McKinsey Digital

QuantumBlack, AI by McKinsey

What every CEO should know about generative AI

Generative AI is evolving at record speed while CEOs are still learning the technology's business value and risks. Here, we offer some of the generative AI essentials.

This article is a collaborative effort by Michael Chui, Roger Roberts, Tanya Rodchenko, Alex Singla, Alex Sukharevsky, Lareina Yee, and Delphine Zurkiya, representing views from the McKinsey Technology Council and QuantumBlack, AI by McKinsey, which are both part of McKinsey Digital.

Image created by Chris Grava / Darby Films using a node-based visual programming language

May 2023

Amid the excitement surrounding generative AI since the release of ChatGPT, Bard, Claude, Midjourney, and other content-creating tools, CEOs are understandably wondering: Is this tech hype, or a game-changing opportunity? And if it is the latter, what is the value to my business?

The public-facing version of ChatGPT reached 100 million users in just two months. It democratized AI in a manner not previously seen while becoming by far the fastest-growing app ever. Its out-of-the-box accessibility makes generative AI different from all AI that came before it. Users don't need a degree in machine learning to interact with or derive value from it; nearly anyone who can ask questions can use it. And, as with other breakthrough technologies such as the personal computer or iPhone, one generative AI platform can give rise to many applications for audiences of any age or education level and in any location with internet access.

All of this is possible because generative AI chatbots are powered by foundation models, which contain expansive neural networks trained on vast quantities of unstructured, unlabeled data in a variety of formats, such as text and audio. Foundation models can be used for a wide range of tasks. In contrast, previous generations of AI models were often "narrow," meaning they could perform just one task, such as predicting customer churn. One foundation model, for example, can create an executive summary for a 20,000-word technical report on quantum computing, draft a go-to-market strategy for a tree-trimming business, and provide five different recipes for the ten ingredients in someone's refrigerator. The downside to such versatility is that, for now, generative AI can sometimes provide less accurate results, placing renewed attention on AI risk management.

With proper guardrails in place, generative AI can not only unlock novel use cases for businesses but also speed up, scale, or otherwise improve existing ones. Imagine a customer sales call, for example. A specially trained AI model could suggest upselling opportunities to a salesperson, but until now those were usually based only on static customer data obtained before the start of the call, such as demographics and purchasing patterns.

A generative AI tool might suggest upselling opportunities to the salesperson in real time based on the actual content of the conversation, drawing from internal customer data, external market trends, and social media influencer data. At the same time, generative AI could offer a first draft of a sales pitch for the salesperson to adapt and personalize.

The preceding example demonstrates the implications of the technology on one job role. But nearly every knowledge worker can likely benefit from teaming up with generative AI. In fact, while generative AI may eventually be used to automate some tasks, much of its value could derive from how software vendors embed the technology into everyday tools (for example, email or word-processing software) used by knowledge workers. Such upgraded tools could substantially increase productivity.

CEOs want to know if they should act now—and, if so, how to start. Some may see an opportunity to leapfrog the competition by reimagining how humans get work done with generative AI applications at their side. Others may want to exercise caution, experimenting with a few use cases and learning more before making any large investments. Companies will also have to assess whether they have the necessary technical expertise, technology and data architecture, operating model, and risk management processes that some of the more transformative implementations of generative AI will require.

The goal of this article is to help CEOs and their teams reflect on the value creation case for generative AI and how to start their journey. First, we offer a generative AI primer to help executives better understand the fast-evolving state of AI and the technical options available. The next section looks at how companies can participate in generative AI through four example cases targeted toward improving organizational effectiveness. These cases reflect what we are seeing among early adopters and shed light on the array of options across the technology, cost, and operating model requirements. Finally, we address the CEO's vital role in positioning an organization for success with generative AI.

Excitement around generative AI is palpable, and C-suite executives rightfully want to move ahead with thoughtful and intentional speed. We hope this article offers business leaders a balanced introduction into the promising world of generative AI.

