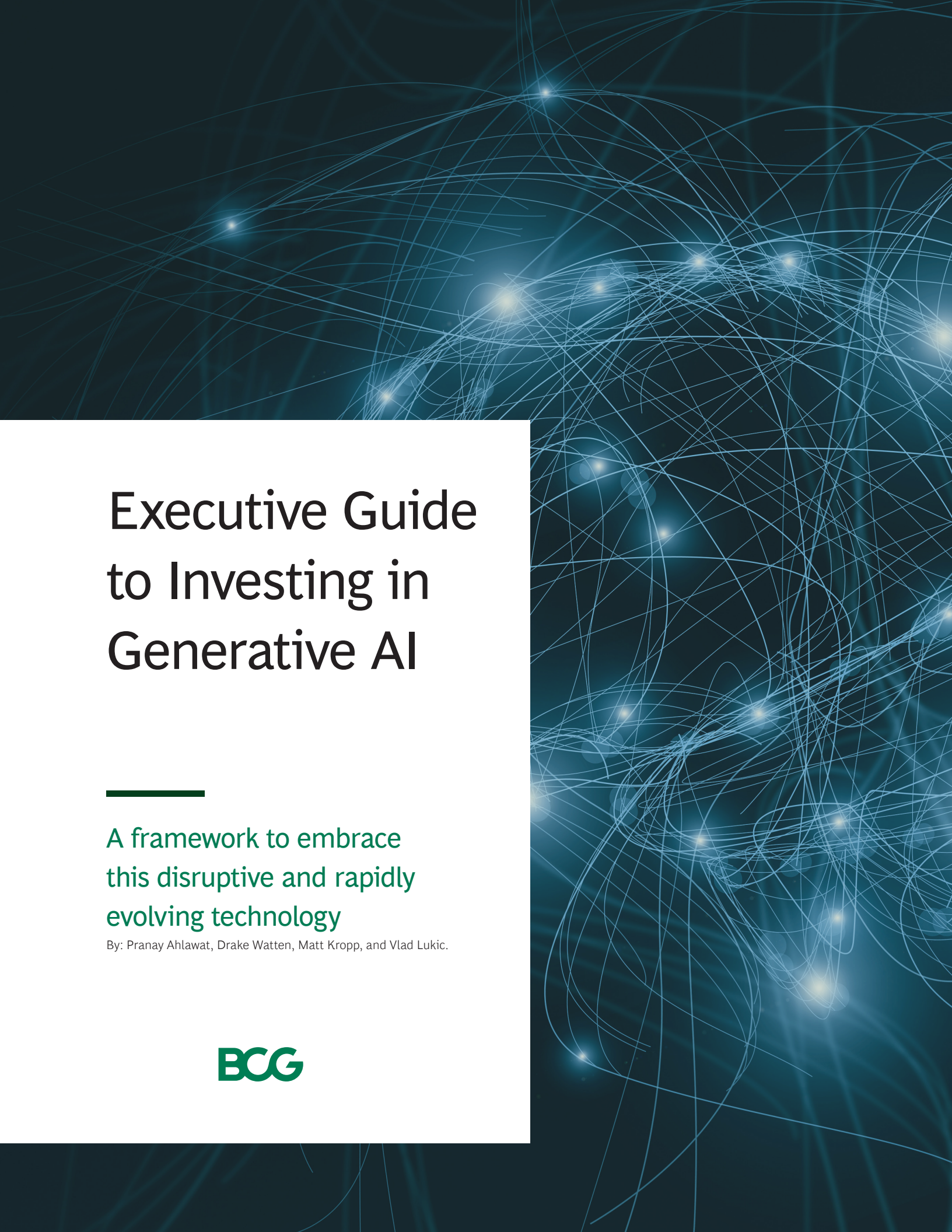

Executive Guide to Investing in Generative AI

A framework to embrace this disruptive and rapidly evolving technology

By: Pranay Ahlawat, Drake Watten, Matt Kropp, and Vlad Lukic.

BCG

Content

Introduction

Generative AI's potential to disrupt industries, revolutionize customer relationships, and change the way knowledge work gets done has made it a strategic imperative across companies. The question is "how" to embrace this technology in a way that is right for your organization and can profitably scale. Based on our research and interviews with executives, there are four key obstacles that executives must navigate today.

