# *Chipmakers Boosts AI a service: Nvidia Launches Cloud Services for NLP Models*

**What's new** : NVIDIA announced early access to **NeMo LLM** and **BioNeMo**, cloud computing services that enables developers to generate text and biological sequences respectively, including the methods trained on web data to work well with a particular users data and task without fine-tuning .Users can deploy a variety of models in the cloud, on-premises, or via on **API**.

## How it works:

The new services are based on NVIDIA's pre-existing NeMo toolkit for Speech Recognition, text-to-speech, and Natural Language Processing.

- NeMo LLM provides access to large Language models including Megatron 530B, T5, and GP-T3. Users can apply two methods of so called **prompt learning** to improve the performance.
- The **prompt learning method called p-tuning** enlists an LSTM to map input tokens representations that elicit better performance from a given model. The LSTM learns this mapping via supervised training on a small number of user-supplied examples
- A **second prompt learning approach**, **prompt tuning**, appends a learned representation of a task to the end of the tokens before feeding them to the model. The representation is learned via supervised training on a small number of user-supplied examples.
- BioNeMo enables users to harness large language models for drug discovery. BioNeMo includes pretrained models such as the molecular-structure model MegaMoIBART, the protein-structure model ESM-1, and the protein-folding model OpenFold.

## BEHIND THE NEWS:

- HuggingFace's accelerated Inference API allows users to implement over 20,000 transformer-based models
- NLP cloud allows users to fine-tune and deploy open-source language models including EleutherAI's GPT-J and GPT-NeoX 20B
- In December 2021, OpenAI enabled customers to fine-tune its large language models, GPT-3.

REFERENCES: Nvidia Launches Cloud Service for NLP Models (deeplearning.ai)