The CEO's Guide to the Generative AI Revolution

MARCH 07, 2023

By François Candelon, Abhishek Gupta, Lisa Krayer, and Leonid Zhukov

READING TIME: 15 MIN

The release of ChatGPT in late 2022 created a groundswell of interest in generative AI. Within hours, users experimenting with this new technology had discovered and shared myriad productivity hacks. In the weeks and months since, organizations have scrambled to keep pace—and to defend against unforeseen complications. Some organizations have already adopted a more formal approach, creating dedicated teams to explore how generative AI can unlock hidden value and improve efficiency.

For CEOs, however, generative AI poses a much bigger challenge. Today's focus might be on productivity gains and technical limitations, but a revolution in business-model innovation is coming. Much as Mosaic, the world's first free web browser, ushered in the internet era and upended the way we work and live, generative AI has the potential to disrupt nearly every industry—promising both competitive advantage and creative destruction. The implication for leaders is clear: today's breathless activity needs to evolve into a generative AI strategy owned by the C-suite.

This is no small task, and CEOs—who are likely several steps removed from the technology itself—may feel uncertain about their next move. But from our perspective, the priority for CEOs isn't to fully immerse themselves in the technology; instead, they should focus on how generative AI will impact their organizations and their industries, and what strategic choices will enable them to exploit opportunities and manage challenges. These choices are centered on three key pillars:

Potential
Identify the uses cases that will differentiate your organization

People
Adapt organizational structures and prepare employees to support deployment

Policies
Set up ethical guardrails and legal protections

Each pillar raises an urgent question for CEOs. What innovations become possible when every employee has access to the seemingly infinite memory generative AI offers? How will this technology change how employees' roles are defined and how they are managed? How do leaders contend with the fact that generative AI models may produce false or biased output?

Clearly, generative AI is a rapidly evolving space, and each of the pillars above involves short- and long-term considerations—and many other unanswered questions. But CEOs need to prepare for the moment when their current business models become obsolete. Here's how to strategize for that future.

Potential: Discover Your Strategic Advantage

AI has never been so accessible. Tools such as ChatGPT, DALL-E 2, Midjourney, and Stable Diffusion allow anyone to create websites, generate advertising strategies, and produce videos—the possibilities are limitless. This "low-code, no-code" quality will also make it easier for organizations to adopt AI capabilities at scale. (See "The Functional Characteristics of Generative AI.")

THE FUNCTIONAL CHARACTERISTICS OF GENERATIVE AI

The transformative potential of generative AI can be summed up by three key functional characteristics.

Seemingly "Infinite" Memory and Pattern Recognition. Because generative AI is trained on huge amounts of data, its memory can appear infinite. For example, ChatGPT has been trained on a massive portion of publicly available information on the internet. To put this in context, as of 2018 the internet generated 2.5 quintillion bytes of new data daily, according to Domo—the equivalent of 1.2 quintillion words. That number is likely much higher today. Generative AI can also create connections (or recognize patterns) between distant concepts in an almost human-like manner.

Low-Code, No-Code Properties. When describing the impact of ChatGPT, Andrej Karpathy, a founding member of OpenAI, said "the hottest new programming language is English." That's because generative AI's natural-language-processing interface allows nonexperts to create applications with little or no coding required. By contrast, coding assistant systems such as Github Copilot still require competent programmers to operate them.

Lack of a Credible Truth Function. Generative AI's "infinite" memory can become an infinite hallucination. In reality, the level of error in today's generative AI systems is an expected characteristic that makes it useful for generating new ideas and content. But because generative AI does not use logic or intelligent thought, instead predicting the most probable next word based on its training data, it should only be used to generate first drafts of content.

Companies are working to make generative AI's output significantly more reliable by using an approach known as reinforcement learning with human feedback; other approaches that combine generative AI with traditional AI and machine learning have also been considered. Improvements to generative AI are expected soon, with some predicting that it will be able to produce final-draft content by 2030.