Fragmented approach to generative AI without a clear business case

Limited understanding of the business value and current feasibility of use cases, as well as unclear prioritization of use cases

Inadequate understanding of generative AI technology trends and implications

Not knowing what technology approach to take and misunderstanding its risks

Not being clear on when to "build" versus "buy"

Prioritizing the wrong use cases to build, when buying an out-of-the-box solution might be a better fit

Not fully understanding the at-scale economics of generative AI

Not factoring the second-order costs of adopting generative AI at scale and building business cases

This article delves into four pivotal questions that executives should ask to better manage these challenges and succeed in their generative AI journeys.

1 *Why do I want to use Generative AI and what business value will it create?*

This is a simple question, but one that gets overlooked in the rush of companies to invest in generative AI. In our research, many companies have started multiple initiatives without fully understanding the entire spectrum of generative AI use cases, sequencing or prioritizing them, and estimating their business impact. We observe many large companies taking a fragmented and uncoordinated approach to generative AI, where different Business Units (BUs) or Lines of Business (LOBs) are driving use cases in the absence of an enterprise-wide strategy. Furthermore, many companies are still discovering the capabilities of current platforms and tools and are failing to understand their limitations or challenges, partly due to the noisy hype surrounding generative AI today.

To cut through the noise, companies must keep three things in mind.

- ***Not all use cases are created equal.*** Organizations need to start with a clear strategy based on the business value of the use case, as opposed to taking a scattered approach across multiple pilots. Some companies are at risk of being disrupted and need to double down on generative AI to build new offerings and value propositions. In other cases, more horizontal use cases may be less strategic, and companies should thoughtfully prioritize these considering taking into consideration levels of investment, ROI, and risk, amongst other things.

- ***It's early days and Generative AI's capabilities are still evolving.*** Despite its enormous potential, generative AI has technical limitations and is still maturing for all use cases, particularly for modalities outside text (such as video), and more industry-focused ones (for example, healthcare, manufacturing). In our research, many customer pilots are running into day-two operational challenges (for example, cybersecurity, machine learning operations, governance), and some are unable to get past pilots and user acceptance testing phases because of results that do not meet the bar. To fully understand the benefits and challenges, organizations should experiment and run pilots within their context, and with their data securely to ensure the outcomes are salient for the objectives they have set.

- ***There are risks and associated challenges with generative AI which may drive second-order costs.*** These risks are well understood and include data hallucinations, bias, cybersecurity, and copyright challenges, amongst others. Many of the companies we interviewed are only starting to grasp these challenges as they scale.

To make thoughtful Generative AI investments that create the most impact, it is important to think through your strategy and technological approach, understand the alternatives, and be equipped to pivot quickly.

2 *What type of Generative AI should I be using and is that a future-proof decision?*

Generative AI is not a one-size-fits-all technology, as it can be deployed in four different ways. The options, which come with their unique tradeoffs and costs, range from using public APIs that are turnkey and available for immediate usage, to building and training a custom model from scratch. What companies choose ought to be driven by six strategic considerations.

Speed to market

Companies that are seeking feature parity with competitors or who operate in industries that are getting disrupted (for example, traditional chatbot platforms or enterprise search vendors) have a much greater need to rapidly respond.

Customization requirements

While out-of-box functionalities from commercial model vendors such as OpenAI, Anthropic, Amazon Titan, etc., might work for multiple use cases, more complex or domain-specialized use cases might benefit from fine-tuning commercial or open models.

Volume and scale of intended use cases

The overall volume and nature of workloads—whether it is spiky and seasonal or consistently high volume—need to be considered, as they impact the mode of deployment.

Data privacy and regulation needs

Customers in industries such as healthcare or financial services will need to evaluate the data residency or privacy controls requirements and decide if those could be met by compliant cloud solutions, or whether they warrant building custom models of a generative AI architecture built for on-premises on-perm.

Latency and performance requirements

Speed of response is a key consideration for certain real-time use cases, such as live summarization, automated trading, etc. Out-of-box capabilities from commercial model vendors might not be sufficient. The ability to optimize the size, and architecture of the foundation model for latency and speed of response may be critical considerations for such use cases.

Operating model implications

Availability of talent is a key factor in determining how a company implements generative AI. Building a custom model is not a viable option for many enterprises due to the steep requirements for data science and machine learning talent. Additionally, the maturity of a company's broader operating model, such as its data pipeline management as well as machine learning operational processes is another key factor.

