



Generative AI in Practice: Examples of Successful Enterprise Deployments



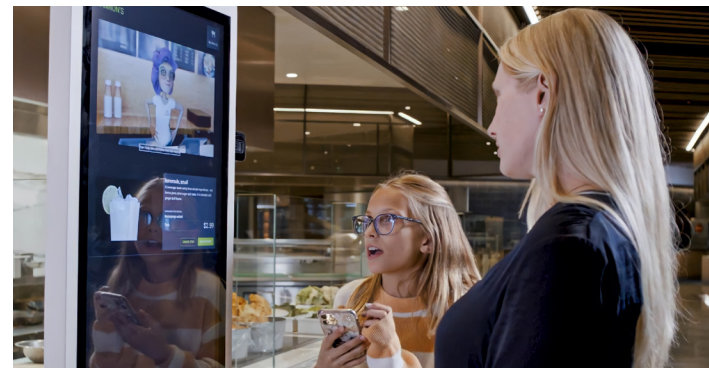
Generative AI and Large Language Models Are Transforming Industry and Society

While traditional AI approaches are based on rules and patterns, **generative AI** uses neural networks to identify the patterns and structures within existing data to generate new and original content. Generative AI is being widely used in applications such as text generation for marketing collateral, real-time translation services, dynamic coding, image creation for gaming characters, and accelerated drug discovery for healthcare. It's unlocking new opportunities to enhance business processes, improving productivity and efficiency.

One of the most popular and advanced applications of generative AI is language. **Large language models (LLMs)** represent a major advancement in artificial intelligence, promising better human and computer interactions. LLM sizes have been increasing 10X every year for the last few years,¹ and as these models grow in complexity and size, so do their

capabilities. For example, **OpenAI's GPT-4**, or the fourth-generation Generative Pretrained Transformer, can generate creative and technical content in the user's writing style, accept visual inputs and generate analyses, create long-form content, and extend conversations with the ability to follow users' intentions—all while being truthful and safe.

Enterprises are now starting to realize the benefits of LLMs, which promise flexible models for few-shot learning. This means models can perform tasks they weren't explicitly trained for, which is important in fields where data is scarce or impossible to obtain (i.e., due to expensive labeled data or privacy concerns) or where the capital and operating expense of training and managing the number of models is insurmountable. Today, we're seeing broader adoption of customized



LLMs that are tailored to understand an enterprise's unique vocabulary, its customer relationships, and the datasets on which the business runs. In the case studies that follow, we'll explore how various organizations built mission-critical LLMs, powered by NVIDIA DGX™ systems and the NVIDIA NeMo™ framework, which is part of NVIDIA AI Enterprise and optimized to run on every DGX system, enabling them to simplify their business and increase customer satisfaction, while achieving the fastest and highest return. DGX systems include DGX H200 and DGX B200, which are powered by NVIDIA Tensor Core GPUs and Intel® Xeon® Processors and expressly designed for training large AI models such as LLMs.

1. Julien Simon. [Large Language Models: A New Moore's Law?](#) Hugging Face. October 26, 2021.



KT, South Korea's leading mobile operator, wanted to improve its GiGA Genie smart speaker and AI Contact Center (AICC) customer services platform. The AI-powered

speaker can control TVs, offer real-time traffic updates, and complete a slew of other home-assistance tasks based on voice commands, while the AICC service offers AI voice agents and other customer service-related applications. KT needed to master the highly complex Korean language as well as English for integration with Amazon Alexa, and scale its services to support over 22 million subscribers. KT trained billion-parameter large language models using NVIDIA DGX SuperPOD™ and the NeMo framework and **achieved 2X faster** training on LLMs versus other frameworks. 3D parallelism techniques in NeMo enabled fast training with the highest throughput, while hyperparameter tools sped model development. KT plans to develop APIs for other tasks like document summarization and classification, emotion recognition, and filtering of inappropriate content.

