5

U.S. Government Accountability Office

Artificial Intelligence: Generative AI Technologies and Their Commercial Applications

GAO-24-106946
Q&A Report to Congressional Requesters
June 20, 2024

Why This Matters

Use of generative artificial intelligence (AI) has exploded to over 100 million users globally due to recently enhanced capabilities and user interest. This growth has created both excitement and debate about the technology's potential to revolutionize entire industries, such as health care, education, and software engineering. Generative AI technology may dramatically increase productivity and transform daily tasks across much of society. However, it may also displace workers, spread disinformation, and present risks to national security and the environment.

For this technology assessment, we were asked to describe generative AI and key aspects of its development. This report is the first in a body of work looking at generative AI. In future reports, we plan to assess best practices and other factors considered for developing and deploying generative AI tools, societal and environmental effects of the use of generative AI, and federal development and adoption of generative AI technologies. To perform this assessment, we conducted literature reviews and interviewed several leading companies developing generative AI technologies. This report provides an overview of how generative AI works, how it differs from other kinds of AI, and examples of its use across various industries.

Key Takeaways

- Generative AI differs from other AI systems in its ability to create novel content, in the vast volumes of data it requires for training, and in the greater size and complexity of its models.
- Generative AI systems employ several model architectures, or underlying structures. These systems, referred to as neural networks, are modeled loosely on the human brain and recognize patterns in data.
- Commercial developers have created a wide range of generative AI models that produce text, code, image, and video outputs. Developers have also created products and services that enhance existing products or support customized development and refinement of models to meet customer needs. Their benefits and risks are still unclear for many applications.

What is generative AI?

Unlike AI technologies that focus on classification and prediction, generative AI can create content such as text, images, audio, or video when prompted by a user. Generative AI systems create responses that are based on data, often text and images sourced from the internet at large. Users may produce outputs from the software by issuing a query commonly known as a prompt. Many of the generative AI systems now available allow users to prompt the system in natural language.

AI models

For the purposes of this report, we use “model” to refer to the result of an algorithm “trained” on a set of data. Training is the iterative process of feeding data (called training data) through an optimization process to improve model performance.

Of the generative models, foundation and frontier models are among the most capable. Foundation models are models trained on a broad array of data that can be adapted to a wide range of tasks – for example, summarizing text, writing code, composing music, or creating images. Foundation models may be fine-tuned or augmented to tailor to customer needs, or they may be integrated into multiple AI systems across a variety of areas. Frontier models are the most advanced foundation models, with respect to new or powerful capabilities, and may pose increased risks related to fact checking or contextual awareness of outputs and responses than foundation models.

Primary differences between generative AI and conventional AI

The ability to create, or generate, novel content sets generative AI apart from other AI. By contrast, conventional AI does tasks of classification and prediction, such as identifying objects in a photograph or forecasting a storm.

On the one hand, creation of novel content increases the risk of undesirable output, such as potentially copyrighted content, offensive text, biased output, misinformation, and explicit imagery. On the other, it has wide flexibility and applicability in the breadth of tasks it can perform.

Another difference is that generative AI typically requires a much larger dataset for training—ranging from millions to trillions of data points. Generative AI model performance generally increases as the size or quality of the training data increase. In addition, developers must consider whether to incorporate user prompts in training and, as with other AI, when and how often to update the training data and model.

Generative model performance also generally increases with larger model size, which is typically indicated by the number of parameters (internal variables of the model that are “learned” through the training data). The smaller generative AI models have millions of parameters while the largest boast hundreds of billions or more.

Secondary differences between generative AI and conventional AI

Generative AI also differs in a few other notable ways. Many generative AI models use natural language as input, meaning any text could be used as a prompt. But this ability also means the quality of the prompt has a significant effect on the response. Prompts serve as the starting point for a model’s generation. Models tend to respond like-for-like to input. For example, a prompt that contains harmful content is more likely to produce harmful output.[1]

Another difference is that it is usually less clear why a generative model produced a certain output. This “black box” effect is notable with the large neural networks underlying generative AI. As a result, reproducibility and accountability are lower, degrading the user’s ability to evaluate and understand the model’s inner workings.