The immediate productivity gains can greatly reduce costs. Generative AI can summarize documents in a matter of seconds with impressive accuracy, for example, whereas a researcher might spend hours on the task (at an estimated $30 to $50 per hour).

But generative AI's democratizing power also means, by definition, that a company's competitors will have the same access and capabilities. Many use cases that rely on existing large language model (LLM)[1] applications—such as productivity improvements for programmers who use Github Copilot and for marketing content developers who use Jasper.ai—will be needed just to keep pace with other organizations. But they won't offer differentiation, because the only variability created will result from users' ability to prompt the system.
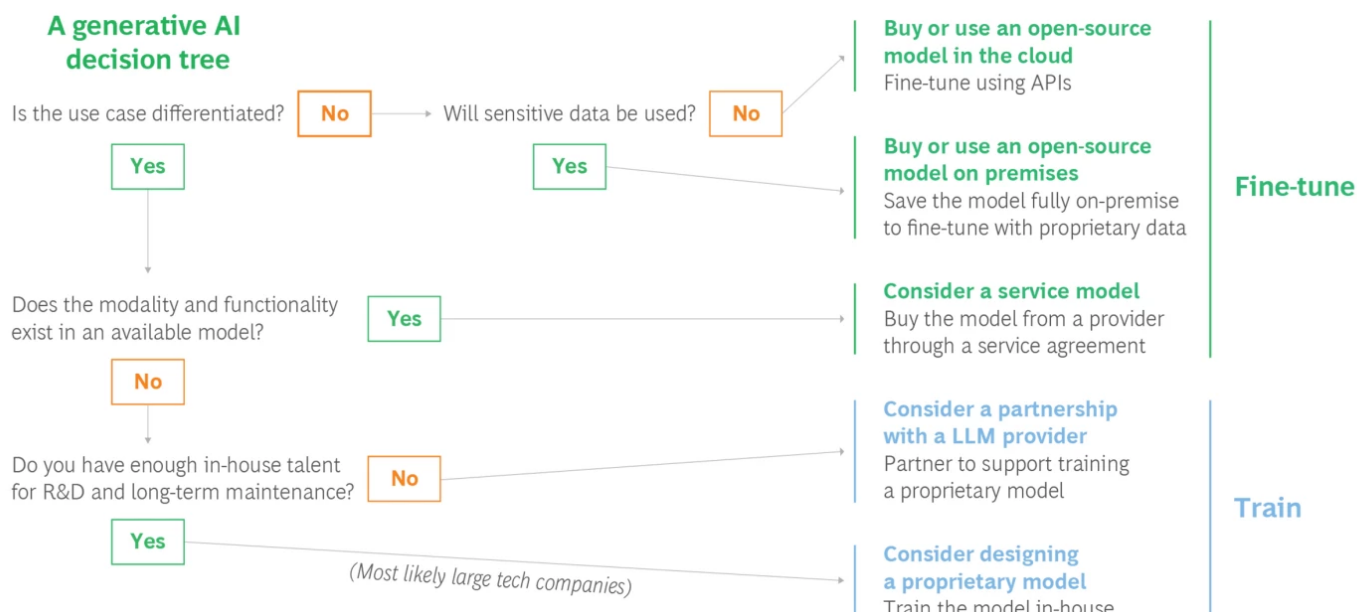
Identify the Right Use Cases

For the CEO, the key is to identify the company's "golden" use cases—those that bring true competitive advantage and create the largest impact relative to existing, best-in-class solutions.

Such use cases can come from any point along the value chain. Some companies will be able to drive growth through improved offerings; Intercom, a provider of customer-service solutions, is running pilots that integrate generative AI into its customer-engagement tool in a move toward automation-first service. Growth can also be found in reduced time-to-market and cost savings—as well as in the ability to stimulate the imagination and create new ideas. In biopharma, for example, much of today's 20-year patent time is consumed by R&D; accelerating this process can significantly increase a patent's value. In February 2021, biotech company Insilico Medicine announced that its AI-generated antifibrotic drug had moved from conceptualization to Phase 1 clinical trials in less than 30 months, for around $2.6 million—several orders of magnitude faster and cheaper than traditional drug discovery.