2. Based on projected time to train 300B tokens on 175B parameter GPT-3 model on DGX H100 systems, NeMo 24.05, FP8

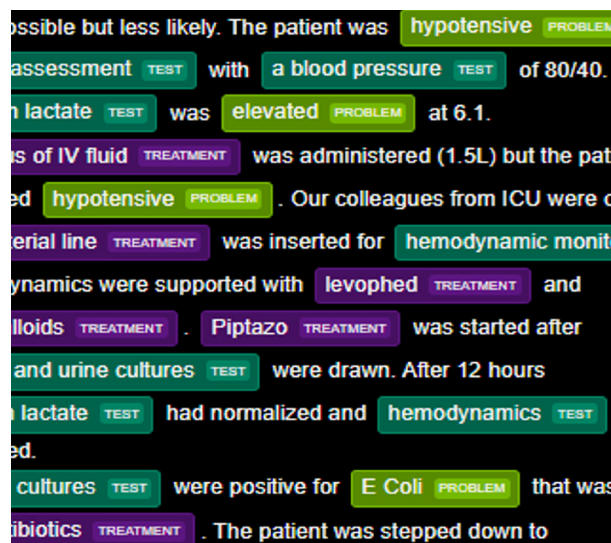
Training a 175B-parameter GPT-3 model on 8 DGX systems takes 2.5 months. Leveraging a DGX SuperPOD with 95 DGX H200 systems powered by NVIDIA H200 Tensor Core GPUs and Intel Xeon Processors can reduce this training time to less than a week, showcasing a dramatic improvement in efficiency and scalability².



The University of Florida's academic health center, UF Health, has teamed up with NVIDIA to develop a transformer model that generates synthetic clinical data.

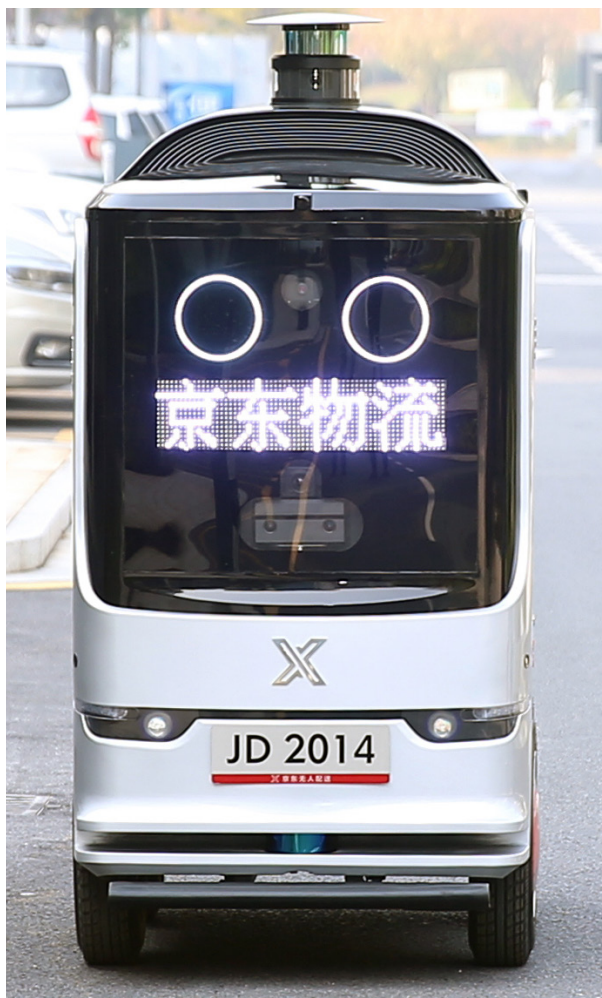
Trained on a decade of data representing more than 2 million patients, SynGatorTron is a language model that can create synthetic patient profiles that mimic the health records from which it has learned. The 5-billion-parameter model is the largest clinical language generator to date. SynGatorTron was developed using the NeMo framework and trained on HiPerGator AI, the university's in-house DGX SuperPOD system. Using this synthetic data, researchers can create tools, models, and tasks with no privacy concerns. These AI-generated records can then be used to supplement and balance out real healthcare datasets used to train other neural networks, so they better represent the population. As 80 percent of information in electronic health records (EHRs) is unstructured in clinical notes, LLM models can be used to easily surface relevant clinical data used in

downstream diagnostic and predictive tasks. These models also improve the patient experience through smarter chatbots and reduce physician burnout through improved transcription and summarization tools, faster documentation, and better decision-support systems.



