# What Are Foundation Models?

Foundation models are AI neural networks trained on massive unlabeled datasets to handle a wide variety of jobs from translating text to analyzing medical images.

Author: Rick Merritt

The mics were live and tape was rolling in the studio where the Miles Davis Quintet was recording dozens of tunes in 1956 for Prestige Records.

When an engineer asked for the next song's title, Davis shot back , "I'll play it, and tell you what it is later."

Like the prolific jazz trumpeter and composer, researchers have been generating AI models at a feverish pace, exploring new architectures and use cases. Focused on plowing new ground, they sometimes leave to others the job of categorizing their work.

A team of more than a hundred Stanford researchers collaborated to do just that in a 214-page paper released in the summer of 2021.

They said transformer models , large language models (LLMs) and other neural networks still being built are part of an important new category they dubbed foundation models.

A foundation model is an AI neural network — trained on mountains of raw data, generally with unsupervised learning — that can be adapted to accomplish a broad range of tasks, the paper said.

"The sheer scale and scope of foundation models from the last few years have stretched our imagination of what's possible," they wrote.

Two important concepts help define this umbrella category: Data gathering is easier, and opportunities are as wide as the horizon.

Foundation models generally learn from unlabeled datasets, saving the time and expense of manually describing each item in massive collections.

Earlier neural networks were narrowly tuned for specific tasks. With a little fine-tuning, foundation models can handle jobs from translating text to analyzing medical images.

Foundation models are demonstrating "impressive behavior," and they're being deployed at scale, the group said on the website of its research center formed to study them. So far, they've posted more than 50 papers on foundation models from in-house researchers alone.

"I think we've uncovered a very small fraction of the capabilities of existing foundation models, let alone future ones," said Percy Liang, the center's director, in the opening talk of the first workshop on foundation models.

In that talk, Liang coined two terms to describe foundation models:

Emergence refers to AI features still being discovered, such as the many nascent skills in foundation models. He calls the blending of AI algorithms and model architectures homogenization , a trend that helped form foundation models. (See chart below.)

The field continues to move fast.

A year after the group defined foundation models, other tech watchers coined a related term — generative AI . It's an umbrella term for transformers, large language models, diffusion models and other neural networks capturing people's imaginations because they can create text, images, music, software and more.

Generative AI has the potential to yield trillions of dollars of economic value, said executives from the venture firm Sequoia Capital who shared their views in a recent AI Podcast .

“We are in a time where simple methods like neural networks are giving us an explosion of new capabilities,” said Ashish Vaswani, an entrepreneur and former senior staff research scientist at Google Brain who led work on the seminal 2017 paper on transformers.

That work inspired researchers who created BERT and other large language models , making 2018 “a watershed moment” for natural language processing, a report on AI said at the end of that year.

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. Then it applied the technology to its search engine so users could ask questions in simple sentences.

In 2020, researchers at OpenAI announced another landmark transformer, GPT-3 . Within weeks, people were using it to create poems, programs, songs, websites and more.

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

Their work also showed how large and compute-intensive these models can be. GPT-3 was trained on a dataset with nearly a trillion words, and it sports a whopping 175 billion parameters, a key measure of the power and complexity of neural networks.

“I just remember being kind of blown away by the things that it could do,” said Liang, speaking of GPT-3 in a podcast .

The latest iteration, ChatGPT — trained on 10,000 NVIDIA GPUs — is even more engaging, attracting over 100 million users in just two months. Its release has been called the iPhone moment for AI because it helped so many people see how they could use the technology.

About the same time ChatGPT debuted, another class of neural networks, called diffusion models, made a splash. Their ability to turn text descriptions into artistic images attracted casual users to create amazing images that went viral on social media.

The first paper to describe a diffusion model arrived with little fanfare in 2015. But like transformers, the new technique soon caught fire.

Researchers posted more than 200 papers on diffusion models last year, according to a list maintained by James Thornton, an AI researcher at the University of Oxford.

In a tweet , Midjourney CEO David Holz revealed that his diffusion-based, text-to-image service has more than 4.4 million users. Serving them requires more than 10,000 NVIDIA GPUs mainly for AI inference, he said in an interview (subscription required).

Hundreds of foundation models are now available. One paper catalogs and classifies more than 50 major transformer models alone (see chart below).

The Stanford group benchmarked 30 foundation models, noting the field is moving so fast they did not review some new and prominent ones.