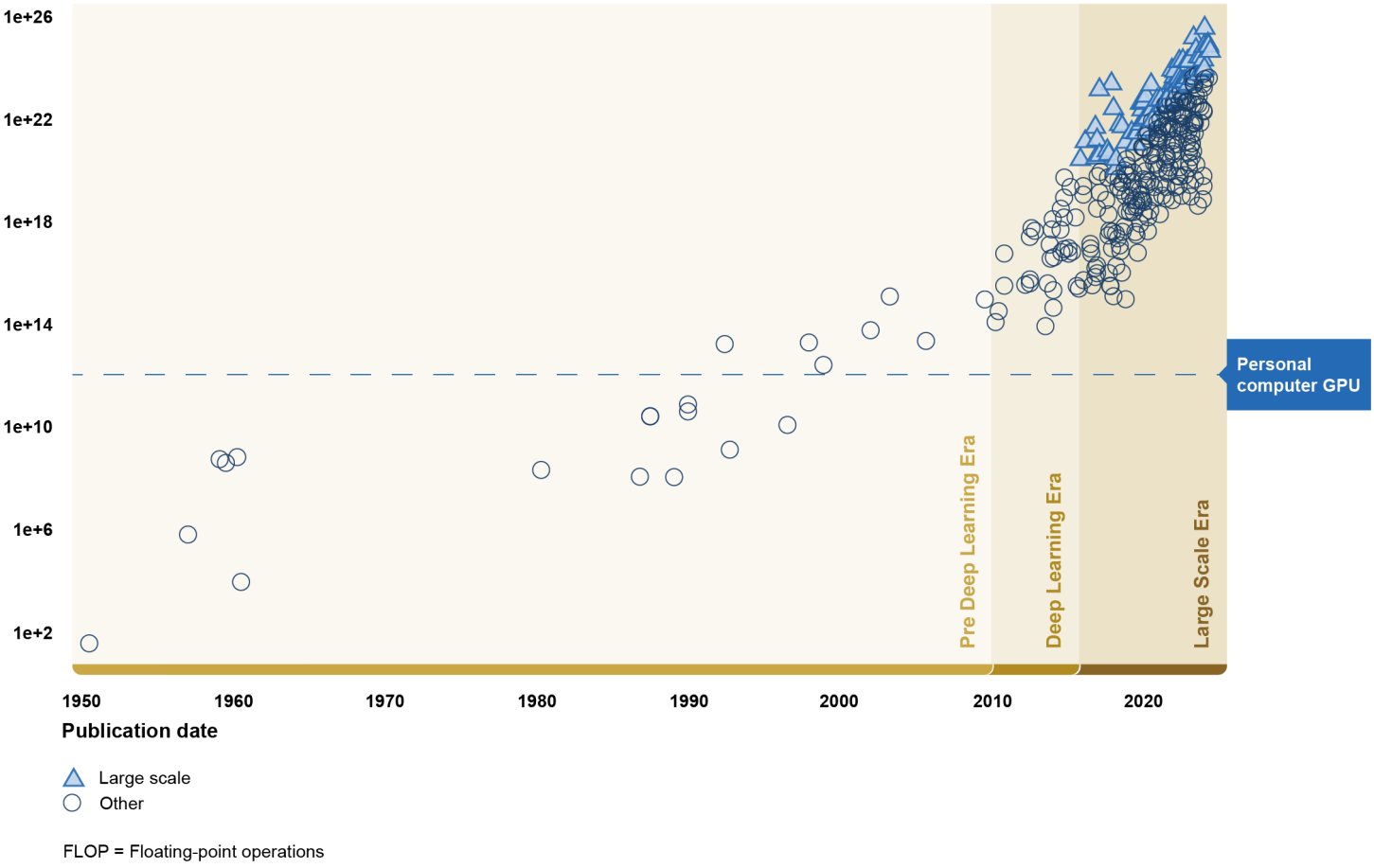
In addition, larger generative AI models may have abilities that are not predicted by extrapolating from those of smaller models. Such “emergent” abilities may be

unintentional or unexpected and may not be apparent until a model is fully developed or deployed.[2]

Trends in generative AI

The history of generative AI starts in the 1950s, but major breakthroughs came within the last decade, with the introduction of particularly effective and widely adopted generative models (see fig. 1). These new models are made possible by advances in computing power (often called compute), new machine learning architectures (see next section), and the vast amounts of data produced in the digital age. Deep learning systems, which consist of neural networks that contain a large number of hidden layers, became prevalent in the 2010s and led to advancements in computer vision and natural language processing. Since 2020, advancements have created AI systems more capable of human-like conversation and generative tasks. In the large-scale era, recent developments include employing a mixture of specialized models and multimodal models that use data types such as image, text, speech, and video inputs and outputs.

