McKinsey Digital

Exploring opportunities in the generative AI value chain

Generative AI is giving rise to an entire ecosystem, from hardware providers to application builders, that will help bring its potential for business to fruition.

This article is a collaborative effort by Tobias Härlin, Gardar Björnsson Rova, Alex Singla, Oleg Sokolov, and Alex Sukharevsky, representing views from McKinsey Digital.

April 2023

Over the course of 2022 and early 2023, tech innovators unleashed generative AI en masse, dazzling business leaders, investors, and society at large with the technology's ability to create entirely new and seemingly human-made text and images.

The response was unprecedented.

In just five days, one million users flocked to ChatGPT, OpenAI's generative AI language model that creates original content in response to user prompts. It took Apple more than two months to reach the same level of adoption for its iPhone. Facebook had to wait ten months and Netflix more than three years to build the same user base.

And ChatGPT isn't alone in the generative AI industry. Stability AI's Stable Diffusion, which can generate images based on text descriptions, garnered more than 30,000 stars on GitHub within 90 days of its release—eight times faster than any previous package.[1]

This flurry of excitement isn't just organizations kicking the tires. Generative AI use cases are already taking flight across industries. Financial services giant Morgan Stanley is testing the technology to help its financial advisers better leverage insights from the firm's more than 100,000 research reports.[2] The government of Iceland has partnered with OpenAI in its efforts to preserve the endangered Icelandic language.[3] Salesforce has integrated the technology into its popular customer-relationship-management (CRM) platform.[4]

The breakneck pace at which generative AI technology is evolving and new use cases are coming to market has left investors and business leaders scrambling to understand the generative AI ecosystem. While deep dives into CEO strategy and the potential economic value that the technology could create globally across industries are forthcoming, here we share a look at the generative AI value chain composition. Our aim is to provide a foundational understanding that can serve as a starting point for assessing investment opportunities in this fast-paced space. Our assessments are based on primary and secondary research, including more than 30 interviews with business founders, CEOs, chief scientists, and business leaders working to commercialize the technology; hundreds of market reports and articles; and proprietary McKinsey research data.

A brief explanation of generative AI

To understand the generative AI value chain, it's helpful to have a basic knowledge of what generative AI is[5] and how its capabilities differ from the "traditional" AI technologies that companies use to, for example, predict client churn, forecast product demand, and make next-best-product recommendations.

A key difference is its ability to create new content. This content can be delivered in multiple modalities, including text (such as articles or answers to questions), images that look like photos or paintings, videos, and 3-D representations (such as scenes and landscapes for video games).

Even in these early days of the technology's development, generative AI outputs have been jaw-droppingly impressive, winning digital-art awards and scoring among or close to the top 10 percent of test takers in numerous tests, including the US bar exam for lawyers and the math, reading, and writing portions of the SATs, a college entrance exam used in the United States.[6]

Most generative AI models produce content in one format, but multimodal models that can, for example, create a slide or web page with both text and graphics based on a user prompt are also emerging.

All of this is made possible by training neural networks (a type of deep learning algorithm) on enormous volumes of data and applying "attention mechanisms," a technique that helps AI models understand what to focus on. With these mechanisms, a generative AI system can identify word patterns, relationships, and the context of a

[1] Guido Appenzeller, Matt Bornstein, Martin Casado, and Yoko Li, "Art isn't dead; it's just machine generated," Andreessen Horowitz, November 16, 2022.
[2] Hugh Son, "Morgan Stanley is testing an OpenAI-powered chatbot for its 16,000 financial advisors," CNBC, March 14, 2023.
[3] "Government of Iceland: How Iceland is using GPT-4 to preserve its language," OpenAI, March 14, 2023.
[4] "Salesforce announces Einstein GPT, the world's first generative AI for CRM," Salesforce, March 7, 2023.
[5] "What is generative AI?" McKinsey, January 19, 2023.
[6] "GPT-4," OpenAI, March 14, 2023.

user's prompt (for instance, understanding that "blue" in the sentence "The cat sat on the mat, which was blue" represents the color of the mat and not of the cat). Traditional AI also might use neural networks and attention mechanisms, but these models aren't designed to create new content. They can only describe, predict, or prescribe something based on existing content.