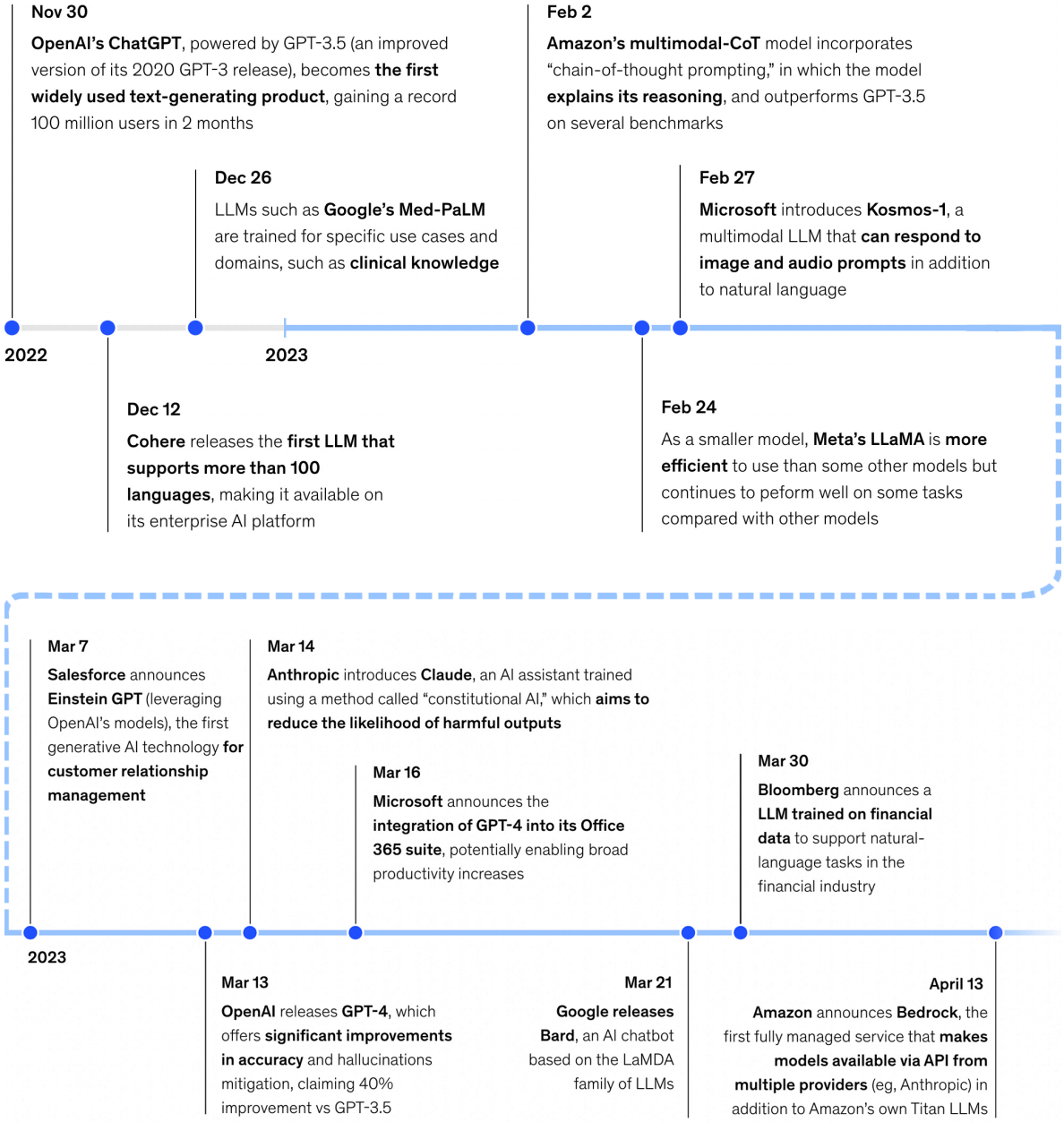
A generative AI primer

Generative AI technology is advancing quickly (Exhibit 1). The release cycle, number of start-ups, and rapid integration into existing software applications are remarkable. In this section, we will discuss the breadth of generative AI

Exhibit 1

Generative AI has been evolving at a rapid pace.

Timeline of some of the major large language model (LLM) developments in the months following ChatGPT's launch



applications and provide a brief explanation of the technology, including how it differs from traditional AI.

More than a chatbot

Generative AI can be used to automate, augment, and accelerate work. For the purposes of this article, we focus on ways generative AI can enhance work rather than on how it can replace the role of humans.

While text-generating chatbots such as ChatGPT have been receiving outsize attention, generative AI can enable capabilities across a broad range of content, including images, video, audio, and computer code. And it can perform several functions in organizations, including classifying, editing, summarizing, answering questions, and drafting new content. Each of these actions has the potential to create value by changing how work gets done at the activity level across business functions and workflows. Following are some examples.

Classify

— A fraud-detection analyst can input transaction descriptions and customer documents into a generative AI tool and ask it to identify fraudulent transactions.

— A customer-care manager can use generative AI to categorize audio files of customer calls based on caller satisfaction levels.

Edit

— A copywriter can use generative AI to correct grammar and convert an article to match a client's brand voice.

— A graphic designer can remove an outdated logo from an image.

Summarize

— A production assistant can create a highlight video based on hours of event footage.

— A business analyst can create a Venn diagram that summarizes key points from an executive's presentation.

Answer questions

— Employees of a manufacturing company can ask a generative AI–based "virtual expert" technical questions about operating procedures.

— A consumer can ask a chatbot questions about how to assemble a new piece of furniture.

Draft

— A software developer can prompt generative AI to create entire lines of code or suggest ways to complete partial lines of existing code.