Figure 1. Compute Required to Train Notable Machine Learning Systems Over Time



Source: "Training Compute of Notable Machine Learning Systems Over Time" by Epoch AI is licensed under CC BY 4.0 DEED (https://epochai.org/data/epochdb?startLargeScaleEra=2015-9-1&splitLargeScaleEra=false&plotRegressions=false&preset=Three%20eras%20of%20compute&systemNames=hide&showDoublingTimes=false). | GAO-24-106946

Several recent surveys have shown increasing awareness and adoption of generative AI in daily life. In March 2023, 58 percent of U.S. adults were aware of ChatGPT, with 14 percent reporting using it for entertainment, to learn something new, or for work.[3] In April 2023, 22 percent of respondents to an annual global survey of business leaders reported using generative AI regularly at work.[4]

What is model architecture, and what types of model architectures are used in generative AI?

Commercial developers employ a variety of architectures to develop generative AI models. Model architecture refers to the underlying structure of a machine learning model. For example, generative AI systems use a neural network.[5] Neural networks consist of an arrangement of interconnected nodes.[6] Generative AI architectures include the following:

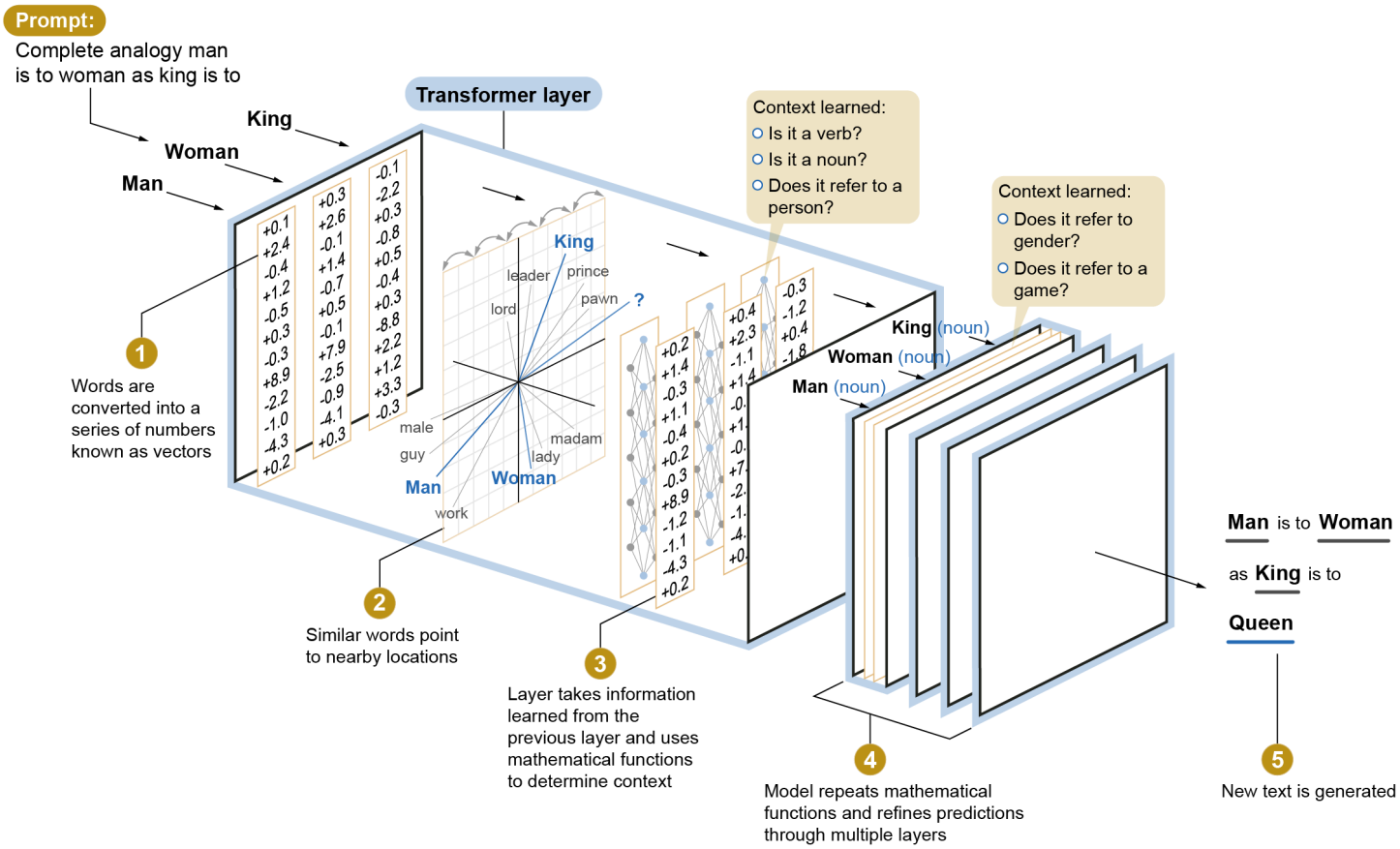
- Variational autoencoders (VAE) use two neural networks. One neural network converts the data to a simpler representation (encoder). The other neural network reconstructs from the simpler representation (decoder)—to understand and optimize data and determine efficient ways of regenerating data. For example, VAEs can be used for novel image generation.

