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.

Author: 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.

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.

Learn more about breakthrough technologies for the era of AI and the metaverse at NVIDIA GTC , running online through Thursday, Sept. 22.

Watch NVIDIA founder and CEO Jensen Huang’s keynote address in replay below: