What is generative AI, what are foundation models, and why do they matter?

Artificial intelligence >

March 8, 2023

By Manish Goyal
Shobhit Varshney
Eniko Rozsa

3 min read

Since its launch in November 2022, OpenAI's ChatGPT has captured the imagination of both consumers and enterprise leaders by demonstrating the potential generative AI has to dramatically transform the ways we live and work. As the scope of its impact on society continues to unfold, business and government organizations are still racing to react, creating policies about employee use of the technology or even restricting access to ChatGPT.

The most prudent among them have been assessing the ways in which they can apply AI to their organizations and preparing for a future that is already here. The most advanced among them are shifting their thinking from AI being a bolt-on afterthought, to reimagining critical workflows with AI at the core.

How generative AI—like ChatGPT—is already transforming businesses

The global generative AI market is approaching an inflection point, with a valuation of USD 8 billion and an estimated CAGR of 34.6% by 2030. With more than 85 million jobs expected to go unfilled by that time, creating more intelligent operations with AI and automation is required to deliver the efficiency, effectiveness and experiences that business leaders and stakeholders expect.

Generative AI presents a compelling opportunity to augment employee efforts and make the enterprise more productive. But as C-Suite leaders research generative AI solutions, they are uncovering more questions: Which use cases will deliver the most value for my business? Which AI technology is best suited for my needs? Is it secure? Is it sustainable? How is it governed? And how do I ensure my AI projects succeed?

Having worked with foundation models for a number of years, IBM Consulting, IBM Technology and IBM Research have developed a grounded point of view on what it takes to derive value from responsibly deploying AI across the enterprise.

Differences between existing enterprise AI in enterprises and new generative AI capabilities

As the name suggests, generative AI *generates* images, music, speech, code, video or text, while it interprets and manipulates pre-existing data. Generative AI is not a new concept: machine-learning techniques behind generative AI have evolved over the past decade. The latest approach is based on a neural network architecture, coined "transformers." Combining transformer architecture with unsupervised learning, large foundation models emerged that outperform existing benchmarks capable of handling multiple data modalities.

These large models are called foundational models, as they serve as the starting point for the development of more advanced and complex models. By building on top of a foundation model, we can create more specialized and sophisticated models tailored to specific use cases or domains. Early examples of models, like GPT-3, BERT, T5 or DALL-E, have shown what's possible: input a short prompt and the system generates an entire essay, or a complex image, based on your parameters.

Large Language Models (LLMs) were explicitly trained on large amounts of text data for NLP tasks and contained a significant number of parameters, usually exceeding 100 million. They facilitate the processing and generation of natural language text for diverse tasks. Each model has its strengths and weaknesses and the choice of which one to use depends on the specific NLP task and the characteristics of the data being analyzed. Choosing the correct LLM to use for a specific job requires expertise in LLMs.

BERT is designed to understand bidirectional relationships between words in a sentence and is primarily used for task classification, question answering and named entity recognition. GPT, on the other hand, is a unidirectional transformer-based model primarily used for text generation tasks such as language translation, summarization, and content creation. T5 is also a transformer-based model, however, it differs from BERT and GPT in that it is trained using a text-to-text approach and can be fine-tuned for various natural language processing tasks such as language translation, summarization and responding to questions.

Acceleration and reduced time to value

Being pre-trained on massive amounts of data, these foundation models deliver huge acceleration in the AI development lifecycle, allowing businesses to focus on fine tuning for their specific use cases. As opposed to building custom NLP models for each domain, foundation models are enabling enterprises to shrink the time to value from months to weeks. In client engagements, IBM

Consulting is seeing up to 70% reduction in time to value for NLP use cases such as call center transcript summarization, analyzing reviews and more.

Deploying foundation models responsibly

Given the cost to train and maintain foundation models, enterprises will have to make choices on how they incorporate and deploy them for their use cases. There are considerations specific to use cases and decision points around cost, effort, data privacy, intellectual property and security. It is possible to use one or more deployment options within an enterprise trading off against these decision points.

Foundation models will dramatically accelerate AI adoption in business by reducing labeling requirements, which will make it easier for businesses to experiment with AI, build efficient AI-driven automation and applications, and deploy AI in a wider range of mission-critical situations. The goal for IBM Consulting is to bring the power of foundation models to every enterprise in a frictionless hybrid-cloud environment.

For more information, see how generative AI can be used to maximize experiences, decision-making and business value, and how IBM Consulting brings a valuable and responsible approach to AI.