- Generative adversarial networks (GAN) also use two neural networks, which compete against each other to generate more authentic new data. For example, GANs can be used for generating novel images and video content.

- Transformers are a type of neural network that applies widely to natural language processing by tracking the relationships between words within sentences to learn context and meaning. For example, transformer models can process and generate text, such as summarizing or generating documents. See figure 2 for a depiction of a simple transformer architecture.

- Diffusion is a technique for creating images where a sample of random noise is generated, then repeatedly fed through a predictive neural network to iteratively remove noise until matching a prompt. For example, diffusion models can be used for image and video generation, which includes correction by replacing missing data.

Figure 2. Simple Transformer Architecture



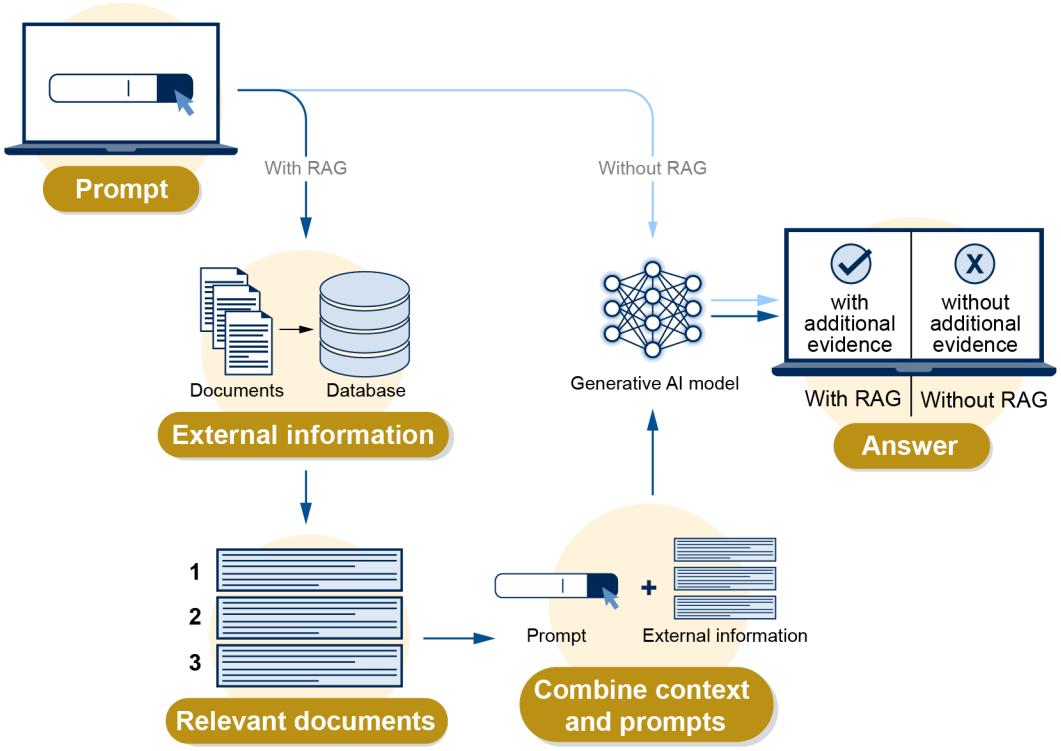
Source: GAO (analysis); GAO adaptation of "But what is a GPT? Visual intro to transformers" by Grant Sanderson (illustration). | GAO-24-106946

What factors enabled commercial development of generative AI systems?

A convergence of factors enabled the rapid development of generative AI. One factor was the availability of large amounts of training data, which generative AI models require to produce high-quality outputs.[7] These data are commonly obtained from publicly available information on the internet. One potential downside is that there may be problems surrounding the use of the generated content if the data used to train the model are covered by copyright protections. In addition, models inherit the characteristics of the data they are trained on, which can include bias, inaccuracy, and impropriety.

Another factor that has enabled rapid development is the creation of a set of techniques to refine model outputs. One such technique is known as retrieval-augmented generation (RAG). RAG enhances the accuracy and reliability of a generative AI model by retrieving contextual information from sources not included in the initial training data. RAG may be implemented as part of initial model training, alongside fine-tuning, or, most commonly, deployed in-line with user prompts. Figure 3 shows an example of how RAG can improve a prompt.

