What Are Large Language Models Used For?

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

Author: 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 .

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.

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.

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.

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.

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.

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.

Learn more about large language models .

Original URL: https://blogs.nvidia.com/blog/2023/01/26/what-are-large-language-models-used-for/