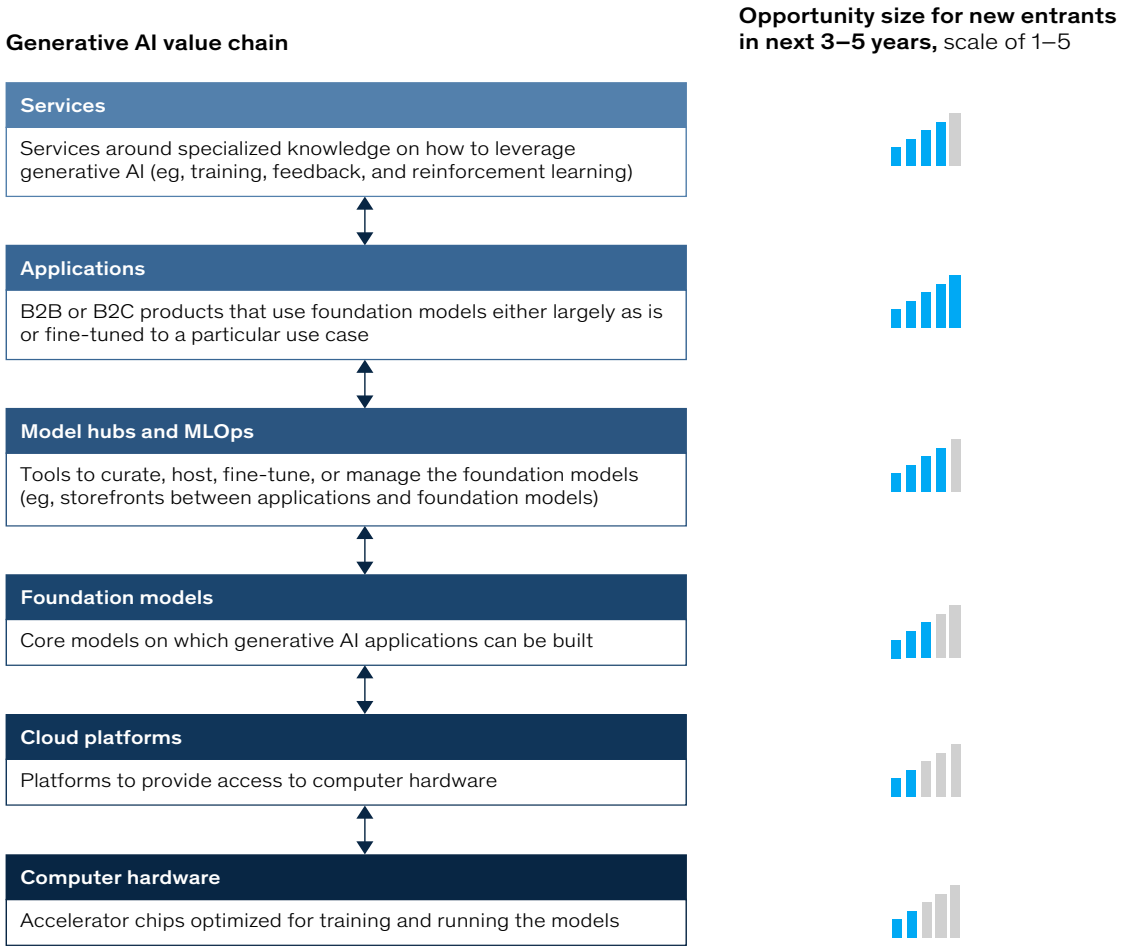
The value chain: Six links, but one outshines them all

As the development and deployment of generative AI systems gets under way, a new value chain is emerging to support the training and use of this powerful technology. At a glance, one might think it's quite similar to a traditional AI value chain. After all, of the six top-level categories—computer hardware, cloud platforms, foundation models, model hubs and machine learning operations (MLOps), applications, and services—only foundation models are a new addition (Exhibit 1).

However, a deeper look reveals some significant differences in market opportunities. To begin with, the underpinnings of generative AI systems are appreciably more complex than most traditional AI systems. Accordingly, the time, cost, and expertise associated with delivering them give rise to significant headwinds for new entrants and small companies across much of the

Exhibit 1

There are opportunities across the generative AI value chain, but the most significant is building end-user applications.



value chain. While pockets of value exist throughout, our research suggests that many areas will continue to be dominated by tech giants and incumbents for the foreseeable future.