Figure 3: Prompt Answer Generation with versus without Retrieval-Augmented Generation



RAG = Retrieval-augmented generation

Source: GAO (analysis and illustrations). | GAO-24-106946

Another technique is reinforcement learning from human feedback (RLHF), which helps a model provide answers that are more meaningful and fit-for-purpose. With RHLF, generative AI models typically undergo further training where humans evaluate and rank the models' outputs, and then change their parameters to better suit the human preferences. An alternative technique, constitutional AI (CAI), provides feedback from a second AI system, which follows a short list of principles. For example, CAI can assess the AI system's initial response to determine whether it is problematic (e.g., harmful, unethical, or illegal). Both RLHF and CAI can guide the behavior of AI systems in either malicious or beneficial directions.

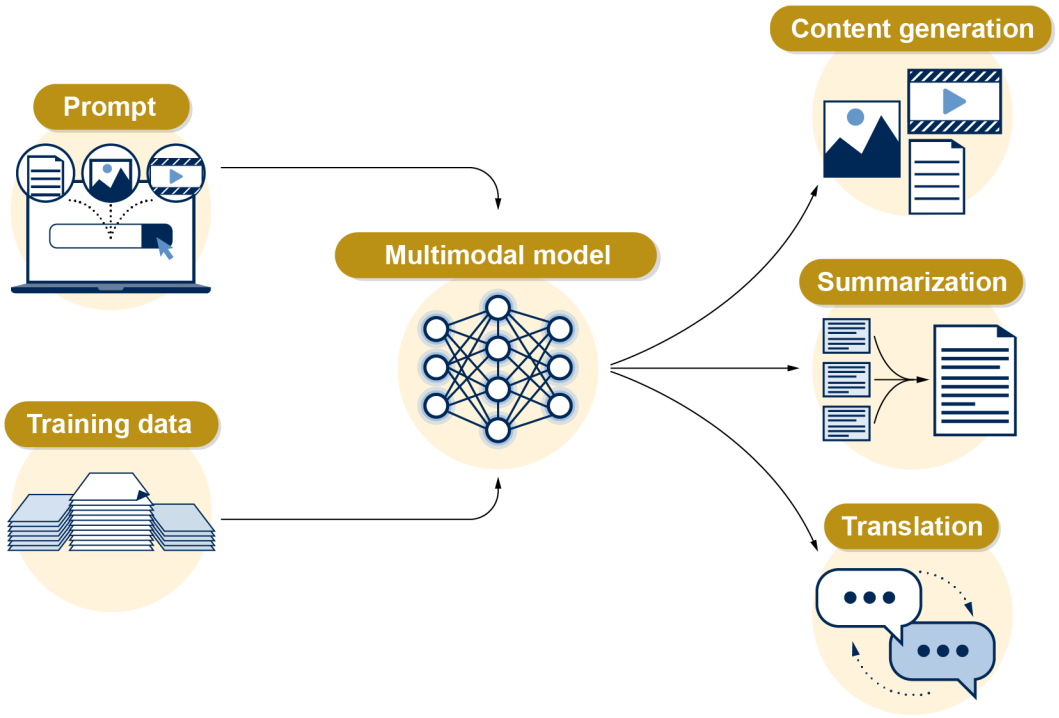
A final factor in the emergence of generative AI is the availability of vast compute capacity. Training one large generative AI model can take tens of thousands of processors running for months and may cost several hundred million dollars. Sustaining the current compute growth rate for training models may present challenges, including cost, limited microchip supply, and the complexities of managing the training process across many processors. To address some of these issues, companies are exploring options such as developing specialized microchips, increasing microchip production capacity, and offering cloud-based model development and deployment options.

What are some examples of commercially developed generative AI products?

Commercial developers have created several types of generative AI tools, including large language models (LLMs), multimodal models, image generation models, and video generation models. These AI tools can receive a prompt in plain language and generate an output (e.g. text, an image, or a video) that is statistically representative of the training data. Some advanced LLMs can also analyze and produce code in various programming languages.

Multimodal models are LLMs that can also accept images, audio, or video as inputs (see fig. 4).

