NVIDIA Launches Large Language Model Cloud Services to Advance AI and Digital Biology

NVIDIA NeMo LLM Service Helps Developers Customize Massive Language Models; NVIDIA BioNeMo Service Helps Researchers Generate and Predict Molecules, Proteins, DNA

GTC—NVIDIA today announced two new large language model cloud AI services — the NVIDIA NeMo Large Language Model Service and the NVIDIA BioNeMo LLM Service — that enable developers to easily adapt LLMs and deploy customized AI applications for content generation, text summarization, chatbots, code development, as well as protein structure and biomolecular property predictions, and more.

The NeMo LLM Service allows developers to rapidly tailor a number of pretrained foundation models using a training method called prompt learning on NVIDIA-managed infrastructure. The NVIDIA BioNeMo Service is a cloud application programming interface (API) that expands LLM use cases beyond language and into scientific applications to accelerate drug discovery for pharma and biotech companies.

“Large language models hold the potential to transform every industry,” said Jensen Huang, founder and CEO of NVIDIA. “The ability to tune foundation models puts the power of LLMs within reach of millions of developers who can now create language services and power scientific discoveries without needing to build a massive model from scratch.”

NeMo LLM Service Boosts Accuracy With Prompt Learning, Accelerates Deployments
With the NeMo LLM Service, developers can use their own training data to customize foundation models ranging from 3 billion parameters up to Megatron 530B, one of the world’s largest LLMs. The process takes just minutes to hours compared with the weeks or months required to train a model from scratch.

Models are customized with prompt learning, which uses a technique called p-tuning. This allows developers to use just a few hundred examples to rapidly tailor foundation models that were originally trained with billions of data points. The customization process generates task-specific prompt tokens, which are then combined with the foundation models to deliver higher accuracy and more relevant responses for specific use cases.

Developers can customize for multiple use cases using the same model and generate many different prompt tokens. A playground feature provides a no-code option to easily experiment and interact with models, further boosting the effectiveness and accessibility of LLMs for industry-specific use cases.

Once ready to deploy, the tuned models can run on cloud instances, on-premises systems or through an API.

BioNeMo LLM Service Enables Researchers to Tap Power of Massive Models
The BioNeMo LLM Service includes two new BioNeMo language models for chemistry and biology applications. It provides support for protein, DNA and biochemical data to help researchers discover patterns and insights in biological sequences.

BioNeMo enables researchers to expand the scope of their work by leveraging models that contain billions of parameters. These larger models can store more information about the structure of proteins, evolutionary relationships between genes, and even generate novel biomolecules for therapeutic applications.

Cloud API Provides Access to Megatron 530B, Other Ready-Made Models
In addition to tuning foundation models, the LLM services include the option to use ready-made and custom models through a cloud API.

This gives developers access to a broad range of pretrained LLMs, including Megatron 530B. It also provides access to T5 and GPT-3 models created with the NVIDIA NeMo Megatron framework — now available in open beta — to support a broad range of applications and multilingual service requirements.

Leaders in automotive, computing, education, healthcare, telecommunications and other industries are using NeMo Megatron to pioneer services for customers in Chinese, English, Korean, Swedish and other languages.

Availability
The NeMo LLM and BioNeMo services and cloud APIs are expected to be available in early access starting next month. Developers can apply now for more details.

The beta release of the NeMo Megatron framework is available from NVIDIA NGC™ and is optimized to run on NVIDIA DGX™ Foundry and NVIDIA DGX SuperPOD™, as well as accelerated cloud instances from Amazon Web Services, Microsoft Azure and Oracle Cloud Infrastructure.

To experience the NeMo Megatron framework, developers can try NVIDIA LaunchPad labs at no charge.

Tune in to Huang's GTC keynote to learn more about large language models powered by NVIDIA AI.

About NVIDIA
Since its founding in 1993, NVIDIA (NASDAQ: NVDA) has been a pioneer in accelerated computing. The company's invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics and ignited the era of modern AI. NVIDIA is now a full-stack computing company with data-center-scale offerings that are reshaping industry. More information at https://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, capabilities, and availability of our products and technologies, including the NeMo LLM Service and the BioNeMo LLM Service; the potential of large language models to transform every industry; the impact of the ability to tune foundation models; and larger models storing more information about the structure of proteins, evolutionary relationships between genes, and generating novel biomolecules for therapeutic applications 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