Based on the above considerations, customers can choose one of four generative AI deployment options.

- Use **public APIs** (for example, Google Translate, AI21 Summarize, Amazon, etc., or AI21 Paraphrase) for simple, and standard tasks that do not require any customization.
- Consume **Models-as-a-Service** from model vendors such as OpenAI, AWS Titan, and Anthropic. Most customers leverage these models with no tuning or customization via techniques like RAG (Retrieval-Augmented Generation).
- Use **open models** such as Falcon or Sable Diffusion which can be used as-is or customized for domain-specific use cases.
- Create and train a **custom model** from scratch (such as Bloomberg GPT) for specialized tasks and maximum performance control.

Generative AI deployment options

	Public APIs and Services	Models-as-a-Service (MaaS)	Open Models	Proprietary Models
Description	Leverage public services e.g., Google Translate, Amazon Polly, Azure Cognitive Services	Build Gen AI applications on managed models e.g., Cohere, Anthropic, OpenAI	Tune or use publicly available or open-source models as is e.g., Falcon, Stable Diffusion	Build custom models from scratch e.g., Bloomberg GPT, Palm etc.
Ability to customize	No ability to customize	Limited – some vendors provide ability to tune models	High degree of customization – can be used as-is, tuned or fully re-trained	Maximum ability to optimize every aspect of stack
Deployment modality	Cloud multi-tenant	Cloud with option to deploy instance in private VPC	Hybrid - cloud or on-premise	Hybrid - cloud or on-premise
Setup Cost[1]	<$0.1M	<$1M (Multitenant cloud) to $7M (Private VPC[2])	$3M (Private VPC[2]) to $30M (On premise)	$10M (Private VPC) to $50M+ (On premise)
Run (1-year)[1]	<$0.1M	<$1M to $6M (Varies by usage and deployment)	$5M (Private VPC[2]) to $1M (On premise)	$7M (Private VPC[2]) to $1M (On premise)
Use-cases	Simple standard tasks, like language translation	Use-cases requiring contextual learning with limited to no customization, e.g., personalized marketing campaigns	Specialized use-cases requiring domain specific or organization specific customizations, e.g., drug discovery acceleration	Use-cases requiring full control over the model for data privacy, customization, latency and performance optimizations, e.g., defense industry use cases

1: Assumed GPT-3 equivalent foundation model with 175B parameters and 300B training tokens. Inferencing assumed for 10B tokens and fine tuning for 20B tokens. Training, tuning and inference costs calculated as API price per token * number of tokens or (FLOPs per token for GPT-3 / FLOPs per second for each machine) * pricing per second for each machine. Total costs estimated based on BCG's 70:20:10 framework – 70% effort in AI implementations spent on talent & org change, 20% on technology and 10% on algorithms. 2. Virtual Private Cloud

These decisions come with their own tradeoffs. Models-as-a-Service can speed up time to market, but it can get expensive at scale and lead to vendor lock-in. On the other hand, leveraging open-source models or building a model from scratch comes with high up-front costs and demanding talent needs.

What complicates this decision even more is the fact that the technology and economics of generative AI are rapidly evolving. To make a future-proof decision, it is imperative for customers to understand where the proverbial "puck" is going and the technology trends influencing that direction. Our research found four key trends:

Bigger is not necessarily better: it is important to find the right model at the right cost/performance

The capabilities of foundation models are quickly converging. Smaller, well-trained models are now delivering comparable performance to larger models at a fraction of the cost, challenging the notion that bigger is always better. For example, Chinchilla, a 70-billion parameter model, delivers the same performance as Gopher, a 280-billion parameter model, but at a 75% lower cost.

Model customization and tuning continue to get easier

Several optimization techniques, such as Low-Rank Adaptation (LoRA), pruning and quantization, are lowering the costs of training, tuning, and inferencing models. This is lowering the technology barriers and time to market to build custom models, making these architectures more viable.

Custom silicon and accelerators are set to create a step change in performance/costs

Innovations in accelerators and silicon are improving price/performance ratios. Custom silicon from Amazon (AWS Trainium and AWS Inferentia) are delivering with up to 30-50% in price performance, and driving more integrated stacks for machine learning.

