# What Are Large Language Models and Why Are They Important?

blogs.nvidia.com/blog/what-are-large-language-models-used-for

## What Are Large Language Models Used For?

Large language models recognize, summarize, translate, predict and generate text and other forms of content.

January 26, 2023 by Angie Lee
AI applications are summarizing articles, writing stories and engaging in long conversations — and large language models are doing the heavy lifting.

A large language model, or LLM, is a deep learning algorithm that can recognize, summarize, translate, predict and generate text and other forms of content based on knowledge gained from massive datasets.

Large language models are among the most successful applications of transformer models. They aren't just for teaching AIs human languages, but for understanding proteins, writing software code, and much, much more.

In addition to accelerating natural language processing applications — like translation, chatbots and AI assistants — large language models are used in healthcare, software development and use cases in many other fields.

## What Are Large Language Models Used For?

Language is used for more than human communication.

Code is the language of computers. Protein and molecular sequences are the language of biology. Large language models can be applied to such languages or scenarios in which communication of different types is needed.

These models broaden AI's reach across industries and enterprises, and are expected to enable a new wave of research, creativity and productivity, as they can help to generate complex solutions for the world's toughest problems.

For example, an AI system using large language models can learn from a database of molecular and protein structures, then use that knowledge to provide viable chemical compounds that help scientists develop groundbreaking vaccines or treatments.

Large language models are also helping to create reimagined search engines, tutoring chatbots, composition tools for songs, poems, stories and marketing materials, and more.

## How Do Large Language Models Work?

Large language models learn from huge volumes of data. As its name suggests, central to an LLM is the size of the dataset it's trained on. But the definition of "large" is growing, along with AI.

Now, large language models are typically trained on datasets large enough to include nearly everything that has been written on the internet over a large span of time.

Such massive amounts of text are fed into the AI algorithm using unsupervised learning — when a model is given a dataset without explicit instructions on what to do with it. Through this method, a large language model learns words, as well as the relationships between and concepts behind them. It could, for example, learn to differentiate the two meanings of the word "bark" based on its context.

And just as a person who masters a language can guess what might come next in a sentence or paragraph — or even come up with new words or concepts themselves — a large language model can apply its knowledge to predict and generate content.

Large language models can also be customized for specific use cases, including through techniques like fine-tuning or prompt-tuning, which is the process of feeding the model small bits of data to focus on, to train it for a specific application.

Thanks to its computational efficiency in processing sequences in parallel, the transformer model architecture is the building block behind the largest and most powerful LLMs.

## Top Applications for Large Language Models

Large language models are unlocking new possibilities in areas such as search engines, natural language processing, healthcare, robotics and code generation.

The popular ChatGPT AI chatbot is one application of a large language model. It can be used for a myriad of natural language processing tasks.

The nearly infinite applications for LLMs also include:

- Retailers and other service providers can use large language models to provide improved customer experiences through dynamic chatbots, AI assistants and more.
- Search engines can use large language models to provide more direct, human-like answers.
- Life science researchers can train large language models to understand proteins, molecules, DNA and RNA.
- Developers can write software and teach robots physical tasks with large language models.

- Marketers can train a large language model to organize customer feedback and requests into clusters, or segment products into categories based on product descriptions.
- Financial advisors can summarize earnings calls and create transcripts of important meetings using large language models. And credit-card companies can use LLMs for anomaly detection and fraud analysis to protect consumers.
- Legal teams can use large language models to help with legal paraphrasing and scribing.

Running these massive models in production efficiently is resource-intensive and requires expertise, among other challenges, so enterprises turn to NVIDIA Triton Inference Server, software that helps standardize model deployment and deliver fast and scalable AI in production.

## When to Use Custom Large Language Models

Many organizations are looking to use custom LLMs tailored to their use case and brand voice. These custom models built on domain-specific data unlock opportunities for enterprises to improve internal operations and offer new customer experiences. Custom models are smaller, more efficient and faster than general-purpose LLMs.

Custom models offer the best solution for applications that involve a lot of proprietary data. One example of a custom LLM is BloombergGPT, homegrown by Bloomberg. It has 50 billion parameters and is targeted at financial applications.

