NVIDIA Brings Large Language AI Models to Enterprises Worldwide

NVIDIA NeMo Megatron Framework; Megatron 530B Customizable Large Language Model; Multi-GPU, Multinode Triton Inference Server Empower Development, Deployment of Language-Based AI to Advance Industries and Science

GTC—NVIDIA today opened the door for enterprises worldwide to develop and deploy large language models (LLM) by enabling them to build their own domain-specific chatbots, personal assistants and other AI applications that understand language with unprecedented levels of subtlety and nuance.

The company unveiled the NVIDIA NeMo Megatron framework for training language models with trillions of parameters, the Megatron 530B customizable LLM that can be trained for new domains and languages, and NVIDIA Triton Inference Server™ with multi-GPU, multinode distributed inference functionality.

Combined with NVIDIA DGX™ systems, these tools provide a production-ready, enterprise-grade solution to simplify the development and deployment of large language models.

“Large language models have proven to be flexible and capable, able to answer deep domain questions, translate languages, comprehend and summarize documents, write stories and compute programs, all without specialized training or supervision,” said Bryan Catanzaro, vice president of Applied Deep Learning Research at NVIDIA. “Building large language models for new languages and domains is likely the largest supercomputing application yet, and now these capabilities are within reach for the world’s enterprises.”

NVIDIA NeMo Megatron and Megatron 530B Speed LLM Development
NVIDIA NeMo Megatron builds on advancements from Megatron, an open-source project led by NVIDIA researchers studying efficient training of large transformer language models at scale. Megatron 530B is the world’s largest customizable language model.

The NeMo Megatron framework enables enterprises to overcome the challenges of training sophisticated natural language processing models. It is optimized to scale out across the large-scale accelerated computing infrastructure of NVIDIA DGX SuperPOD™.

NeMo Megatron automates the complexity of LLM training with data processing libraries that ingest, curate, organize and clean data. Using advanced technologies for data, tensor and pipeline parallelization, it enables the training of large language models to be distributed efficiently across thousands of GPUs. Enterprises can use the NeMo Megatron framework to train LLMs for their specific domains and languages.

NVIDIA Triton Inference Server Powers Real-Time LLM Inference
New multi-GPU, multinode features in the latest NVIDIA Triton Inference Server — announced separately today — enable LLM inference workloads to scale across multiple GPUs and nodes with real-time performance. The models require more memory than is available in a single GPU or even a large server with multiple GPUs, and inference must run quickly to be useful in applications.

With Triton Inference Server, Megatron 530B can run on two NVIDIA DGX systems to shorten the processing time from over a minute on a CPU server to half a second, making it possible to deploy LLMs for real-time applications.

Massive Custom Language Models Developed Worldwide
Among early adopters building large language models with NVIDIA DGX SuperPOD are SiDi, JD Explore Academy and VinBrain.

SiDi, one of Brazil’s largest AI research and development institutes, has adapted the Samsung virtual assistant for use by the nation’s 200 million Brazilian Portuguese speakers.

“The SiDi team has extensive experience developing AI virtual assistants and chatbots, which require both powerful AI performance and specialized software that is trained and adapted to the changing nuances of human language,” said John Yi, CEO of SiDi. “NVIDIA DGX SuperPOD is ideal for powering the advanced work of our team to help us bring world-leading AI services to Portuguese speakers in Brazil.”

JD Explore Academy**,** the research and development division of JD.com, a leading supply chain-based technology and service provider, is utilizing NVIDIA DGX SuperPOD to develop NLP for the application of smart customer service, smart retail, smart logistics, IoT, healthcare and more.

VinBrain, a Vietnam-based healthcare AI company, has used a DGX SuperPOD to develop and deploy a clinical language model for radiologists and telehealth in 100 hospitals, where it is used by over 600 healthcare practitioners.

Availability
Enterprises can experience developing and deploying large language models at no charge in curated labs with NVIDIA LaunchPad, announced separately today.

Organizations can apply to join the early access program for the NVIDIA NeMo Megatron accelerated framework for training large language models.

NVIDIA Triton is available from the NVIDIA NGC™ catalog, a hub for GPU-optimized AI software that includes frameworks, toolkits, pretrained models and Jupyter Notebooks, and as open source code from the Triton GitHub repository.

Triton is also included in the NVIDIA AI Enterprise software suite, which is optimized, certified and supported by NVIDIA. Enterprises can use the software suite to run language model inference on mainstream accelerated servers in on-prem data centers and private clouds.

NVIDIA DGX SuperPOD and NVIDIA DGX systems are available from NVIDIA's global resellers, which can provide pricing to qualified customers upon request.

Register for free to learn more during NVIDIA GTC, taking place online through Nov. 11. Watch NVIDIA founder and CEO Jensen Huang's GTC keynote address streaming on Nov. 9 and in replay.

About NVIDIA
NVIDIA's (NASDAQ: NVDA) invention of the GPU in 1999 sparked the growth of the PC gaming market and has redefined modern computer graphics, high performance computing and artificial intelligence. The company's pioneering work in accelerated computing and AI is reshaping trillion-dollar industries, such as transportation, healthcare and manufacturing, and fueling the growth of many others. 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 our products and services, including the NVIDIA NeMo Megatron framework, NVIDIA Triton Inference Server, NVIDIA DGX SuperPOD and NVIDIA DGX systems; and building large language models for new languages and domains as likely to be the largest supercomputing application yet, and these capabilities being within reach for the world's enterprises 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.

Shannon McPhee
+1-310-920-9642
smcphee@nvidia.com