Once leaders identify their golden use cases, they will need to work with their technology teams to make strategic choices about whether to fine-tune existing LLMs or to train a custom model. (See Exhibit 1.)

Exhibit 1 - Choosing the Right Generative AI Model

A generative AI decision tree

Is the use case differentiated? **No** → Will sensitive data be used? **No** → **Buy or use an open-source model in the cloud** Fine-tune using APIs

Yes ↓

Will sensitive data be used? **Yes** → **Buy or use an open-source model on premises** Save the model fully on-premise to fine-tune with proprietary data

Fine-tune

Does the modality and functionality exist in an available model? **Yes** → **Consider a service model** Buy the model from a provider through a service agreement

No ↓

Do you have enough in-house talent for R&D and long-term maintenance? **No** → **Consider a partnership with a LLM provider** Partner to support training a proprietary model

Train

Yes → *(Most likely large tech companies)* → **Consider designing a proprietary model** Train the model in-house

Source: BCG Henderson Institute.

Note: API = application programming interface; LLM = large language model.

Fine-Tuning an Existing Model. Adapting existing open-source or paid models is cost effective—in a 2022 experiment, Snorkel AI found that it cost between $1,915 and $7,418 to fine-tune a LLM model to complete a complex legal classification. Such an application could save hours of a lawyer's time, which can cost up to $500 per hour.

Fine-tuning can also jumpstart experimentation, whereas using in-house capabilities will siphon off time, talent, and investment. And it will prepare companies for the future, when generative AI is likely to evolve into a model like cloud services: a company purchases the solution with the expectation of achieving quality at scale from the cloud provider's standardization and reliability.

But there are downsides to this approach. Such models are completely dependent on the functionality and domain knowledge of the core model's training data; they are also restricted to available modalities, which today are comprised mostly of language models. And they offer limited options for protecting proprietary data—for example, fine-tuning LLMs that are stored fully on premises.

Training a New or Existing Model. Training a custom LLM will offer greater flexibility, but it comes with high costs and capability requirements: an estimated $1.6 million to train a 1.5-billion-parameter model with two configurations and 10 runs per configuration, according to AI21 Labs. To put this investment in context, AI21 Labs estimated that Google spent approximately $10 million for training BERT and OpenAI spent $12 million on a single training run for GPT-3.[2] (Note that it takes multiple rounds of training for a successful LLM.)

These costs—as well as data center, computing, and talent requirements—are significantly higher than those associated with other AI models, even when managed through a partnership. The bar to justify this investment is high, but for a truly differentiated use case, the value generated from the model could offset the cost.

Plan Your Investment

Leaders will need to carefully assess the timing of such an investment, weighing the potential costs of moving too soon on a complex project for which the talent and technology aren't yet ready against the risks of falling behind. Today's generative AI is still limited by its propensity for error and should primarily be implemented for use cases with a high tolerance for variability. CEOs will also need to consider new funding mechanisms for data and infrastructure—whether, for example, the budget should come from IT, R&D, or another source—if they determine that custom development is a critical and time-sensitive need.

The "fine-tune versus train" debate has other implications when it comes to long-term competitive advantage. Previously, most research on generative AI was public and models were provided through open-source channels. Because this research is now moving behind closed doors, open-source models are already falling far behind state-of-the-art solutions. In other words, we're on the brink of a generative AI arms race. (See "The Future State of the LLM Market.")

THE FUTURE OF THE LLM MARKET

Until recently, most generative AI research has been publicly accessible. But many companies are choosing to stop or delay publishing their research findings and are keeping model architectures as proprietary knowledge. (For example, GPT-2 was open-source but GPT-3 is proprietary.)

The next improvements to generative models with vast number of users will likely come from logs of their user interaction, giving these models a significant competitive advantage over new entrants. This reality, combined with the heavy data, infrastructure, and talent costs required to train LLMs, means that the LLM market has both economy and quality of scale. Advances in generative AI therefore might be limited to large companies, while the democratization of AI development for small and medium-sized enterprises could be limited to nondifferentiated use cases.