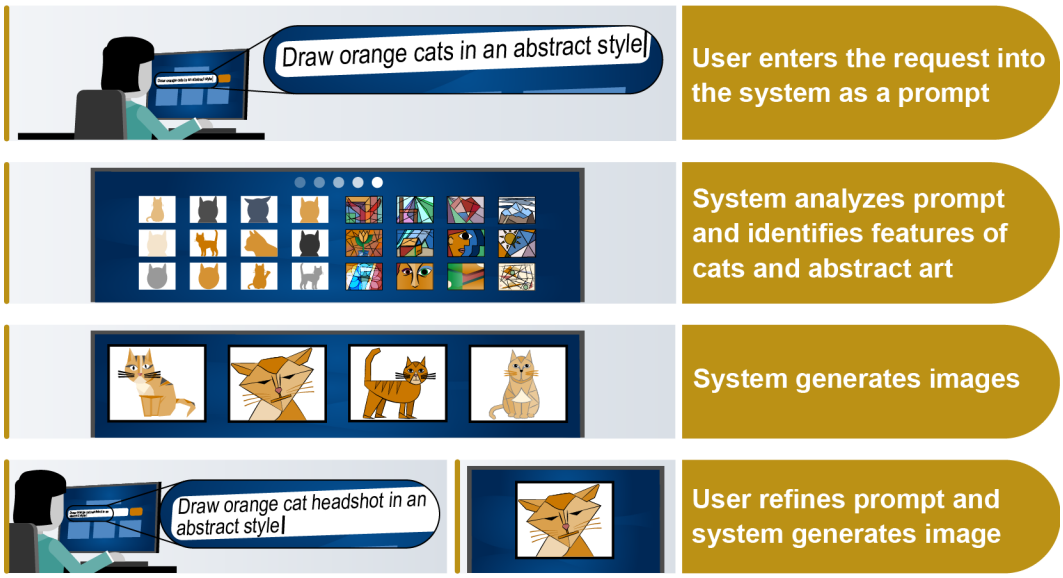
Figure 4: Process Flow for Multimodal Generative AI Models



Source: GAO (analysis and illustrations). | GAO-24-106946

Image generation models take text or images as input and generate or edit images (see fig.5).

Figure 5: Example of a Generative AI System Creating an Image from Prompts



Source: GAO (analysis and illustrations). | GAO-24-106946

Along with image generation models, some commercial developers also have created video generation models. Video generation models can take text, images or video as input and generate or edit videos. As of April 2024, five of the companies we selected for interviews had image generation models (Amazon, Google, Meta, Open AI, and Stability AI) and four had video generation models (Google, Meta, OpenAI, and Stability AI). Many commercial developers have developed at least one LLM, and several have developed multimodal, image, and video generation models. See table 1 for more details.

Table 1. Capabilities of Selected Commercially Developed Generative Artificial Intelligence Models, as of April 2024

Company	Model[a]	Release date	Capabilities								
			Text-to-				Code-to-text	Image-to-			Audio-to-text	Video-to-text
			text	code	image	video		text	image	video		
Amazon	Titan Text Express	Sept. 2023	x	x								
	Titan Image Generator	Nov. 2023			x							
Anthropic	Claude 2.1	Nov. 2023	x									
	Claude 3	Mar. 2024	x					x				
Google	Gemini (Ultra, Pro, Nano)	Dec. 2023	x		x			x	x		x	x
	PaLM 2	May. 2023	x	x			x					
	Imagen 2	Dec. 2023			x							
	Lumiere	Jan. 2024				x				x		
Microsoft	Phi 3	Apr. 2024	x	x			x					
	Florence	Mar. 2023						x				x
Meta	Llama 3	Apr. 2024	x	x								
	Code Llama	Aug. 2023	x				x					
	Emu	Sep. 2023			x							
	Emu Edit	Nov. 2023			x				x			
	Emu Video	Nov. 2023				x				x		
OpenAI	GPT-4 Turbo	Nov. 2023	x	x			x	x				
	GPT-3.5 Turbo	Nov. 2023	x	x			x					
	DALL-E 3	Oct. 2023			x							
	Sora	Feb. 2024				x						
Stability AI	Stable Beluga 1	July 2023	x									
	Stable Beluga 2	July 2023	x									
	Stable LM 2	Apr. 2024	x									
	Stable Code	Jan. 2024		x								
	Stable Diffusion SDXL 1.0	July 2023			x							
	Stable Video Diffusion	Nov. 2023								x		

Source: GAO summary of publicly available information. | GAO-24-106946

Note: Text-to-text means that a model takes text as an input prompt and generates a text output. The other capabilities in the table can be read in the same way.

[a]We identified models in this table from selected generative AI commercial developers with whom we met. For a more detailed discussion of our selection criteria, see “How GAO Did This Study” below.

Other products and services

Some companies also use generative AI in their products and services. They can use the models to generate product descriptions and images for advertising, a more conversational search engine experience, improved web search results, and product comparisons. In software development, generative AI models can generate code suggestions in real-time based on a developer’s comments in natural language and on existing code.