The generative AI application market is the section of the value chain expected to expand most rapidly and offer significant value-creation opportunities to both incumbent tech companies and new market entrants. Companies that use specialized or proprietary data to fine-tune applications can achieve a significant competitive advantage over those that don't.

Computer hardware

Generative AI systems need knowledge—and lots of it—to create content. OpenAI's GPT-3, the generative AI model underpinning ChatGPT, for example, was trained on about 45 terabytes of text data (akin to nearly one million feet of bookshelf space).[7]

It's not something traditional computer hardware can handle. These types of workloads require large clusters of graphic processing units (GPUs) or tensor processing units (TPUs) with specialized "accelerator" chips capable of processing all that data across billions of parameters in parallel.

Once training of this foundational generative AI model is completed, businesses may also use such clusters to customize the models (a process called "tuning") and run these power-hungry models within their applications. However, compared with the initial training, these latter steps require much less computational power.

While there are a few smaller players in the mix, the design and production of these specialized AI processors is concentrated. NVIDIA and Google dominate the chip design market, and one player, Taiwan Semiconductor Manufacturing Company Limited (TSMC), produces almost all of the accelerator chips. New market entrants face high start-up costs for research and development. Traditional hardware designers must develop the specialized skills, knowledge, and computational capabilities necessary to serve the generative AI market.

Cloud platforms

GPUs and TPUs are expensive and scarce, making it difficult and not cost-effective for most businesses to acquire and maintain this vital hardware platform on-premises. As a result, much of the work to build, tune, and run large AI models occurs in the cloud. This enables companies to easily access computational power and manage their spend as needed.

Unsurprisingly, the major cloud providers have the most comprehensive platforms for running generative AI workloads and preferential access to the hardware and chips. Specialized cloud challengers could gain market share, but not in the near future and not without support from a large enterprise seeking to reduce its dependence on hyperscalers.

Foundation models

At the heart of generative AI are foundation models. These large deep learning models are pretrained to create a particular type of content and can be adapted to support a wide range of tasks. A foundation model is like a Swiss Army knife—it can be used for multiple purposes. Once the foundation model is developed, anyone can build an application on top of it to leverage its content-creation capabilities. Consider OpenAI's GPT-3 and GPT-4, foundation models that can produce human-quality text. They power dozens of applications, from the much-talked-about chatbot ChatGPT to software-as-a-service (SaaS) content generators Jasper and Copy.ai.

Foundation models are trained on massive data sets. This may include public data scraped from Wikipedia, government sites, social media, and books, as well as private data from large databases. OpenAI, for example, partnered with Shutterstock to train its image model on Shutterstock's proprietary images.[8]

[7] "What is generative AI?" January 19, 2023; and Kindra Cooper, "OpenAI GPT-3: Everything you need to know," Springboard, November 1, 2021.
[8] "Shutterstock partners with OpenAI and leads the way to bring AI-generated content to all," Shutterstock, October 25, 2022.

Developing foundation models requires deep expertise in several areas. These include preparing the data, selecting the model architecture that can create the targeted output, training the model, and then tuning the model to improve output (which entails labeling the quality of the model's output and feeding it back into the model so it can learn).

Today, training foundation models in particular comes at a steep price, given the repetitive nature of the process and the substantial computational resources required to support it. In the beginning of the training process, the model typically produces random results. To improve its next output so it is more in line with what is expected, the training algorithm adjusts the weights of the underlying neural network. It may need to do this millions of times to get to the desired level of accuracy. Currently, such training efforts can cost millions of dollars and take months. Training OpenAI's GPT-3, for example, is estimated to cost $4 million to $12 million.[9] As a result, the market is currently dominated by a few tech giants and start-ups backed by significant investment (Exhibit 2). However, there is work in progress toward making smaller models that can deliver effective results for some tasks and training that is more efficient, which could eventually open the market to more entrants. We already see that some start-ups have achieved certain success in developing their own models—Cohere, Anthropic, and AI21, among others, build and train their own large language models (LLMs). Additionally, there is a scenario where most big companies would want to have LLMs working in their environments—such as for a higher level of data security and privacy, among other reasons—and some players (such as Cohere) already offer this kind of service around LLMs.

