Simplifying Access to Large Language Models with NVIDIA Nemo Framework and Service

What's new: *NVIDIA has announced two services:* **Nemo LLM**, *and* **BioNeMo**. ***NVIDIA NeMo Megatron, an end-to-end framework for training and deploying LLMs, is now available to developer around the world in open beta.***

From recent advances in large language models (LLM) have fueled state-of-art performance for NLP applications such as virtual scribes in health care, interactive virtual assistances, and many more.

So to *simply access to LLMs, NVIDIA has announced NeMo LLM and BioNeMo.*

NeMo LLM for customizing and using LLMs and

BioNeMo that which expands scientific applications of LLMs for pharmaceuticals and biotechnology industries.

NeMo LLM Service:

NVIDIA NeMo LLM service provides the fastest path to customize foundation LLMs and deploy them at scale leveraging the NVIDIA managed cloud API or through private and public clouds.

NVIDIA and community built foundation models can be customized using prompt learning capabilities. Now, the promise of LLM serving several use cases with a single model is realized.

Prompt learning capabilities, which are compute efficient techniques, embedding context in user queries to enable greater accuracy in specific use cases. These techniques require just a few hundred samples to achieve high accuracy.

Developers can build applications ranging from text summarization, to paraphrasing, to story generation, and many others, for specific domains and use cases. Minimal compute and technical expertise are required.

Megatron 530B model is one of the world's largest LLMs, with 530 billion parameters based on GPT-3 architecture.

BioNeMo service

BioNeMo service, built on NeMo Megatorn

BioNeMo ***is an AI powered drug discovery cloud service and frame work built on NVIDIA NeMo Megatorn for training and deploying large bimolecular transformer AI models at supercomputing scale***. The service includes pre trained large language models (LLMs) and native **support for common file formats** for proteins, DNA, RNA, and chemistry, providing data loaders for SMILES for molecular structures and FASTA for amino acid and nucleotide sequences.

Cloud environment for AI based drug discovery workflows. Chemists, biologists, and Ai drug discovery researches can generate novel therapeutics; understand the properties and structure, and function; and ultimately predict binding to a drug target.

Today, the BioNeMo service supports state-of-the-art transformer-based models for both chemistry and proteomics. Support for DNA-based workflows is coming soon. The ESM-1 architecture provides equivalent capabilities for proteins, and OpenFold is supported for ease of use and scaling of workflows for predictions of protein structures.

The platform enables an end to end modular drug discovery workflow to accelerate research and better understand proteins, genes, and other molecules.

NeMo Megatron

NVIDIA has announced new updates to NVIDIA NeMo Megatron, and end-to-end frame work for training and deploying LLM up to **trillion of paramaters.** NeMo Megatron is now available to developers in **open beta**, on **several cloud platforms** including **Microsoft Azure**, **Amazon web Service**, and **Oracle Cloud Infrastructure**, as well as **NVIDIA DGX SuperPODs** and **NVIDIA DGX Foundry**.

NeMo Megatron is available as a containerized framework on NGC, offering an easy, effective and cost effective path to build and deploy LLMs. It consist of an end-to-end workflow for automated distributed data processing; training large-scale customized GPT-3. T5, and multilingual T5 (mT5) models; and deploying models for inference at scale.

Its hyper parameter tool enables **custom model development,** automatically searching for the best hyper parameter configurations for both training and inference, on any given distributed GPU cluster configuration.

Large scale models are made **practical**, **delivering high training efficiency**, using techniques such as tensor, data, pipeline parallelism, and sequence parallelism, alongside selective activation recomputing. It is also equipped with **prompt learning techniques** that performance and **few shot tasks**.

Reference: Annamalai Chockalingam author NVIDIA [Author: Annamalai Chockalingam | NVIDIA Technical Blog]