— A marketing manager can use generative AI to draft various versions of campaign messaging.

As the technology evolves and matures, these kinds of generative AI can be increasingly integrated into enterprise workflows to automate tasks and directly perform specific actions (for example, automatically sending summary notes at the end of meetings). We already see tools emerging in this area.

How generative AI differs from other kinds of AI

As the name suggests, the primary way in which generative AI differs from previous forms of AI or analytics is that it can generate new content efficiently, often in "unstructured" forms (for example, written text or images) that aren't naturally represented in tables with rows and columns (see sidebar "Glossary" for a list of terms associated with generative AI).

The underlying model that enables generative AI to work is called a foundation model. Transformers are key components of foundation models—GPT actually stands for generative pre-trained transformer. A transformer is a type of artificial neural network that is trained using deep learning, a term that alludes to the many (deep) layers within neural networks. Deep learning has powered many of the recent advances in AI.

However, some characteristics set foundation models apart from previous generations of deep

Glossary

Application programming interface (API) is a way to programmatically access (usually external) models, data sets, or other pieces of software.

Artificial intelligence (AI) is the ability of software to perform tasks that traditionally require human intelligence.

Deep learning is a subset of machine learning that uses deep neural networks, which are layers of connected "neurons" whose connections have parameters or weights that can be trained. It is especially effective at learning from unstructured data such as images, text, and audio.

Fine-tuning is the process of adapting a pretrained foundation model to perform better in a specific task. This entails a relatively short period of training on a labeled data set, which is much smaller than the data set the model was initially trained on. This additional training allows the model to learn and adapt to the nuances, terminology, and specific patterns found in the smaller data set.

Foundation models (FM) are deep learning models trained on vast quantities of unstructured, unlabeled data that can be used for a wide range of tasks out of the box or adapted to specific tasks through fine-tuning. Examples of these models are GPT-4, PaLM, DALL·E 2, and Stable Diffusion.

Generative AI is AI that is typically built using foundation models and has capabilities that earlier AI did not have, such as the ability to generate content. Foundation models can also be used for non-generative purposes (for example, classifying user sentiment as negative or positive based on call transcripts) while offering significant improvement over earlier models. For simplicity, when we refer to generative AI in this article, we include all foundation model use cases.

Graphics processing units (GPUs) are computer chips that were originally developed for producing computer graphics (such as for video games) and are also useful for deep learning applications. In contrast, traditional machine learning and other analyses usually run on *central processing units (CPUs)*, normally referred to as a computer's "processor."

Large language models (LLMs) make up a class of foundation models that can process massive amounts of unstructured text and learn the relationships between words or portions of words, known as tokens. This enables LLMs to generate natural language text, performing tasks such as summarization or knowledge extraction. GPT-4 (which underlies ChatGPT) and LaMDA (the model behind Bard) are examples of LLMs.

Machine learning (ML) is a subset of AI in which a model gains capabilities after it is trained on, or shown, many example data points. Machine learning algorithms detect patterns and learn how to make predictions and recommendations by processing data and experiences, rather than by receiving explicit programming instruction. The algorithms also adapt and can become more effective in response to new data and experiences.

MLOps refers to the engineering patterns and practices to scale and sustain AI and ML. It encompasses a set of practices that span the full ML life cycle (data management, development, deployment, and live operations). Many of these practices are now enabled or optimized by supporting software (tools that help to standardize, streamline, or automate tasks).

Prompt engineering refers to the process of designing, refining, and optimizing input prompts to guide a generative AI model toward producing desired (that is, accurate) outputs.

Structured data are tabular data (for example, organized in tables, databases, or spreadsheets) that can be used to train some machine learning models effectively.

Transformers are key components of foundation models. They are artificial neural networks that use special mechanisms called "attention heads" to understand context in sequential data, such as how a word is used in a sentence.

Unstructured data lack a consistent format or structure (for example, text, images, and audio files) and typically require more advanced techniques to extract insights.

learning models. To start, they can be trained on extremely large and varied sets of unstructured data. For example, a type of foundation model called a large language model can be trained on vast amounts of text that is publicly available on the internet and covers many different topics. While other deep learning models can operate on sizable amounts of unstructured data, they are usually trained on a more specific data set. For example, a model might be trained on a specific set of images to enable it to recognize certain objects in photographs.

In fact, other deep learning models often can perform only one such task. They can, for example, either classify objects in a photo or perform another function such as making a prediction. In contrast, one foundation model can perform both of these functions and generate content as well. Foundation models amass these capabilities by learning patterns and relationships from the broad training data they ingest, which, for example, enables them to predict the next word in a sentence. That's how ChatGPT can answer questions about varied topics and how DALL·E 2 and Stable Diffusion can produce images based on a description.