It's important to note that many questions have yet to be answered regarding ownership and rights over the data used in the development of this nascent technology—as well as over the outputs produced—which may influence how the technology evolves (see sidebar, "Some of the key issues shaping generative AI's future").

Some of the key issues shaping generative AI's future

Amid the enormous enthusiasm, many questions have emerged surrounding generative AI technology, the answers to which will likely shape future development and use. Following are three of the most important questions to consider when evaluating how the generative AI ecosystem will evolve:

— *Can copyrighted or personal data be used to train models?* When training foundation models, developers typically "scrape" data from the internet. This can sometimes include copyrighted images, news articles, social media data, personal data protected by the General Data Protection Regulation (GDPR), and more. Current laws and regulations are ambiguous in terms of the implications of such practices. Precedents will likely evolve to place limits on scraping proprietary data that may be posted online or enable data owners to restrict or opt out of search indexes so their data can't easily be found online. New compensation models for data owners will also likely emerge.

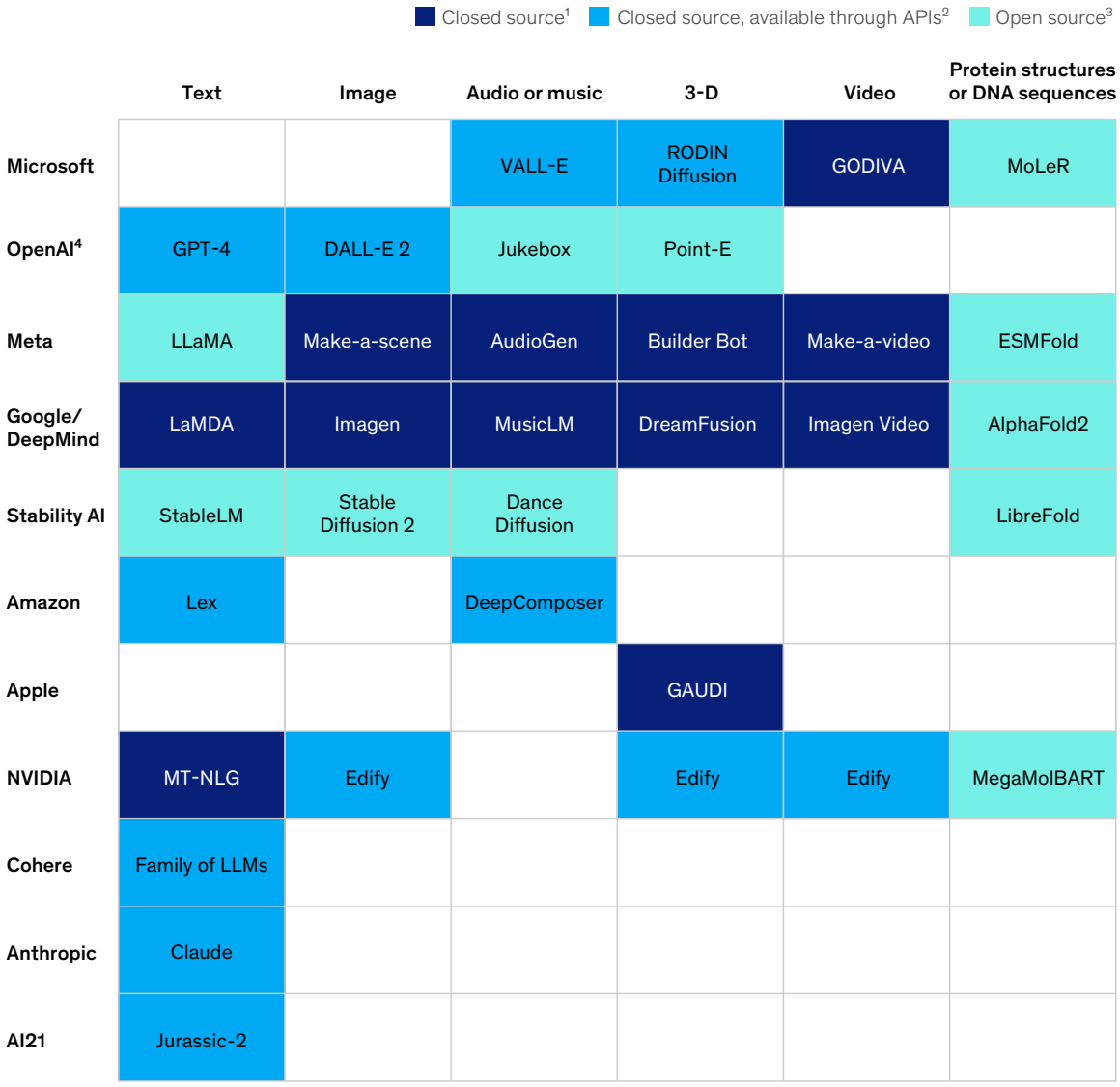
— *Who owns the creative outputs?* Current laws and regulations also do not clearly answer who owns the copyright on the final "output" of a generative AI system. Several potential actors can share or own exclusive rights to the final outputs, such as the data set owner, model developer, platform owner, prompt creator, or the designer who manually refines and delivers the final generative AI output.