AI Sweden is the Swedish national center for applied artificial intelligence. AI Sweden is digitizing Sweden's history, as captured in the media and publications in the Swedish Royal Library, and building language models from this unstructured data that can be commercialized in enterprise applications. Its goal is to train large-scale Swedish LLM models and make them publicly accessible through open source, thereby accelerating LLM industry applications in Sweden. AI Sweden is addressing the problem of low-resource languages, or languages lacking many data resources, by combining datasets from morphologically similar languages. Using the Berzelius supercomputer powered by DGX SuperPOD and the NeMo framework, AI Sweden is able to train, serve, and leverage the power of a 100-billion-parameter model for Nordic languages (Swedish, Norwegian, Danish, and Icelandic) and push the boundaries of applied AI in businesses and the public sector.

Powered by NVIDIA Blackwell GPUs and Intel Xeon Processors, the NVIDIA DGX B200 delivers 3X AI training and 15X AI inference performance on Mixture of Experts (MoE) LLM with 1.8T parameters over prior generation.



JD.com is a leading supply chain-based technology and service provider and the largest online retailer in China. JD Explore Academy, its research and development division, is utilizing DGX SuperPOD to develop LLMs for the application of smart customer service, smart retail, smart logistics, IoT, healthcare, and more. JD is training GPT-3-like, large-scale models to investigate how to transfer knowledge—both efficiently and securely—from large-scale datasets to the parameters of a pretraining model to improve downstream LLM tasks, like sentiment analysis, dialogue, and translation. JD was able to train its 5-billion-parameter model easily using the NeMo framework's out-of-the-box recipes, which come with all the required hyperparameters making manual tuning unnecessary. JD was able to achieve 8X faster training on DGX SuperPOD versus a cluster of GPU-based legacy servers powered by NVIDIA. This technology is helping JD develop the next generation of super deep learning technology, which links data across its various businesses like retail, healthcare, intelligent supply chain, digital intelligence, and industrial services.



A leading European telecommunications provider with over 80 million daily users across Europe and Asia wanted to improve customer loyalty and prevent churn.

The company also offers services outside of its core network-related business, including finance and banking, cloud computing, and ecommerce. To lower costs and improve employee efficiency, the provider wanted to develop virtual assistants to support various functions (e.g., customer support, billing) and to support its various businesses beyond telecommunications (e.g., pay TV, banking). The company leveraged DGX SuperPOD and the NeMo framework to train an over-3-billion-parameter model and has deployed this into production. This technology enabled the company to develop different chatbots with slightly different interfaces and domains. Using automatic speech recognition and synthesis, the provider created chatbots that sound humanlike and can understand both vocabulary and intonation. There are 160 nationalities within the company's network, so it even developed a different model adapted to each region. Today, 80 percent of calls leverage this AI service. The provider is also able to develop custom LLM services to other verticals such as healthcare, where it's providing a mobile application for online consultation to a chain of clinics.

Deliver Enterprise-Grade, Superhuman Language and Visual Understanding

Generative AI can transform how enterprises operate by enabling unparalleled quality for language, audio, and visual-based services such as summary generation, auto completion, intent detection, automatic dialogue generation, translation, and many more. As LLMs are applied to new use cases, their complexity and size has increased exponentially.

LLMs require computing on a scale that goes beyond mainstream, commonplace NLP apps to state-of-the-art applications, customized for each enterprise and delivered at the speed of business. The NVIDIA DGX platform provides optimized AI software, infrastructure, and access to AI experts, so you can stand up your own world-class LLM

architecture in record time. For optimal deployment of NeMo models, use **NVIDIA NIM™** to speed deployment with ease of use, manageability, and security.

- > Learn more about **NVIDIA DGX SuperPOD**, the turnkey AI data center solution.
- > Start using the **NVIDIA NeMo framework** to build state-of-the-art AI models.
- > Learn more about **NVIDIA NIM**, accelerated inference microservices.

Ready to Get Started?

To accelerate generative AI development on NVIDIA x86-based DGX systems, powered by NVIDIA Tensor Core GPUs and Intel Xeon Processors, visit nvidia.com/dgx

© 2024 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, DGX, DGX SuperPOD, and NeMo are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated. All other trademarks are property of their respective owners. 3411830. AUG24