Given the versatility of a foundation model, companies can use the same one to implement multiple business use cases, something rarely achieved using earlier deep learning models. A foundation model that has incorporated information about a company's products could potentially be used both for answering customers' questions and for supporting engineers in developing updated versions of the products. As a result, companies can stand up applications and realize their benefits much faster.

However, because of the way current foundation models work, they aren't naturally suited to all applications. For example, large language models can be prone to "hallucination," or answering questions with plausible but untrue assertions (see sidebar "Using generative AI responsibly"). Additionally, the underlying reasoning or sources for a response are not always provided. This means companies should be careful of integrating generative AI without human oversight in applications where errors can cause harm or where explainability is needed. Generative AI is also currently unsuited for directly analyzing large amounts of tabular data or solving advanced numerical-optimization problems. Researchers are working hard to address these limitations.

The emerging generative AI ecosystem

While foundation models serve as the "brain" of generative AI, an entire value chain is emerging to support the training and use of this technology (Exhibit 2).[1] Specialized hardware provides the extensive compute power needed to train the models. Cloud platforms offer the ability to tap this hardware. MLOps and model hub providers offer the tools, technologies, and practices an organization needs to adapt a foundation model and deploy it within its end-user applications. Many companies are entering the market to offer applications built on top of foundation models that enable them to perform a specific task, such as helping a company's customers with service issues.

The first foundation models required high levels of investment to develop, given the substantial computational resources required to train them and the human effort required to refine them. As a result, they were developed primarily by a few tech giants, start-ups backed by significant investment, and some open-source research collectives (for example, BigScience). However, work is under way on both smaller models that can deliver effective results for some tasks and training that's more efficient. This could eventually open the market to more entrants. Some start-ups have already succeeded in developing their own models—for example, Cohere, Anthropic, and AI21 Labs build and train their own large language models.

Putting generative AI to work

CEOs should consider exploration of generative AI a must, not a maybe. Generative AI can create value in a wide range of use cases. The economics

[1] For more, see "Exploring opportunities in the generative AI value chain," McKinsey, April 26, 2023.

Exhibit 2

A value chain supporting generative AI systems is developing quickly.

and technical requirements to start are not prohibitive, while the downside of inaction could be quickly falling behind competitors. Each CEO should work with the executive team to reflect on where and how to play. Some CEOs may decide that generative AI presents a transformative opportunity for their companies, offering a chance to reimagine everything from research and development to marketing and sales to customer operations. Others may choose to start small and scale later. Once the decision is made, there are technical pathways that AI experts can follow to execute the strategy, depending on the use case.

Much of the use (although not necessarily all of the value) from generative AI in an organization will come from workers employing features embedded in the software they already have. Email systems will provide an option to write the first drafts of messages. Productivity applications will create the first draft of a presentation based on a description. Financial software will generate a prose description of the notable features in a financial report. Customer-relationship-management systems will suggest ways to interact

Using generative AI responsibly

Generative AI poses a variety of risks. CEOs will want to design their teams and processes to mitigate those risks from the start—not only to meet fast-evolving regulatory requirements but also to protect their business and earn consumers' digital trust (we offer recommendations on how to do so later in this article).[1]

Fairness: Models may generate algorithmic bias due to imperfect training data or decisions made by the engineers developing the models.

Intellectual property (IP): Training data and model outputs can generate significant IP risks, including infringing on copyrighted, trademarked, patented, or otherwise legally protected materials. Even when using a provider's generative AI tool, organizations will need to understand what data went into training and how it's used in tool outputs.

Privacy: Privacy concerns could arise if users input information that later ends up in model outputs in a form that makes individuals identifiable. Generative AI could also be used to create and disseminate malicious content such as disinformation, deepfakes, and hate speech.

Security: Generative AI may be used by bad actors to accelerate the sophistication and speed of cyberattacks. It also can be manipulated to provide malicious outputs. For example, through a technique called prompt injection, a third party gives a model new instructions that trick the model into delivering an output unintended by the model producer and end user.

Explainability: Generative AI relies on neural networks with billions of parameters, challenging our ability to explain how any given answer is produced.

Reliability: Models can produce different answers to the same prompts, impeding the user's ability to assess the accuracy and reliability of outputs

Organizational impact: Generative AI may significantly affect the workforce, and the impact on specific groups and local communities could be disproportionately negative.

Social and environmental impact: The development and training of foundation models may lead to detrimental social and environmental consequences, including an increase in carbon emissions (for example, training one large language model can emit about 315 tons of carbon dioxide).[2]

[1] Jim Boehm, Liz Grennan, Alex Singla, and Kate Smaje, "Why digital trust truly matters," McKinsey, September 12, 2022.
[2] Ananya Ganesh, Andrew McCallum, and Emma Strubell, "Energy and policy considerations for deep learning in NLP," Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, June 5, 2019.

with customers. These features could accelerate the productivity of every knowledge worker.