The jury is still out, but this dynamic appears comparable to the "search-engine wars." Several large companies invested heavily in search solutions, but Google's user-

friendliness and accuracy helped set it apart from competitors. Once users preferred Google, other engines could not keep up—because every search request Google received made it better and smarter. Soon, all other B2C solutions faded away. A similar winner-take-all situation could play out in the LLM market, with the big, early entrants eventually owning the models and having full control over access.

> "
>
> **A winner-take-all situation could play out in the LLM market.**

It's worth noting, however, that Google did not achieve the same level of success in the enterprise search market, which has unique requirements and challenges compared to B2C. At the enterprise level, search-engines lack the scale to build domain expertise and lack the volume of user data to build that capability. Similarly, businesses will get the most value out of LLMs that are trained on their proprietary data and that have modalities that drive unique use cases. This could make it difficult for any single player to dominate the B2B market.

There is also the potential for companies and governments to fund open-source models to keep them state of the art—similar to how IBM funded Linux.

These market dynamics have key implications for CEOs as they make customization and implementation decisions:

- It is unlikely that any single LLM provider will dominate the B2B market; the key for companies is to find large models with the modality and functionality that match their golden use cases or use cases that require sensitive data.

- While training LLMs is an option for large businesses, the quality of scale could make purchasing solutions more reliable (similar to cloud).

- If choosing to train in-house, be wary of relying too much on individual researchers. If only a small number of people have the expertise to advance and maintain the model, this will cause a single point of failure if those researchers choose to leave.

As research accelerates and becomes more and more proprietary, and as the algorithms become increasingly complex, it will be challenging to keep up with state-of-the-art models. Data scientists will need special training, advanced skills, and deep expertise to understand how the models work—their capabilities, limitations, and utility for new business use cases. Large players that want to remain independent while using the latest AI technology will need to build strong internal tech teams.

People: Prepare Your Workforce

Like existing forms of artificial intelligence, generative AI is a disruptive force for humans. In the near term, CEOs need to work with their leadership teams as well as HR leaders to determine how this transformation should unfold within their organizations—redefining employees' roles and responsibilities and adjusting operating models accordingly.

Redefine Roles and Responsibilities

Some AI-related shifts have already occurred. Traditional AI and machine-learning algorithms (sometimes incorrectly referred to as analytical AI), which use powerful logic or statistics to analyze data and automate or augment decision making, have enabled people to work more autonomously and managers to increasingly focus on team dynamics and goal setting.

Now generative AI, in its capacity as a first-draft content generator, will augment many roles by increasing productivity, performance, and creativity. Employees in more clerical roles, such as paralegals and marketers, can use generative AI to create first drafts, allowing them to spend more of their time refining content and identifying new solutions. Coders will be able to focus on activities such as improving code quality on tight timelines and ensuring compliance with security requirements.

Of course, these changes cannot (and should not) happen in a vacuum. CEOs need to be aware of the effect that AI has on employees' emotional well-being and professional identity. Productivity improvements are often conflated with reduction in overall staff, and AI has already stoked concern among employees; many college graduates believe AI will make their job irrelevant in a few years. But it's also possible that AI will create as many jobs as it will displace.

The impact of AI is thus a critical culture and workforce issue, and CEOs should work with HR to understand how roles will evolve. As AI initiatives roll out, regular pulse checks should be conducted to

track employee sentiment; CEOs will also need to develop a transparent change-management initiative that will both help employees embrace their new AI coworkers and ensure employees retain autonomy. The message should be that humans aren't going anywhere—and in fact are needed to deploy AI effectively and ethically. (See Exhibit 2.)