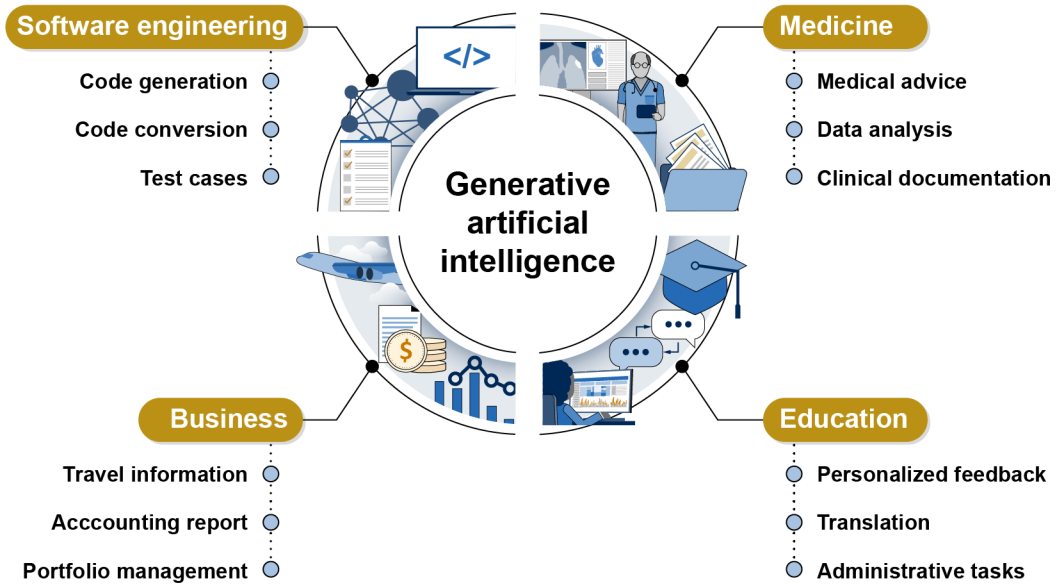
Other companies have systems and platforms designed to support the development of generative AI. For example, representatives from one of the companies told us that the company provides products and services, such as

data center infrastructure, processing hardware, and development environments to help businesses build AI models with the data of those businesses.

How can generative AI be used in selected industries?

Generative AI has many potential applications in fields such as software engineering, business, education, and health care (see fig. 6). The information below is based on a literature review of future applications. We did not independently confirm these capabilities or assess the maturity of these applications. One expert said that many generative AI products require supervision of trained staff to prevent bad outcomes from edge cases. We plan to assess potential societal and environmental effects of generative AI in a future technology assessment.

Figure 6: Summary of Potential Applications for Generative Artificial Intelligence



Source: GAO (analysis and illustrations). | GAO-24-106946

Generative AI in software engineering

Generative AI has potential uses in software engineering, such as generating software, code conversion, and software testing. For example, it has the potential to generate new software and convert code to another programming language. Additionally, AI tools possess the capability to generate test cases and automate testing functions. Generative AI may also enable people without software engineering skills to develop software prototypes. However, generative AI could also enable people of all skill levels to develop malicious software.

Generative AI in business

In business, generative AI has the potential to assist with travel arrangements, analyze accounting data, and help manage investments. For example, generative AI may help answer traveler inquiries by generating information about destinations, attractions, events, travel services, local customs, and visa requirements. In the accounting and auditing sectors, generative AI may automatically create coherent, informative, and well-structured financial reports based on historical data, including balance sheets, income statements, and tax documents. This process may significantly reduce the operational risks of manual errors. In the finance sector, generative AI may play a role in digital advisory services and provide portfolio management services without significant human

involvement. However, generative AI may also contribute to worker displacement, especially in jobs with routine tasks, such as administration or basic analysis.

Generative AI in education

Generative AI also has potential applications in education, such as personalized learning, language learning, and automated administrative tasks. Generative AI may be used to provide personalized feedback on writing assignments, such as essays and research papers. This feedback may cover issues such as grammar, organization, and content. Generative AI also may be used to support language learners by providing personalized feedback on grammar, vocabulary, and assistance in language translation in a classroom setting. Furthermore, generative AI may be used to automate administrative tasks, such as grading assessments and answering frequently asked questions, freeing up teachers to focus on other tasks. However, generative AI may produce poor quality papers or provide biased information.