But generative AI can also be more transformative in certain use cases. Following, we look at four examples of how companies in different industries are using generative AI today to reshape how work is done within their organization.[2] The examples range from those requiring minimal resources to resource-intensive undertakings. (For a quick comparison of these examples and more technical detail, see Exhibit 3.)

[2] These examples are amalgamations of cases culled from our client work and public examples rather than reflective of exact events in one particular company.

Exhibit 3

The organizational requirements for generative AI range from low to high, depending on the use case.

Use case example	Technical pathway	Costs	Tech talent	Proprietary data	Process adjustments
Changing the work of software engineering	Use software-as-a-service (SaaS) tool	Many SaaS tools offer fixed-fee subscriptions of $10 to $30 per user per month; some products have usage-based pricing	Little technical talent is needed—potentially for selecting the right solution and light integration work	Because the model is used as is, no proprietary data is needed	Processes largely remain the same, but workers should systematically check model results for accuracy and appropriateness
Helping relationship managers keep up with the pace of public information and data	Build software layers on model API	Up-front investment is needed to develop the user interface, integrate the solution, and build postprocessing layers Running costs for API usage and software maintenance	Software development, product management, and database integration capabilities are needed, which require at least 1 data scientist, machine learning engineer, data engineer, designer, and front-end developer	Because the model is used as is, no proprietary data is needed	Processes may be needed to enable storage of prompts and results, and guardrails may be needed to limit usage for risk or cost reasons
Freeing up customer support representatives' time for higher-value activities	Fine-tune open-source model in-house	Initial costs ~2x more than building on API due to increased human capital costs required for data cleaning and labeling and model fine-tuning Higher running costs for model maintenance and cloud computing	Experienced data science and engineering team with machine learning operations (MLOps) knowledge and resources to check or create labeled data needed	A proprietary, labeled data set is required to fine-tune the model, although in some cases it can be relatively small	Processes for triaging and escalating issues to humans are needed, as well as periodic assessments of model safety
Accelerating the pace at which research scientists can identify relevant cell features for drug discovery	Train a foundation model from scratch	Initial costs ~10–20x more than building on API due to up-front human capital and tech infrastructure costs Running costs for model maintenance and cloud computing similar to the above	Requires large data science and engineering team with PhD-level knowledge of subject matter, best-practice MLOps, and data and infrastructure management skills	Foundation models can be trained on large publicly available data, although long-term differentiation comes from adding owned labeled or unlabeled data (which is easier to collect)	Including the above, when training on external data, thorough legal review is needed to prevent IP issues

Changing the work of software engineering

The first example is a relatively low-complexity case with immediate productivity benefits because it uses an off-the-shelf generative AI solution and doesn't require in-house customization.

The biggest part of a software engineer's job is writing code. It's a labor-intensive process that requires extensive trial and error and research into private and public documentation. At this company, a shortage of skilled software engineers has led to a large backlog of requests for features and bug fixes.

To improve engineers' productivity, the company is implementing an AI-based code-completion product that integrates with the software the engineers use to code. This allows engineers to write code descriptions in natural language, while the AI suggests several variants of code blocks that will satisfy the description. Engineers can select one of the AI's proposals, make needed refinements, and click on it to insert the code.

Our research has shown that such tools can speed up a developer's code generation by as much as 50 percent. It can also help in debugging, which may improve the quality of the developed product. But today, generative AI cannot replace skilled software engineers. In fact, more-experienced engineers appear to reap the greatest productivity benefits from the tools, with inexperienced developers seeing less impressive—and sometimes negative—results. A known risk is that the AI-generated code may contain vulnerabilities or other bugs, so software engineers must be involved to ensure the quality and security of the code (see the final section in this article for ways to mitigate risks).

The cost of this off-the-shelf generative AI coding tool is relatively low, and the time to market is short because the product is available and does not require significant in-house development. Cost varies by software provider, but fixed-fee subscriptions range from $10 to $30 per user per month. When choosing a tool, it's important to discuss licensing and intellectual property issues with the provider to ensure the generated code doesn't result in violations.

Supporting the new tool is a small cross-functional team focused on selecting the software provider and monitoring performance, which should include checking for intellectual property and security issues. Implementation requires only workflow and policy changes. Because the tool is purely off-the-shelf software as a service (SaaS), additional computing and storage costs are minimal or nonexistent.

Helping relationship managers keep up with the pace of public information and data
Companies may decide to build their own generative AI applications, leveraging foundation models (via APIs or open models), instead of using an off-the-shelf tool. This requires a step up in investment from the previous example but facilitates a more customized approach to meet the company's specific context and needs.