Exhibit 2 - Key Questions to Shape Your Adoption of Generative AI

Organization and Operating Model	Talent and Skills	Culture and Change
What short-term actions will manage the cadence of initiatives across the organization?	What new skills and talent will help achieve long-term advantage?	What training will ensure that my workforce uses generative AI appropriately and ethically?
How will existing roles and responsibilities change because of generative AI?	Should I upskill, reskill, or resize the workforce?	How can we adapt professional identity to encourage AI adoption?
How should I organize my departments for efficient collaboration with AI?	What new competencies will managers need to lead an AI-augmented workforce?	How can we encourage a culture of where humans and AI collaborate as colleagues?
Where should the large language models and data scientists sit within the organization?	How can we adapt recruiting and training to ensure a strong workforce in 10 years?	How can management communication create positive momentum?

Source: BCG Henderson Institute.

As AI adoption accelerates, CEOs need to learn as they go and use those lessons to develop a strategic workforce plan—in fact, they should start creating this plan now and adapt it as the technology evolves. This is about more than determining how certain job descriptions will change—it's about ensuring that the company has the right people and management in place to stay competitive and make the most out of their AI investments. Among the questions CEOs should ask as they assess their company's strengths, weaknesses, and priorities are:

- What competencies will project leaders need to ensure that individual contributors' work is of sufficient quality?

- How can CEOs create the optimal experience curve to produce the right future talent pipeline— ensuring, for example, that employees at a more junior level are upskilled in AI augmentation and that supervisors are prepared to lead an AI-augmented workforce?

- How should training and recruiting be adjusted to build a high-performing workforce now and in the future?

Adjust Your Operating Model

We expect that agile (or bionic) models will remain the most effective and scalable in the long term, but with centralized IT and R&D departments staffed with experts who can train and customize LLMs. This centralization should ensure that employees who work with similar types of data have access to the same data sets. When data is siloed within individual departments—an all-too-common occurrence—companies will struggle to realize generative AI's true potential. But under the right conditions, generative AI has the power to eliminate the compromise between agility and scale.

Because of the increased importance of data science and engineering, many companies will benefit from having a senior executive role (for example, a chief AI officer) oversee the business and technical requirements for AI initiatives. This executive should place small data-science or engineering teams within each business unit to adapt models for specific tasks or applications. Technical teams will thus have the domain expertise and direct contact to support individual contributors, ideally limiting the distance between the platform or tech leaders and individual contributors to one layer.

Structurally, this could involve department-focused teams with cross-functional members (for example, sales teams with sales reps and dedicated technical support) or, preferably, cross-departmental and cross-functional teams aligned to the business and technical platforms.

Policies: Protect Your Business

Generative AI lacks a credible truth function, meaning that it doesn't know when information is factually incorrect. The implications of this characteristic, also referred to as "hallucination," can range from humorous foibles to damaging or dangerous errors. But generative AI also presents other critical risks for companies, including copyright infringement; leaks of proprietary data; and unplanned functionality that is discovered after a product release, also known as capability overhang. (See Exhibit 3.) For example, Riffusion used a text-to-image model, Stable Diffusion, to create new music by converting music data into spectrograms.

Exhibit 3 - Generative AI Poses a Range of Risks

Challenging

Ease of mitigation

- Energy use and environmental harm
- Capability overhang
- Biased outputs
- Copyright infringement
- Lack of a truth function
- Sophisticated phishing and fraud
- Leaks of proprietary data

Simple

Low risk   **Severity of business risk**   High risk

Source: BCG Henderson Institute.

Prepare for Risk

Companies need policies that help employees use generative AI safely and that limit its use to cases for which its performance is within well-established guardrails. Experimentation should be encouraged; however, it is important to track all experiments across the organization and avoid "shadow experiments" that risk exposing sensitive information. These policies should also guarantee clear data ownership, establish review processes to prevent incorrect or harmful content from being published, and protect the proprietary data of the company and its clients.

Another near-term imperative is to train employees how to use generative AI within the scope of their expertise. Generative AI's low-code, no-code properties may make employees feel overconfident in their ability to complete a task for which they lack the requisite background or skills; marketing staff, for example, might be tempted to bypass corporate IT rules and write code to build a new marketing tool. About 40% of code generated by AI is insecure, according to NYU's Center for Cybersecurity—and because most employees are not qualified to assess code vulnerabilities, this creates a significant security risk. AI assistance in writing code also creates a quality risk, according to a Stanford University study, because coders can become overconfident in AI's ability to avoid vulnerabilities.