— *How will organizations manage the quality of generative AI outputs?* We have already seen numerous examples of systems providing inaccurate, inflammatory, biased, or plagiarized content. It's not clear whether models will be able to eliminate such outputs. Ultimately, all companies developing generative AI applications will need processes for assessing outputs at the use case level and determining where the potential harm should limit commercialization.

[9] Kif Leswing and Jonathan Vanian, "ChatGPT and generative AI are booming, but the costs can be extraordinary," CNBC, March 13, 2023; and Toby McClean, "Machines are learning from each other, but it's a good thing," *Forbes*, February 3, 2021.

Exhibit 2

Examples of generative AI models from some of the early providers show there are many options available for each modality, several of which are open source.



	Text	Image	Audio or music	3-D	Video	Protein structures or DNA sequences
Microsoft			VALL-E	RODIN Diffusion	GODIVA	MoLeR
OpenAI[4]	GPT-4	DALL-E 2	Jukebox	Point-E		
Meta	LLaMA	Make-a-scene	AudioGen	Builder Bot	Make-a-video	ESMFold
Google/ DeepMind	LaMDA	Imagen	MusicLM	DreamFusion	Imagen Video	AlphaFold2
Stability AI	StableLM	Stable Diffusion 2	Dance Diffusion			LibreFold
Amazon	Lex		DeepComposer			
Apple				GAUDI		
NVIDIA	MT-NLG	Edify		Edify	Edify	MegaMolBART
Cohere	Family of LLMs					
Anthropic	Claude					
AI21	Jurassic-2					

Note: List of products are provided for informational purposes only and do not reflect an endorsement from McKinsey & Company.
[1]"Closed source" defined as: model not publicly available, access is typically granted through strict process, and usage may be governed by NDA or other contract.
[2]"Closed source, available through APIs" defined as: source code of model is not available to the public, but the model is often accessible via API, where usage is typically governed by licensing agreements.
[3]"Open source" defined as: code of models available to the public and can be either freely used, distributed, and modified by anyone or restricted for non-commercial use.
[4]OpenAI is backed by significant Microsoft investments.

Model hubs and MLOps

To build applications on top of foundation models, businesses need two things. The first is a place to store and access the foundation model. Second, they may need specialized MLOps tooling, technologies, and practices for adapting a foundation model and deploying it within their end-user applications. This includes, for example, capabilities to incorporate and label additional training data or build the APIs that allow applications to interact with it.

Model hubs provide these services. For closed-source models in which the source code is not made available to the public, the developer of the foundation model typically serves as a model hub. It will offer access to the model via an API through a licensing agreement. Sometimes the provider will also deliver MLOps capabilities so the model can be tuned and deployed in different applications.

For open-source models, which provide code that anyone can freely use and modify, independent model hubs are emerging to offer a spectrum of services. Some may act only as model aggregators, providing AI teams with access to different foundation models, including those customized by other developers. AI teams can then download the models to their servers and fine-tune and deploy them within their application. Others, such as Hugging Face and Amazon Web Services, may provide access to models and end-to-end MLOps capabilities, including the expertise to tune the foundation model with proprietary data and deploy it within their applications. This latter model fills a growing gap for companies eager to leverage generative AI technology but lacking the in-house talent and infrastructure to do so.