Generative AI in health care

In health care, generative AI has potential applications in areas such as data analysis, medical advice, and clinical documentation. Generative AI may be used in patient data analysis and advising health care professionals on recommended courses of action. Another potential application of generative AI models in health care is the automation of clinical documentation that provides clinical administration support. For example, clinicians may leverage generative AI capabilities to generate draft clinical notes more swiftly and accurately. Despite potential benefits, generative AI in a medical setting presents potential privacy issues related to patients' health data.

How GAO Did This Study

To describe the technologies that enable the development of generative AI tools, we gathered information regarding the companies' various models, tools, products, and services that enable the development of generative AI. We selected the following commercial developers of generative AI: Amazon, Anthropic, Google, Meta, Microsoft, Nvidia Corporation, OpenAI, and Stability. These companies are among the leading AI organizations that, in 2023, made voluntary commitments to the White House to manage risks posed by AI. We also reviewed relevant publicly available documentation, such as white papers, model cards, and guidance documents to identify further information regarding the companies' generative AI products. Additionally, we interviewed representatives of those selected commercial developers of generative AI.

In addition, we conducted a literature search using a variety of databases. We conducted our search to identify relevant publications such as government reports, conference papers, and scholarly and peer-reviewed publications from 2022 through 2024. We analyzed these sources to identify terminology related to generative AI, understand the technologies that enable the development of generative AI models, identify primary and secondary differences of generative AI from conventional AI systems (i.e., models focused on classification and prediction), identify various model architectures employed by generative AI models, and determine potential applications of generative AI across selected sectors. We selected the sectors based on prevalence of information within our literature search results that discussed potential applications of generative AI.

We conducted our work from July 2023 to June 2024 in accordance with all sections of GAO's Quality Assurance Framework that are relevant to technology assessments. The framework requires that we plan and perform the engagement

to obtain sufficient and appropriate evidence to meet our stated objectives and to discuss any limitations to our work. We believe that the information and data obtained, and the analysis conducted, provide a reasonable basis for any findings and conclusions in this product.

List of Addressees

The Honorable Gary C. Peters
Chairman
Committee on Homeland Security and Governmental Affairs
United States Senate

The Honorable Ed Markey
United States Senate

We are sending copies of this report to appropriate congressional committees. In addition, the report is available at no charge on the GAO website at https://www.gao.gov.

GAO Contact Information

For more information, contact: Brian Bothwell, Director, Science, Technology Assessment, and Analytics, bothwellb@gao.gov, (202) 512-6888 or Kevin Walsh, Director, Information Technology and Cybersecurity, walshk@gao.gov, (202) 512-6151.

Sarah Kaczmarek, Acting Managing Director, kaczmareks@gao.gov, (202) 512-4800.

A. Nicole Clowers, Managing Director, Congressional Relations, ClowersA@gao.gov, (202) 512-4400.

Staff Acknowledgments: R. Scott Fletcher (Assistant Director), Jessica Steele (Assistant Director), Sean Manzano (Analyst-in-Charge), Owen Baron, Christopher Cooper, Nathan Hanks, Igor Koshelev, Anika McMillon, Ben Shouse, Andrew Stavisky, Ashley Stewart, and Wes Wilhelm.

Connect with GAO on Facebook, Flickr, Twitter, and YouTube. Subscribe to our RSS Feeds or Email Updates. Listen to our Podcasts.

Visit GAO on the web at https://www.gao.gov.

This work of the United States may include copyrighted material, details at https://www.gao.gov/copyright.

Endnotes

[1]Additional training and prompt filtering are some of the mitigation strategies that can limit this trend. Models may still produce harmful output if users input carefully crafted prompts that exploit the generative AI, according to experts.

[2]Some experts attribute the apparent property of emergent abilities to the choice of metrics.

[3]Emily A. Vogels. Pew Research Center, *A majority of Americans have heard of ChatGPT, but few have tried it themselves* (Washington, D.C.: May 24, 2023). https://www.pewresearch.org/short-reads/2023/05/24/a-majority-of-americans-have-heard-of-chatgpt-but-few-have-tried-it-themselves/

[4]McKinsey Global Institute, McKinsey & Company, *The state of AI in 2023: Generative AI's breakout year* (2023). https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year

[5]GAO, *Science & Tech Spotlight: Deepfakes*, GAO-20-379SP (Washington, D.C., Feb. 20, 2020).

[6]GAO, *Artificial Intelligence in Intelligence in Natural Hazard Modeling: Severe Storms, Hurricanes, Floods, and Wildfires*, GAO-24-106213 (Washington, D.C., Dec. 14, 2023).

[7]GAO, *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning Technologies for Medical Diagnostics*, GAO-22-104629 (Washington, D.C., Sep. 29, 2022).