Leaders therefore need to encourage all employees, especially coders, to retain a healthy skepticism of AI-generated content. Company policy should dictate that employees only use data they fully understand and that all content generated by AI is thoroughly reviewed by data owners. Generative AI applications (such as Bing Chat) have already started implementing the ability to reference source data, and this function can be expanded to identify data owners.

Ensure Quality and Security

Leaders can adapt existing recommendations regarding responsible publication to guide releases of generative AI content and code. They should mandate robust documentation and set up an institutional review board to review *a priori* considerations of impact, akin to the processes for publishing scientific research. Licensing for downstream uses, such as the Responsible AI License (RAIL), presents another mechanism for managing the generative AI's lack of a truth function.

Finally, leaders should caution employees against using public chatbots for sensitive information. All information typed into generative AI tools will be stored and used to continue training the model; even Microsoft, which has made significant investments in generative AI, has warned its employees not to share sensitive data with ChatGPT.

Today, companies have few ways to leverage LLMs without disclosing data. One option for data privacy is to store the full model on premises or on a dedicated server. (BLOOM, an open-source model from Hugging Face's BigScience group, is the size of GPT-3 but only requires roughly 512 gigabytes of storage.) This may limit the ability to use state-of-the-art solutions, however. Beyond sharing proprietary data, there are other data concerns when using LLMs, including protecting personally identifiable information. Leaders should consider leveraging cleaning techniques such as named entity recognition to remove person, place, and organization names. As LLMs mature, solutions to protect sensitive information will also gain sophistication—and CEOs should regularly update their security protocols and policies.

Generative AI presents unprecedented opportunities. But it also forces CEOs to grapple with towering unknowns, and to do so in a space that may feel unfamiliar or uncomfortable. Crafting an effective strategic approach to generative AI can help distinguish the signal from the noise. Leaders who are prepared to reimagine their business models—identifying the right opportunities, organizing their workforce and operating models to support generative AI innovation, and ensuring that experimentation doesn't come at the expense of security and ethics—can create long-term competitive advantage.

ABOUT BOSTON CONSULTING GROUP

Boston Consulting Group partners with leaders in business and society to tackle their most important challenges and capture their greatest opportunities. BCG was the pioneer in business strategy when it was founded in 1963. Today, we work closely with clients to embrace a transformational approach aimed at benefiting all stakeholders—empowering organizations to grow, build sustainable competitive advantage, and drive positive societal impact.

Our diverse, global teams bring deep industry and functional expertise and a range of perspectives that

question the status quo and spark change. BCG delivers solutions through leading-edge management consulting, technology and design, and corporate and digital ventures. We work in a uniquely collaborative model across the firm and throughout all levels of the client organization, fueled by the goal of helping our clients thrive and enabling them to make the world a better place.

For information or permission to reprint, please contact BCG at permissions@bcg.com. To find the latest BCG content and register to receive e-alerts on this topic or others, please visit bcg.com. Follow Boston Consulting Group on Facebook and Twitter.

1   Large language models, also known as foundation models, are deep- learning algorithms that can recognize, summarize, translate, predict, and generate content based on its training data. Today these models are mostly trained on text, images, and audio, but they can also go beyond language and images into signals, biological data, and more. Models trained on data beyond language are called multimodal models.
2   "How Generative AI Is Changing Creative Work," Harvard Business Review, November 14, 2022.

Authors

François Candelon

MANAGING DIRECTOR & SENIOR PARTNER; GLOBAL DIRECTOR, BCG HENDERSON INSTITUTE

Paris

Abhishek Gupta

SENIOR SOLUTION DELIVERY MANAGER, RESPONSIBLE AI

Montreal

Lisa Krayer

PROJECT LEADER

Washington, DC

Leonid Zhukov

VICE PRESIDENT - DATA SCIENCE

New York