



NATURAL LANGUAGE UNDERSTANDING IN PRACTICE: EXAMPLES OF SUCCESSFUL ENTERPRISE DEPLOYMENTS



LARGE LANGUAGE MODELS ARE TRANSFORMING INDUSTRY AND SOCIETY

Natural language understanding (NLU) can be a huge help to any business: organizations of every size collect written texts—from legal documents, patient records, financial documents, code, websites, emails, and more. The ability to leverage that unstructured information and improve customer satisfaction is crucial to staying competitive.

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 years¹, and as these models grow in complexity and size, so do their capabilities. For example, **OpenAI's GPT-3**, or the third-generation Generative Pretrained Transformer, can perform specialized tasks such as answering questions, writing essays, summarizing text, translating languages, and generating computer code.

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 Natural Language Processing (NLP) that's tailored to understand an enterprise's unique vocabulary, their customer relationships, and the datasets on which their business runs. In the case studies that follow, we'll explore how various organizations built mission-critical NLP, powered by NVIDIA DGX SuperPOD™ and NVIDIA NeMo Megatron, enabling them to simplify their business and increase customer satisfaction, while achieving the fastest and highest return.

¹JULIEN SIMON, LARGE LANGUAGE MODELS: A NEW MOORE'S LAW? HUGGING FACE, OCTOBER 26, 2021.





JD.com is a leading supply chain-based technology and service provider and the largest online retailer in China. JD Explore Academy, their research and development division, is utilizing NVIDIA DGX SuperPOD to develop NLP for the application of smart customer service, smart retail, smart logistics, IoT, healthcare, and more. They're 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 NLP tasks, like sentiment analysis, dialogue, and translation. JD was able to train their 5-billion-parameter model easily using NVIDIA NeMo Megatron's out-of-the-box recipes, which come with all the required hyperparameters so they don't need to tune them themselves. JD was able to achieve 8X faster training on DGX SuperPOD powered by NVIDIA A100 Tensor Core GPUs and AMD EPYC CPUs vs a cluster of NVIDIA V100 GPU-based systems. This technology is helping JD develop the next generation of super deep learning technology, which links data across their various businesses like retail, healthcare, intelligent supply chain, digital intelligence, and industrial services.



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. Their goal is to train large-scale Swedish NLP models and make them publicly accessible through open source, thereby accelerating NLP 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 NVIDIA NeMo Megatron 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 languages) and push the boundaries of applied AI in businesses and the public sector.



nephritis and urinary symptoms. A preliminary assessment was established. Other causes of hypotensive were possible but less likely. The patient was hypotensive. A preliminary assessment with a blood pressure test of 6.1. Serum lactate was elevated at 6.1. A bolus of IV fluid was administered (1.5L) but the patient remained hypotensive. Our colleagues from ICU inserted an arterial line for hemodynamic monitoring. Hemodynamics were supported with levophed. Levofloxacins and Piptazo was started. Blood and urine cultures were drawn. After 12 hours, serum lactate had normalized and hemodynamics improved. Blood cultures were positive for E. Coli. The patient was stepped down to the medical ward. The patient was stepped down to the medical ward to complete a total of 14 days of treatment.



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 NVIDIA Megatron framework and trained on HiPerGator-AI, the university's in-house NVIDIA DGX SuperPOD system, powered by NVIDIA A100 Tensor Core GPUs and AMD EPYC CPUs. 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) are unstructured in clinical notes, NLP 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.



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 their core network-related business, including finance and banking, cloud computing, and e-commerce. 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 their various businesses beyond telecommunications (e.g., pay TV, banking). They leveraged NVIDIA DGX SuperPOD and NVIDIA NeMo Megatron to train an over-3-billion-parameter model and have deployed this into production. This technology enabled them to develop different chatbots with slightly different interfaces and domains. Using automatic speech recognition and synthesis, the provider created chatbots that sound human-like and can understand both vocabulary and intonation. There are 160 nationalities within their network, so they 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 NLP services to other verticals such as healthcare, where they're providing a mobile application for online consultation to a chain of clinics.

DELIVER ENTERPRISE-GRADE, SUPERHUMAN LANGUAGE UNDERSTANDING

LLMs can transform how enterprises operate by enabling unparalleled quality for language-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 DGX SuperPOD NLP solution for enterprise provides optimized hardware topology, advanced algorithms, and access to AI experts, so you can stand up your own world-class NLP architecture in record time.

- > Learn more about **NVIDIA DGX SuperPOD**, the turnkey AI data center solution powered by NVIDIA A100 Tensor Core GPUs and AMD EPYC CPUs.
- > Start using the **NVIDIA NeMo Megatron** framework to build state-of-the-art AI models.

Test drive NeMo Megatron on DGX systems with a free, short-term trial on NVIDIA Base Command: www.nvidia.com/try-base-command

Partner
Logo