In this example, a large corporate bank wants to use generative AI to improve the productivity of relationship managers (RMs). RMs spend considerable time reviewing large documents, such as annual reports and transcripts of earnings calls, to stay informed about a client's situation and priorities. This enables the RM to offer services suited to the client's particular needs.

The bank decided to build a solution that accesses a foundation model through an API. The solution scans documents and can quickly provide synthesized answers to questions posed by RMs. Additional layers around the foundation model are built to streamline the user experience, integrate the tool with company systems, and apply risk and compliance controls. In particular, model outputs must be verified, much as an organization would check the outputs of a junior analyst, because some large language models have been known to hallucinate. RMs are also trained to ask questions in a way that will provide the most accurate answers from the solution (called prompt engineering), and processes are put in place to streamline validation of the tool's outputs and information sources.

In this instance, generative AI can speed up an RM's analysis process (from days to hours), improve job satisfaction, and potentially capture insights the RM might have otherwise overlooked.

The development cost comes mostly from the user interface build and integrations, which require time from a data scientist, a machine learning engineer or data engineer, a designer, and a front-end developer. Ongoing expenses include software maintenance and the cost of using APIs. Costs depend on the model choice and third-party vendor fees, team size, and time to minimum viable product.

Freeing up customer support representatives for higher-value activities

The next level of sophistication is fine-tuning a foundation model. In this example, a company uses a foundation model optimized for conversations and fine-tunes it on its own high-quality customer chats and sector-specific questions and answers. The company operates in a sector with specialized terminology (for example, law, medicine, real estate, and finance). Fast customer service is a competitive differentiator.

This company's customer support representatives handle hundreds of inbound inquiries a day. Response times were sometimes too high, causing user dissatisfaction. The company decided to introduce a generative AI customer-service bot to handle most customer requests. The goal was a swift response in a tone that matched the company brand and customer preferences. Part of the process of fine-tuning and testing the foundation model includes ensuring that responses are aligned with the domain-specific language, brand promise, and tone set for the company; ongoing monitoring is required to verify the performance of the system across multiple dimensions, including customer satisfaction.

The company created a product road map consisting of several waves to minimize potential model errors. In the first wave, the chatbot was piloted internally. Employees were able to give "thumbs up" or "thumbs down" answers to the model's suggestions, and the model was able to learn from these inputs. As a next step, the model "listened" to customer support conversations and offered suggestions. Once the technology was tested sufficiently, the second wave began, and the model was shifted toward customer-facing use cases with a human in the loop. Eventually, when leaders are completely confident in the technology, it can be largely automated.

In this case, generative AI freed up service representatives to focus on higher-value and complex customer inquiries, improved representatives' efficiency and job satisfaction, and increased service standards and customer satisfaction. The bot has access to all internal data on the customer and can "remember" earlier conversations (including phone calls), representing a step change over current customer chatbots.

To capture the benefits, this use case required material investments in software, cloud infrastructure, and tech talent, as well as higher degrees of internal coordination in risk and operations. In general, fine-tuning foundation models costs two to three times as much as building one or more software layers on top of an API. Talent and third-party costs for cloud computing (if fine-tuning a self-hosted model) or for the API (if fine-tuning via a third-party API) account for the increased costs. To implement the solution, the company needed help from DataOps and MLOps experts as well as input from other functions such as product management, design, legal, and customer service specialists.

Accelerating drug discovery

The most complex and customized generative AI use cases emerge when no suitable foundation models are available and the company needs to build one from scratch. This situation may arise in specialized sectors or in working with unique data sets that are significantly different from the data used to train existing foundation models, as this pharmaceutical example demonstrates. Training a foundation model from scratch presents substantial technical, engineering, and resource challenges. The additional return on investment from using a higher-performing model should outweigh the financial and human capital costs.

In this example, research scientists in drug discovery at a pharmaceutical company had to decide which experiments to run next, based on microscopy images. They had a data set of millions of these images, containing a wealth of visual information on cell features that are relevant to drug discovery but difficult for a human to interpret. The images were used to evaluate potential therapeutic candidates.

The company decided to create a tool that would help scientists understand the relationship between drug chemistry and the recorded microscopy outcomes to accelerate R&D efforts. Since such multimodal models are still in infancy, the company decided to train its own instead. To build the model, team members employed both real-world images that are used to train image-based foundational models and their large internal microscopy image data set.

The trained model added value by predicting which drug candidates might lead to favorable outcomes and by improving the ability to accurately identify relevant cell features for drug discovery. This can lead to more efficient and effective drug discovery processes, not only improving time to value but also reducing the number of inaccurate, misleading, or failed analyses.

In general, training a model from scratch costs ten to 20 times more than building software around a model API. Larger teams (including, for example, PhD-level machine learning experts) and higher compute and storage spending account for the differences in cost. The projected cost of training a foundation model varies widely based on the desired model performance level and modeling complexity. Those factors influence the required size of the data set, team composition, and compute resources. In this use case, the engineering team and the ongoing cloud expenses accounted for the majority of costs.