## Where to Find Large Language Models

In June 2020, OpenAI released GPT-3 as a service, powered by a 175-billion-parameter model that can generate text and code with short written prompts.

In 2021, NVIDIA and Microsoft developed Megatron-Turing Natural Language Generation 530B, one of the world’s largest models for reading comprehension and natural language inference, which eases tasks like summarization and content generation.

And HuggingFace last year introduced BLOOM, an open large language model that’s able to generate text in 46 natural languages and over a dozen programming languages.

Another LLM, Codex, turns text to code for software engineers and other developers.

NVIDIA offers tools to ease the building and deployment of large language models:

- NVIDIA NeMo LLM Service provides a fast path to customizing large language models and deploying them at scale using NVIDIA’s managed cloud API, or through private and public clouds.

- NVIDIA NeMo framework, part of the NVIDIA AI platform, enables easy, efficient, cost-effective training and deployment of large language models. Designed for enterprise application development, NeMo provides an end-to-end workflow for automated distributed data processing; training large-scale, customized model types including GPT-3 and T5; and deploying these models for inference at scale.
- NVIDIA BioNeMo is a domain-specific managed service and framework for large language models in proteomics, small molecules, DNA and RNA. It's built on NVIDIA NeMo for training and deploying large biomolecular transformer AI models at supercomputing scale.

## Challenges of Large Language Models

Scaling and maintaining large language models can be difficult and expensive.

Building a foundational large language model often requires months of training time and millions of dollars.

And because LLMs require a significant amount of training data, developers and enterprises can find it a challenge to access large-enough datasets.

Due to the scale of large language models, deploying them requires technical expertise, including a strong understanding of deep learning, transformer models and distributed software and hardware.

Many leaders in tech are working to advance development and build resources that can expand access to large language models, allowing consumers and enterprises of all sizes to reap their benefits.

## Beyond Words: Large Language Models Expand AI's Horizon

The powerful models making waves in natural language processing are rippling across fields from healthcare to robotics and beyond.

October 10, 2022 by Rick Merritt
Back in 2018, BERT got people talking about how machine learning models were learning to read and speak. Today, large language models, or LLMs, are growing up fast, showing dexterity in all sorts of applications.

They're, for one, speeding drug discovery, thanks to research from the Rostlab at Technical University of Munich, as well as work by a team from Harvard, Yale and New York University and others. In separate efforts, they applied LLMs to interpret the strings of amino acids that make up proteins, advancing our understanding of these building blocks of biology.

It's one of many inroads LLMs are making in healthcare, robotics and other fields.

## A Brief History of LLMs

Transformer models — neural networks, defined in 2017, that can learn context in sequential data — got LLMs started.

Researchers behind BERT and other transformer models made 2018 “a watershed moment” for natural language processing, a report on AI said at the end of that year. “Quite a few experts have claimed that the release of BERT marks a new era in NLP,” it added.