Startup NLP Cloud , a member of the NVIDIA Inception program that nurtures cutting-edge startups, says it uses about 25 large language models in a commercial offering that serves airlines, pharmacies and other users. Experts expect that a growing share of the models will be made open source on sites like Hugging Face’s model hub .

Foundation models keep getting larger and more complex, too.

That’s why — rather than building new models from scratch — many businesses are already customizing pretrained foundation models to turbocharge their journeys into AI.

One venture capital firm lists 33 use cases for generative AI, from ad generation to semantic search.

Major cloud services have been using foundation models for some time. For example, Microsoft Azure worked with NVIDIA to implement a transformer for its Translator service. It helped disaster workers understand Haitian Creole while they were responding to a 7.0 earthquake.

In February, Microsoft announced plans to enhance its browser and search engine with ChatGPT and related innovations. “We think of these tools as an AI copilot for the web,” the announcement said.

Google announced Bard , an experimental conversational AI service. It plans to plug many of its products into the power of its foundation models like LaMDA, PaLM, Imagen and MusicLM.

“AI is the most profound technology we are working on today,” the company’s blog wrote.

Startup Jasper expects to log $75 million in annual revenue from products that write copy for companies like VMware. It’s leading a field of more than a dozen companies that generate text, including Writer, an NVIDIA Inception member.

Other Inception members in the field include Tokyo-based rinna that’s created chatbots used by millions in Japan. In Tel Aviv, Tabnine runs a generative AI service that’s automated up to 30% of the code written by a million developers globally.

Researchers at startup Evozyne used foundation models in NVIDIA BioNeMo to generate two new proteins . One could treat a rare disease and another could help capture carbon in the atmosphere.

BioNeMo, a software platform and cloud service for generative AI in drug discovery, offers tools to train, run inference and deploy custom biomolecular AI models. It includes MegaMolBART , a generative AI model for chemistry developed by NVIDIA and AstraZeneca.

“Just as AI language models can learn the relationships between words in a sentence, our aim is that neural networks trained on molecular structure data will be able to learn the relationships between atoms in real-world molecules,” said Ola Engkvist, head of molecular AI, discovery sciences and R&D; at AstraZeneca, when the work was announced .

Separately, the University of Florida’s academic health center collaborated with NVIDIA researchers to create GatorTron . The large language model aims to extract insights from massive volumes of clinical data to accelerate medical research.

A Stanford center is applying the latest diffusion models to advance medical imaging. NVIDIA also helps healthcare companies and hospitals use AI in medical imaging , speeding diagnosis of deadly diseases.

Another new framework, NVIDIA NeMo framework , aims to let any business create its own billion- or trillion-parameter transformers to power custom chatbots, personal assistants and other AI applications.

It created the 530-billion parameter Megatron-Turing Natural Language Generation model ( MT-NLG ) that powers TJ, the Toy Jensen avatar that gave part of the keynote at NVIDIA GTC last year.

Foundation models — connected to 3D platforms like NVIDIA Omniverse — will be key to simplifying development of the metaverse , the 3D evolution of the internet. These models will power applications and assets for entertainment and industrial users.

Factories and warehouses are already applying foundation models inside digital twins, realistic simulations that help find more efficient ways to work.

Foundation models can ease the job of training autonomous vehicles and robots that assist humans on factory floors and logistics centers. They also help train autonomous vehicles by creating realistic environments like the one below.

New uses for foundation models are emerging daily, as are challenges in applying them.

Several papers on foundation and generative AI models describing risks such as:

amplifying bias implicit in the massive datasets used to train models,

introducing inaccurate or misleading information in images or videos, and

violating intellectual property rights of existing works.

“Given that future AI systems will likely rely heavily on foundation models, it is imperative that we, as a community, come together to develop more rigorous principles for foundation models and guidance for their responsible development and deployment,” said the Stanford paper on foundation models.

Current ideas for safeguards include filtering prompts and their outputs, recalibrating models on the fly and scrubbing massive datasets.

“These are issues we’re working on as a research community,” said Bryan Catanzaro, vice president of applied deep learning research at NVIDIA. “For these models to be truly widely deployed, we have to invest a lot in safety.”

It’s one more field AI researchers and developers are plowing as they create the future.

Original URL: https://blogs.nvidia.com/blog/2023/03/13/what-are-foundation-models/