The company found that major updates to its tech infrastructure and processes would be needed, including access to many GPU instances to train the model, tools to distribute the training across many systems, and best-practice MLOps to limit cost and project duration. Also, substantial data-processing work was required for collection, integration (ensuring files of different data sets are in the same format and resolution), and cleaning (filtering low-quality data, removing duplicates, and ensuring distribution is in line with the intended use). Since the foundation model was trained from scratch, rigorous testing of the final model was needed to ensure that output was accurate and safe to use.

Lessons CEOs can take away from these examples

The use cases outlined here offer powerful takeaways for CEOs as they embark on the generative AI journey:

— Transformative use cases that offer practical benefits for jobs and the workplace already exist. Companies across sectors, from pharmaceuticals to banking to retail, are standing up a range of use cases to capture value creation potential. Organizations can start small or large, depending on their aspiration.

— Costs of pursuing generative AI vary widely, depending on the use case and the data required for software, cloud infrastructure, technical expertise, and risk mitigation. Companies must take into account risk issues, regardless of use case, and some will require more resources than others.

— While there is merit to getting started fast, building a basic business case first will help companies better navigate their generative AI journeys.

Considerations for getting started

The CEO has a crucial role to play in catalyzing a company's focus on generative AI. In this closing section, we discuss strategies that CEOs will want to keep in mind as they begin their journey. Many of them echo the responses of senior executives to previous waves of new technology. However, generative AI presents its own challenges, including managing a technology moving at a speed not seen in previous technology transitions.

Organizing for generative AI

Many organizations began exploring the possibilities for traditional AI through siloed experiments. Generative AI requires a more deliberate and coordinated approach given its unique risk considerations and the ability of foundation models to underpin multiple use cases across an organization. For example, a model fine-tuned using proprietary material to reflect the enterprise's brand identity could be deployed across several use cases (for example, generating personalized marketing campaigns and product descriptions) and business functions, such as product development and marketing.

To that end, we recommend convening a cross-functional group of the company's leaders (for example, representing data science, engineering, legal, cybersecurity, marketing, design, and other business functions). Such a group can not only help identify and prioritize the highest-value use cases but also enable coordinated and safe implementation across the organization.

Reimagining end-to-end domains versus focusing on use cases

Generative AI is a powerful tool that can transform how organizations operate, with particular impact in certain business domains within the value chain (for example, marketing for a retailer or operations for a manufacturer). The ease of deploying generative AI can tempt organizations to apply it to sporadic use cases across the business. It is important to have a perspective on the family of use cases by domain that will have the most transformative potential across business functions. Organizations are reimagining the target state enabled by generative AI working in

sync with other traditional AI applications, along with new ways of working that may not have been possible before.

Enabling a fully loaded technology stack

A modern data and tech stack is key to nearly any successful approach to generative AI. CEOs should look to their chief technology officers to determine whether the company has the required technical capabilities in terms of computing resources, data systems, tools, and access to models (open source via model hubs or commercial via APIs).

For example, the lifeblood of generative AI is fluid access to data honed for a specific business context or problem. Companies that have not yet found ways to effectively harmonize and provide ready access to their data will be unable to fine-tune generative AI to unlock more of its potentially transformative uses. Equally important is to design a scalable data architecture that includes data governance and security procedures. Depending on the use case, the existing computing and tooling infrastructure (which can be sourced via a cloud provider or set up in-house) might also need upgrading. A clear data and infrastructure strategy anchored on the business value and competitive advantage derived from generative AI will be critical.

Building a 'lighthouse'

CEOs will want to avoid getting stuck in the planning stages. New models and applications are being developed and released rapidly. GPT-4, for example, was released in March 2023, following the release of ChatGPT (GPT-3.5) in November 2022 and GPT-3 in 2020. In the world of business, time is of the essence, and the fast-paced nature of generative AI technology demands that companies move quickly to take advantage of it. There are a few ways executives can keep moving at a steady clip.

Although generative AI is still in the early days, it's important to showcase internally how it can affect a company's operating model, perhaps through a "lighthouse approach." For example, one way forward is building a "virtual expert" that enables frontline workers to tap proprietary sources of knowledge and offer the most relevant content to customers. This has the potential to increase productivity, create enthusiasm, and enable an organization to test generative AI internally before scaling to customer-facing applications.

As with other waves of technical innovation, there will be proof-of-concept fatigue and many examples of companies stuck in "pilot purgatory." But encouraging a proof of concept is still the best way to quickly test and refine a valuable business case before scaling to adjacent use cases. By focusing on early wins that deliver meaningful results, companies can build momentum and then scale out and up, leveraging the multipurpose nature of generative AI. This approach could enable companies to promote broader AI adoption and create the culture of innovation that is essential to maintaining a competitive edge. As outlined above, the cross-functional leadership team will want to make sure such proofs of concept are deliberate and coordinated.