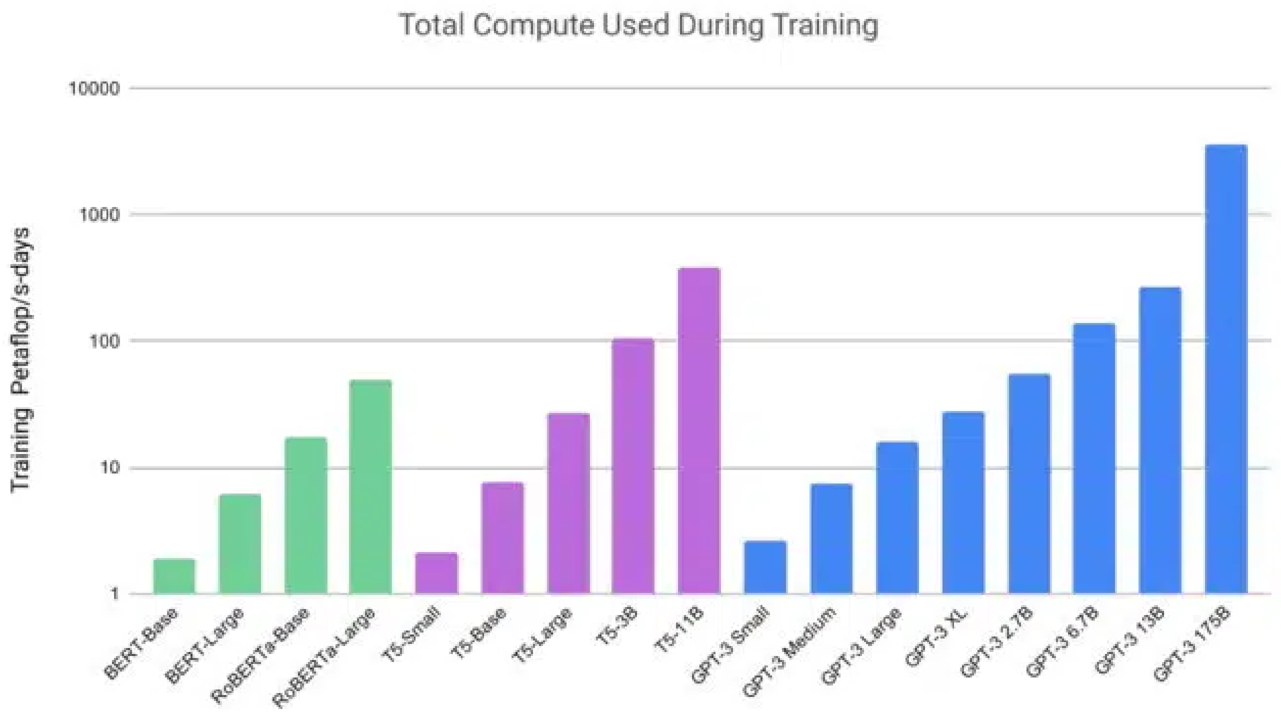
Developed by Google, BERT (aka Bidirectional Encoder Representations from Transformers) delivered state-of-the-art scores on benchmarks for NLP. In 2019, it announced BERT powers the company’s search engine.

Google released BERT as open-source software, spawning a family of follow-ons and setting off a race to build ever larger, more powerful LLMs.

For instance, Meta created an enhanced version called RoBERTa, released as open-source code in July 2017. For training, it used “an order of magnitude more data than BERT,” the paper said, and leapt ahead on NLP leaderboards. A scrum followed.

## Scaling Parameters and Markets

For convenience, score is often kept by the number of an LLM’s parameters or weights, measures of the strength of a connection between two nodes in a neural network. BERT had 110 million, RoBERTa had 123 million, then BERT-Large weighed in at 354 million, setting a new record, but not for long.



As LLMs expanded into new applications, their size and computing requirements grew.

In 2020, researchers at OpenAI and Johns Hopkins University announced GPT-3, with a whopping 175 billion parameters, trained on a dataset with nearly a trillion words. It scored well on a slew of language tasks and even ciphered three-digit arithmetic.

"Language models have a wide range of beneficial applications for society," the researchers wrote.

## Experts Feel 'Blown Away'

Within weeks, people were using GPT-3 to create poems, programs, songs, websites and more. Recently, GPT-3 even wrote an academic paper about itself.

"I just remember being kind of blown away by the things that it could do, for being just a language model," said Percy Liang, a Stanford associate professor of computer science, speaking in a podcast.

GPT-3 helped motivate Stanford to create a center Liang now leads, exploring the implications of what it calls foundational models that can handle a wide variety of tasks well.

## Toward Trillions of Parameters

Last year, NVIDIA announced the Megatron 530B LLM that can be trained for new domains and languages. It debuted with tools and services for training language models with trillions of parameters.

“Large language models have proven to be flexible and capable … able to answer deep domain questions without specialized training or supervision,” Bryan Catanzaro, vice president of applied deep learning research at NVIDIA, said at that time.

