A 140- node DGX SuperPOD provides 700 PFLOPS of AI computing, a multi-rail high-performance InfiniBand network optimized with Magnum IO™ NCCL and SHARPV2 In-Network Acceleration. This optimized topology coupled with the Full-Stack Solution for NeMo Megatron enables businesses to rapidly create custom-tailored NLP for their most important missions that understands their unique data and knows their customers intimately.

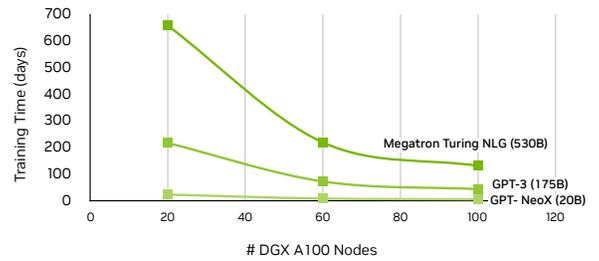
Streamline LLM Development Workflows

NVIDIA Base Command Platform accelerates the process of prototyping and scaling LLMs by providing a cloud-hosted control plane and a single pane-of-glass view for managing NLP training projects. Developers can use the platform to simplify and accelerate the time taken to coordinate the infrastructure supporting LLM development, gaining greater productivity and faster return on investment on AI projects. Businesses can experience Base Command Platform by trying it within **NVIDIA LaunchPad**, or paired with their on-premises DGX infrastructure.

Train Every Large Language Model with NVIDIA DGX Infrastructure and NeMo Megatron

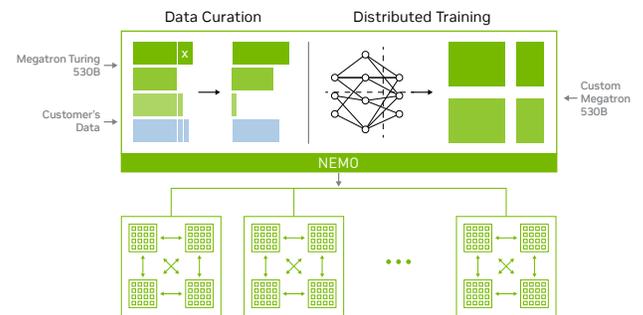
Train Every LLM

Validated Convergence | Scalable



Fastest LLM Training

In-Cloud, On-Prem

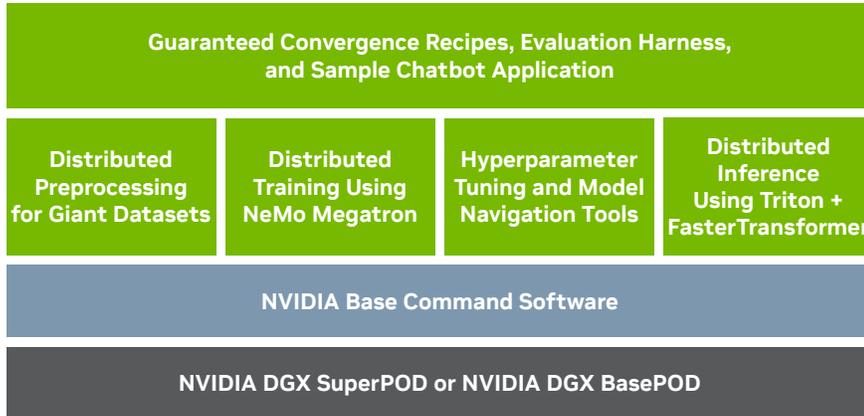


NVIDIA DGX SuperPOD delivers LLM applications for multiple languages and industries



Turnkey Experience for Rapid Deployment

Get a full-stack data center platform that includes industry leading computing, storage, networking, software, I/O acceleration, and management tools in no time. Enterprises conducting cutting-edge NLP work can get up and running, from initial engagement to power on of their DGX infrastructure, in as little as three months.



The Full-Stack Solution for NVIDIA NeMo Megatron on NVIDIA DGX Systems

Direct Access to World-Class NLP Experts

NVIDIA DGX infrastructure comes with access to dedicated expertise from install to infrastructure management to scaling workloads to streamlined production AI. Partner with a global team of AI-fluent practitioners who have built a wealth of experience over the last decade and have done many successful AI infrastructure deployments, including DGX customers on the **TOP500** list of the world's fastest supercomputers.

Ready to Get Started?

Test drive NeMo Megatron on DGX systems with a free, short-term trial on NVIDIA Base Command:

www.nvidia.com/try-base-command.

Learn more about NVIDIA NeMo Megatron:

developer.nvidia.com/nemo/megatron.

Learn how to accelerate LLMs on NVIDIA DGX systems, at:

www.nvidia.com/dgx.

Get Going in Three Easy Steps



Step 01

Pretrain large language models on massive amounts of text using NVIDIA NeMo Megatron on NVIDIA DGX SuperPOD or DGX BasePOD



Step 02

Fine-tune on variety of downstream tasks to get state-of-the-art accuracy



Step 03

Run the fastest inference using unique tools like NVIDIA Triton Inference Server with FasterTransformer backend



Pro Tip

To streamline and accelerate LLM development, organizations can use NVIDIA Base Command Platform which provides end to end management of workflow, giving developers a simplified experience that scales productivity and enables more models to be deployed in production. Base Command Platform is available with on-premises DGX infrastructure.