Applications

While one foundation model is capable of performing a wide variety of tasks, the applications built on top of it are what enable a specific task to be completed—for example, helping a business's customers with service issues or drafting marketing emails (Exhibit 3). These applications may be developed by a new market entrant seeking to deliver a novel offering, an existing solution provider working to add innovative capabilities to its current offerings, or a business looking to build a competitive advantage in its industry.

There are many ways that application providers can create value. At least in the near term, we see one category of applications offering the greatest potential for value creation. And we expect applications developed for certain industries and functions to provide more value in the early days of generative AI.

Applications built from fine-tuned models stand out

Broadly, we find that generative AI applications fall into one of two categories. The first represents instances in which companies use foundation models largely as is within the applications they build—with some customizations. These could include creating a tailored user interface or adding guidance and a search index for documents that help the models better understand common customer prompts so they can return a high-quality output.

The second category represents the most attractive part of the value chain: applications that leverage fine-tuned foundation models—those that have been fed additional relevant data or had their parameters adjusted—to deliver outputs for a particular use case. While training foundation models requires massive amounts of data, is extremely expensive, and can take months, fine-tuning foundation models requires less data, costs less, and can be completed in days, putting it within reach of many companies.

Application builders may amass this data from in-depth knowledge of an industry or customer needs. For example, consider Harvey, the generative AI application created to answer legal questions. Harvey's developers fed legal data sets into OpenAI's GPT-3 and tested different prompts to enable the tuned model to generate legal documents that were far better than those that the original foundation model could create.

Exhibit 3

There are many applications of generative AI across modalities.

Modality	Application	Example use cases
Text	Content writing	• Marketing: creating personalized emails and posts • Talent: drafting interview questions, job descriptions
	Chatbots or assistants	• Customer service: using chatbots to boost conversion on websites
	Search	• Making more natural web search • Corporate knowledge: enhancing internal search tools
	Analysis and synthesis	• Sales: analyzing customer interactions to extract insights • Risk and legal: summarizing regulatory documents
Code	Code generation	• IT: accelerating application development and quality with automatic code recommendations
	Application prototype and design	• IT: quickly generating user interface designs
	Data set generation	• Generating synthetic data sets to improve AI models quality
Image	Stock image generator	• Marketing and sales: generating unique media
	Image editor	• Marketing and sales: personalizing content quickly
Audio	Text to voice generation	• Trainings: creating educational voiceover
	Sound creation	• Entertainment: making custom sounds without copyright violations
	Audio editing	• Entertainment: editing podcast in post without having to rerecord
3-D or other	3-D object generation	• Video games: writing scenes, characters • Digital representation: creating interior-design mockups and virtual staging for architecture design
	Product design and discovery	• Manufacturing: optimizing material design • Drug discovery: accelerating R&D process
Video	Video creation	• Entertainment: generating short-form videos for TikTok • Training or learning: creating video lessons or corporate presentations using AI avatars
	Video editing	• Entertainment: shortening videos for social media • E-commerce: adding personalization to generic videos • Entertainment: removing background images and background noise in post
	Voice translation and adjustments	• Video dubbing: translating into new languages using AI-generated or original-speaker voices • Live translation: for corporate meetings, video conferencing • Voice cloning: replicating actor voice or changing for studio effect such as aging
	Face swaps and adjustments	• Virtual effects: enabling rapid high-end aging; de-aging; cosmetic, wig, and prosthetic fixes • Lip syncing or “visual” dubbing in post-production: editing footage to achieve release in multiple ratings or languages • Face swapping and deep-fake visual effects • Video conferencing: real-time gaze correction

Note: This list is not exhaustive.

Organizations could also leverage proprietary data from daily business operations. A software developer that has tuned a generative AI chatbot specifically for banks, for instance, might partner with its customers to incorporate data from call-center chats, enabling them to continually elevate the customer experience as their user base grows.