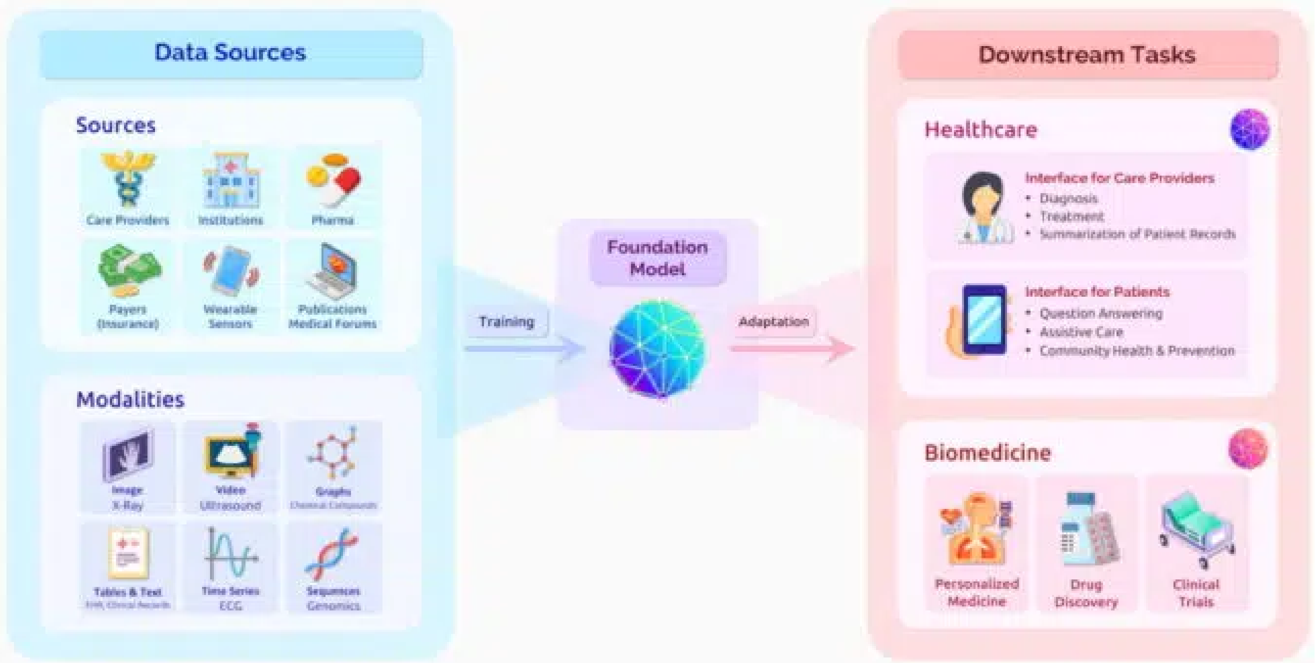
Making it even easier for users to adopt the powerful models, the NVIDIA Nemo LLM service debuted in September at GTC. It’s an NVIDIA-managed cloud service to adapt pretrained LLMs to perform specific tasks.

## Transformers Transform Drug Discovery

The advances LLMs are making with proteins and chemical structures are also being applied to DNA.

Researchers aim to scale their work with NVIDIA BioNeMo, a software framework and cloud service to generate, predict and understand biomolecular data. Part of the NVIDIA Clara Discovery collection of frameworks, applications and AI models for drug discovery, it supports work in widely used protein, DNA and chemistry data formats.

NVIDIA BioNeMo features multiple pretrained AI models, including the MegaMolBART model, developed by NVIDIA and AstraZeneca.



In their paper on foundational models, Stanford researchers projected many uses for LLMs in healthcare.

## LLMs Enhance Computer Vision

Transformers are also reshaping computer vision as powerful LLMs replace traditional convolutional AI models. For example, researchers at Meta AI and Dartmouth designed TimeSformer, an AI model that uses transformers to analyze video with state-of-the-art results.

Experts predict such models could spawn all sorts of new applications in computational photography, education and interactive experiences for mobile users.

In related work earlier this year, two companies released powerful AI models to generate images from text.