Balancing risk and value creation

As our four detailed use cases demonstrate, business leaders must balance value creation opportunities with the risks involved in generative AI. According to our recent Global AI Survey, most organizations don't mitigate most of the risks associated with traditional AI, even though more than half of organizations have already adopted the technology.[3] Generative AI brings renewed attention to many of these same risks, such as the potential to perpetuate bias hidden in training data, while presenting new ones, such as its propensity to hallucinate.

As a result, the cross-functional leadership team will want to not only establish overarching ethical principles and guidelines for generative AI use but also develop a thorough understanding of the risks presented by each potential use case.

[3] "The state of AI in 2022—and a half decade in review," McKinsey, December 6, 2022.

It will be important to look for initial use cases that both align with the organization's overall risk tolerance and have structures in place to mitigate consequential risk. For example, a retail organization might prioritize a use case that has slightly lower value but also lower risk—such as creating initial drafts of marketing content and other tasks that keep a human in the loop. At the same time, the company might set aside a higher-value, high-risk use case such as a tool that automatically drafts and sends hyperpersonalized marketing emails. Such risk-forward practices can enable organizations to establish the controls necessary to properly manage generative AI and maintain compliance.

CEOs and their teams will also want to stay current with the latest developments in generative AI regulation, including rules related to consumer data protection and intellectual property rights, to protect the company from liability issues. Countries may take varying approaches to regulation, as they often already do with AI and data. Organizations may need to adapt their working approach to calibrate process management, culture, and talent management in a way that ensures they can handle the rapidly evolving regulatory environment and risks of generative AI at scale..

Applying an ecosystem approach to partnerships

Business leaders should focus on building and maintaining a balanced set of alliances. A company's acquisitions and alliances strategy should continue to concentrate on building an ecosystem of partners tuned to different contexts and addressing what generative AI requires at all levels of the tech stack, while being careful to prevent vendor lock-in.

Partnering with the right companies can help accelerate execution. Organizations do not have to build out all applications or foundation models themselves. Instead, they can partner with generative AI vendors and experts to move more quickly. For instance, they can team up with model providers to customize models for a specific sector, or partner with infrastructure providers that offer support capabilities such as scalable cloud computing.

Companies can use the expertise of others and move quickly to take advantage of the latest generative AI technology. But generative AI models are just the tip of the spear: multiple additional elements are required for value creation.

Focusing on required talent and skills

To effectively apply generative AI for business value, companies need to build their technical capabilities and upskill their current workforce. This requires a concerted effort by leadership to identify the required capabilities based on the company's prioritized use cases, which will likely extend beyond technical roles to include a talent mix across engineering, data, design, risk, product, and other business functions.

As demonstrated in the use cases highlighted above, technical and talent needs vary widely depending on the nature of a given implementation—from using off-the-shelf solutions to building a foundation model from scratch. For example, to build a generative model, a company may need PhD-level machine learning experts; on the other hand, to develop generative AI tools using existing models and SaaS offerings, a data engineer and a software engineer may be sufficient to lead the effort.

In addition to hiring the right talent, companies will want to train and educate their existing workforces. Prompt-based conversational user interfaces can make generative AI applications easy to use. But users still need to optimize their prompts, understand the technology's limitations, and know where and when they can acceptably integrate the application into their workflows. Leadership should provide clear guidelines on the use of generative AI tools and offer ongoing education and training to keep employees apprised of their risks. Fostering a culture of self-driven research and experimentation can also encourage employees to innovate processes and products that effectively incorporate these tools.

Businesses have been pursuing AI ambitions for years, and many have realized new revenue streams, product improvements, and operational efficiencies. Much of the successes in these areas have stemmed from AI technologies that remain the best tool for a particular job, and businesses should continue scaling such efforts. However, generative AI represents another promising leap forward and a world of new possibilities. While the technology's operational and risk scaffolding is still being built, business leaders know they should embark on the generative AI journey. But where and how should they start? The answer will vary from company to company as well as within an organization. Some will start big; others may undertake smaller experiments. The best approach will depend on a company's aspiration and risk appetite. Whatever the ambition, the key is to get under way and learn by doing.

Michael Chui is a partner at the McKinsey Global Institute and a partner in McKinsey's Bay Area office, where **Roger Roberts** is a partner, **Tanya Rodchenko** is an associate partner, and **Lareina Yee**, chair of the McKinsey Technology Council, is a senior partner. **Alex Singla**, a senior partner in the Chicago office, and **Alex Sukharevsky**, a senior partner in the London office, are global leaders of QuantumBlack, AI by McKinsey. **Delphine Zurkiya** is a senior partner in the Boston office.