Finally, companies may create proprietary data from feedback loops driven by an end-user rating system, such as a star rating system or a thumbs-up, thumbs-down rating system. OpenAI, for instance, uses the latter approach to continuously train ChatGPT, and OpenAI reports that this helps to improve the underlying model. As customers rank the quality of the output they receive, that information is fed back into the model, giving it more "data" to draw from when creating a new output—which improves its subsequent response. As the outputs improve, more customers are drawn to use the application and provide more feedback, creating a virtuous cycle of improvement that can result in a significant competitive advantage.

In all cases, application developers will need to keep an eye on generative AI advances. The technology is moving at a rapid pace, and tech giants continue to roll out new versions of foundation models with even greater capabilities. OpenAI, for instance, reports that its recently introduced GPT-4 offers "broader general knowledge and problem-solving abilities" for greater accuracy. Developers must be prepared to assess the costs and benefits of leveraging these advances within their application.

Pinpointing the first wave of application impact by function and industry

While generative AI will likely affect most business functions over the longer term, our research suggests that information technology, marketing and sales, customer service, and product development are most ripe for the first wave of applications.

- *Information technology.* Generative AI can help teams write code and documentation. Already, automated coders on the market have improved developer productivity by more than 50 percent, helping to accelerate software development. [10]

- *Marketing and sales.* Teams can use generative AI applications to create content for customer outreach. Within two years, 30 percent of all outbound marketing messages are expected to be developed with the assistance of generative AI systems.[11]

- *Customer service.* Natural-sounding, personalized chatbots and virtual assistants can handle customer inquiries, recommend swift resolution, and guide customers to the information they need. Companies such as Salesforce, Dialpad, and Ada have already announced offerings in this area.

- *Product development.* Companies can use generative AI to rapidly prototype product designs. Life sciences companies, for instance, have already started to explore the use of generative AI to help generate sequences of amino acids and DNA nucleotides to shorten the drug design phase from months to weeks.[12]

In the near term, some industries can leverage these applications to greater effect than others. The media and entertainment industry can become more efficient by using generative AI to produce unique content (for example, localizing movies without the need for hours of human translation) and rapidly develop ideas for new content and visual effects for video games, music, movie story lines, and news articles. Banking, consumer, telecommunications, life sciences, and technology companies are expected to experience outsize operational efficiencies given their considerable investments in IT, customer service, marketing and sales, and product development.

[10] *GitHub Product Blog*, "Research: Quantifying GitHub Copilot's impact on developer productivity and happiness," blog entry by Eirini Kalliamvakou, September 7, 2022.
[11] Jackie Wiles, "Beyond ChatGPT: The future of generative AI for enterprises," Gartner, January 26, 2023.
[12] *NVIDIA Developer Technical Blog*, "Build generative AI pipelines for drug discovery with NVIDIA BioNeMo Service," blog entry by Vanessa Braunstein, March 21, 2023; and Alex Ouyang and Abdul Latif Jameel, "Speeding up drug discovery with diffusion generative models," MIT News, March 31, 2023.

Services

As with AI in general, dedicated generative AI services will certainly emerge to help companies fill capability gaps as they race to build out their experience and navigate the business opportunities and technical complexities. Existing AI service providers are expected to evolve their capabilities to serve the generative AI market. Niche players may also enter the market with specialized knowledge for applying generative AI within a specific function (such as how to apply generative AI to customer service workflows), industry (for instance, guiding pharmaceutical companies on the use of generative AI for drug discovery), or capability (such as how to build effective feedback loops in different contexts).

While generative AI technology and its supporting ecosystem are still evolving, it is already quite clear that applications offer the most significant value-creation opportunities. Those who can harness niche—or, even better, proprietary—data in fine-tuning foundation models for their applications can expect to achieve the greatest differentiation and competitive advantage. The race has already begun, as evidenced by the steady stream of announcements from software providers—both existing and new market entrants—bringing new solutions to market. In the weeks and months ahead, we will further illuminate value-creation prospects in particular industries and functions as well as the impact generative AI could have on the global economy and the future of work.

Tobias Härlin and **Gardar Björnsson Rova** are partners in McKinsey's Stockholm office, where **Oleg Sokolov** is an associate partner; **Alex Singla** is a senior partner in the Chicago office; and **Alex Sukharevsky** is a senior partner in the London office.