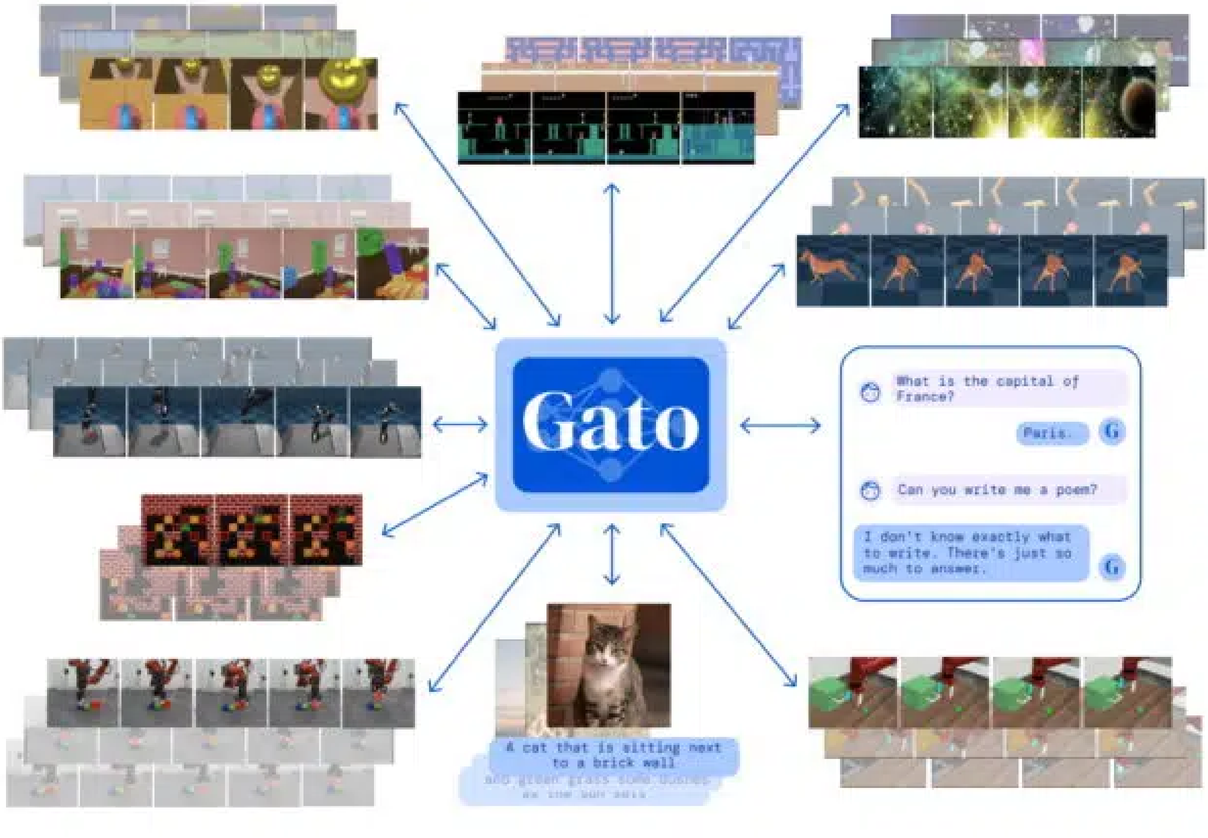
OpenAI announced DALL-E 2, a transformer model with 3.5 billion parameters designed to create realistic images from text descriptions. And recently, Stability AI, based in London, launched Stability Diffusion,

## Writing Code, Controlling Robots

LLMs also help developers write software. Tabnine — a member of NVIDIA Inception, a program that nurtures cutting-edge startups — claims it's automating up to 30% of the code generated by a million developers.

Taking the next step, researchers are using transformer-based models to teach robots used in manufacturing, construction, autonomous driving and personal assistants.

For example, DeepMind developed Gato, an LLM that taught a robotic arm how to stack blocks. The 1.2-billion parameter model was trained on more than 600 distinct tasks so it could be useful in a variety of modes and environments, whether playing games or animating chatbots.



What is the capital of France?

Paris.

Can you write me a poem?

I don't know exactly what to write. There's just so much to answer.

A cat that is sitting next to a brick wall

and green grass some bushes as the sun sets

The Gato LLM can analyze robot actions and images as well as text.

“By scaling up and iterating on this same basic approach, we can build a useful general-purpose agent,” researchers said in a paper posted in May.

It’s another example of what the Stanford center in a July paper called a paradigm shift in AI. “Foundation models have only just begun to transform the way AI systems are built and deployed in the world,” it said.

Learn how companies around the world are implementing LLMs with NVIDIA Triton for many use cases.

## AI Esperanto: Large Language Models Read Data With NVIDIA Triton

Companies bringing natural language processing to many markets are turning to Triton for AI inference.

October 5, 2022 by Rick Merritt
Julien Salinas wears many hats. He’s an entrepreneur, software developer and, until lately, a volunteer fireman in his mountain village an hour’s drive from Grenoble, a tech hub in southeast France.

He’s nurturing a two-year old startup, NLP Cloud, that’s already profitable, employs about a dozen people and serves customers around the globe. It’s one of many companies worldwide using NVIDIA software to deploy some of today’s most complex and powerful AI models.

NLP Cloud is an AI-powered software service for text data. A major European airline uses it to summarize internet news for its employees. A small healthcare company employs it to parse patient requests for prescription refills. An online app uses it to let kids talk to their favorite cartoon characters.

## Large Language Models Speak Volumes

It’s all part of the magic of natural language processing (NLP), a popular form of AI that’s spawning some of the planet’s biggest neural networks called large language models. Trained with huge datasets on powerful systems, LLMs can handle all sorts of jobs such as recognizing and generating text with amazing accuracy.

NLP Cloud uses about 25 LLMs today, the largest has 20 billion parameters, a key measure of the sophistication of a model. And now it’s implementing BLOOM, an LLM with a whopping 176 billion parameters.

Running these massive models in production efficiently across multiple cloud services is hard work. That’s why Salinas turns to NVIDIA Triton Inference Server.

## High Throughput, Low Latency

“Very quickly the main challenge we faced was server costs,” Salinas said, proud his self-funded startup has not taken any outside backing to date.

“Triton turned out to be a great way to make full use of the GPUs at our disposal,” he said.

For example, NVIDIA A100 Tensor Core GPUs can process as many as 10 requests at a time — twice the throughput of alternative software — thanks to FasterTransformer, a part of Triton that automates complex jobs like splitting up models across many GPUs.

FasterTransformer also helps NLP Cloud spread jobs that require more memory across multiple NVIDIA T4 GPUs while shaving the response time for the task.

Customers who demand the fastest response times can process 50 tokens — text elements like words or punctuation marks — in as little as half a second with Triton on an A100 GPU, about a third of the response time without Triton.

“That’s very cool,” said Salinas, who’s reviewed dozens of software tools on his personal blog.

## Touring Triton's Users

Around the globe, other startups and established giants are using Triton to get the most out of LLMs.

Microsoft's Translate service helped disaster workers understand Haitian Creole while responding to a 7.0 earthquake. It was one of many use cases for the service that got a 27x speedup using Triton to run inference on models with up to 5 billion parameters.

NLP provider Cohere was founded by one of the AI researchers who wrote the seminal paper that defined transformer models. It's getting up to 4x speedups on inference using Triton on its custom LLMs, so users of customer support chatbots, for example, get swift responses to their queries.

NLP Cloud and Cohere are among many members of the NVIDIA Inception program, which nurtures cutting-edge startups. Several other Inception startups also use Triton for AI inference on LLMs.

Tokyo-based rinna created chatbots used by millions in Japan, as well as tools to let developers build custom chatbots and AI-powered characters. Triton helped the company achieve inference latency of less than two seconds on GPUs.

In Tel Aviv, Tabnine runs a service that's automated up to 30% of the code written by a million developers globally (see a demo below). Its service runs multiple LLMs on A100 GPUs with Triton to handle more than 20 programming languages and 15 code editors.

Twitter uses the LLM service of Writer, based in San Francisco. It ensures the social network's employees write in a voice that adheres to the company's style guide. Writer's service achieves a 3x lower latency and up to 4x greater throughput using Triton compared to prior software.

If you want to put a face to those words, Inception member Ex-human, just down the street from Writer, helps users create realistic avatars for games, chatbots and virtual reality applications. With Triton, it delivers response times of less than a second on an LLM with 6 billion parameters while reducing GPU memory consumption by a third.

It's another example of how LLMs are expanding AI's horizons.

Triton is widely used, in part, because its versatile. The software works with any style of inference and any AI framework — and it runs on CPUs as well as NVIDIA GPUs and other accelerators.

## A Full-Stack Platform

Back in France, NLP Cloud is now using other elements of the NVIDIA AI platform.

For inference on models running on a single GPU, it's adopting NVIDIA TensorRT software to minimize latency. "We're getting blazing-fast performance with it, and latency is really going down," Salinas said.

The company also started training custom versions of LLMs to support more languages and enhance efficiency. For that work, it's adopting NVIDIA Nemo Megatron, an end-to-end framework for training and deploying LLMs with trillions of parameters.

The 35-year-old Salinas has the energy of a 20-something for coding and growing his business. He describes plans to build private infrastructure to complement the four public cloud services the startup uses, as well as to expand into LLMs that handle speech and text-to-image to address applications like semantic search.

"I always loved coding, but being a good developer is not enough: You have to understand your customers' needs," said Salinas, who posted code on GitHub nearly 200 times last year.

## No Hang Ups With Hangul: KT Trains Smart Speakers, Customer Call Centers With NVIDIA AI

South Korea's leading mobile operator builds billion-parameter large language models trained with the NVIDIA DGX SuperPOD platform and NeMo framework.

September 20, 2022 by Angie Lee
South Korea's most popular AI voice assistant, GiGA Genie, has conversed with 8 million people.

The AI-powered speaker from telecom company KT can control TVs, offer real-time traffic updates and complete a slew of other home-assistance tasks based on voice commands. It has mastered its conversational skills in the highly complex Korean language thanks to large language models (LLMs) — machine learning algorithms that can recognize, understand, predict and generate human languages based on huge text datasets.

The company's models are built using the NVIDIA DGX SuperPOD data center infrastructure platform and the NeMo framework for training and deploying LLMs with billions of parameters.

The Korean language, known as Hangul, reliably shows up in lists of the world's most challenging languages. It includes four types of compound verbs, and words are often composed of two or more roots.

KT — South Korea's leading mobile operator with over 22 million subscribers — improved the smart speaker's understanding of such words by developing LLMs. And through integration with Amazon Alexa, GiGA Genie can converse with users in English, too.

“With transformer-based models, we’ve achieved significant quality improvements for the GiGA Genie smart speaker, as well as our customer services platform AI Contact Center, or AICC,” said Hwijung Ryu, LLM development team lead at KT.

AICC is an all-in-one, cloud-based platform that offers AI voice agents and other customer service-related applications.

It can receive calls and provide requested information — or quickly connect customers to human agents for answers to more detailed inquiries. AICC without human intervention manages more than 100,000 calls daily across Korea, according to Ryu.

“LLMs enable GiGA Genie to gain better language understanding and generate more human-like sentences, and AICC to reduce consultation times by 15 seconds as it summarizes and classifies inquiry types more quickly,” he added.

## Training Large Language Models

Developing LLMs can be an expensive, time-consuming process that requires deep technical expertise and full-stack technology investments.

The NVIDIA AI platform simplified and sped up this process for KT.

“We trained our LLM models more effectively with NVIDIA DGX SuperPOD’s powerful performance — as well as NeMo’s optimized algorithms and 3D parallelism techniques,” Ryu said. “NeMo is continuously adopting new features, which is the biggest advantage we think it offers in improving our model accuracy.”

3D parallelism — a distributed training method in which an extremely large-scale deep learning model is partitioned across multiple devices — was crucial for training KT’s LLMs. NeMo enabled the team to easily accomplish this task with the highest throughput, according to Ryu.

“We considered using other platforms, but it was difficult to find an alternative that provides full-stack environments — from the hardware level to the inference level,” he added. “NVIDIA also provides exceptional expertise from product, engineering teams and more, so we easily solved several technical issues.”

Using hyperparameter optimization tools in NeMo, KT trained its LLMs 2x faster than with other frameworks, Ryu said. These tools allow users to automatically find the best configurations for LLM training and inference, easing and speeding the development and deployment process.

KT is also planning to use the NVIDIA Triton Inference Server to provide an optimized real-time inference service, as well as NVIDIA Base Command Manager to easily monitor and manage hundreds of nodes in its AI cluster.

“Thanks to LLMs, KT can release competitive products faster than ever,” Ryu said. “We also believe that our technology can drive innovation from other companies, as it can be used to improve their value and create innovative products.”

KT plans to release more than 20 natural language understanding and natural language generation APIs for developers in November. The application programming interfaces can be used for tasks including document summarization and classification, emotion recognition, and filtering of potentially inappropriate content.