Overall generative AI stack is maturing and driving democratization

Tools and open-source libraries such as LangChain and HuggingFace Transformers are reducing the barriers to entry to building complex generative AI applications.

As companies evaluate different deployment options, it is crucial to assess the technical, financial, and organizational impacts of each option. They must think of vendor lock-in risks, as well as technology risks of making wrong platform bets. In our research, we found that organizations are experimenting with multiple options at the same time, depending on use cases. They are using Models-as-a-Service (MaaS) to start, focusing on lower-risk use cases (for example, summarizations, knowledge management), while also dabbling in open-source or custom models for specialized use cases like anti-money laundering and fraud detection. The value and risk of generative AI must be understood within an organization's context.

3 *Do I really need to build this capability, or should I wait and buy?*

Beyond choosing the right technology, companies need to ask if building a solution is warranted at all. Given the rapidly changing technology landscape and where we are in the generative AI adoption cycle, it is important for organizations to be realistic about what can be achieved, and be clear about what use cases to build and which ones to buy.

In our experience, there are multiple types of scenarios where investing aggressively and building solutions early is advisable. The first scenario is one in which generative AI incrementally enhances the core offering or business model (for example, chat-based e-commerce, built-in generative AI tools in creative software such as Adobe Firefly, and AI-enabled workflow tools for enterprise software such as Salesforce Einstein-GPT). A second scenario is when generative AI creates a novel offering or opens new markets, such as drug discovery in biopharma. Yet another scenario might be if your industry is facing disruption (for example, information services, legacy NLP (Natural Language Processing), and chatbot platforms), moving rapidly and adopting generative AI might be the only option.

On the other end of the spectrum, buying out-of-the-box solutions might be a better option for more standardized use cases where solutions already exist (for example, coding or writing assistants), or horizontal use cases where platform solutions will likely emerge (for example, customer support). Waiting, continuing to evaluate, and keeping one's options open can be a good strategic decision as it increases flexibility, keeps a company focused, and reduces the technological risk of making wrong investments.

There are other vertical and efficiency-driven use cases (for example, network automation in telcos), where companies need to consider their business value, assess their current ability to execute (such as lack of specialized machine learning talent), and estimate costs to build. While prioritizing these use cases can yield significant value, it remains crucial for companies to continuously assess the rapidly evolving tech landscape and explore commercial solutions that not only expedite the advancement of their use case but also effectively manage risks and minimize investments.

Regardless of approach, as a general principle, companies should have a very high bar for building proprietary foundation models, as this will only make economic sense at a very large scale. This option is more suited for hyperscalers, given extremely specialized talent and large upfront capital are required to develop and train state-of-the-art foundation models.

4 What is the Total Cost of Ownership (TCO) equation and how will it evolve over time?

There are two types of costs of generative AI: primary and secondary. Our research suggests that most companies consistently underestimate the latter.

Primary costs, which are better understood, include fixed set-up costs (for example, hardware, data curation, costs of engineering, training, and tuning) and variable costs (such as consumption costs for APIs, and inference).

The secondary costs of generative AI include a lot of hidden, and hard-to-estimate costs, ranging from maintenance costs (for example, repaying technical debt, re-training, incremental testing), risk management, organizational change management, as well as legal costs.

The total costs can vary significantly between deployment options (as illustrated in the exhibit above). Public API options have the lowest setup and run costs (<$0.1M setup costs for API integration). Maas (Models as a Service) options are generally used without customization and typically have lower setup costs (<$1M) including setting up data pipelines and vector databases, but the run costs at scale can get expensive, reaching up to $6M in some cases. Building proprietary models can be prohibitively expensive—sometimes costing upwards of $50M depending on the type of model, data curation, and architectural investments—but they provide the most architectural control over run costs.

Like any large-scale digital transformation, the biggest overlooked secondary cost is organizational change management. BCG's 10:20:70 framework explains the relative investments needed to implement AI at scale: 10% on algorithms, 20% on technology, and 70% on organizational change. The last cost of organizational change is not only the biggest investment, but also the hardest to estimate and fraught with the risks of internal change.

Compliance and legal costs are also increasing due to the new risks introduced by generative AI. In our research, some organizations are experiencing a 25% increase in legal, testing, and compliance costs for their generative AI use cases. We do expect these costs to improve over time, as companies go up the experience and adoption curve.

Understanding the TCO for generative AI requires modeling different scenarios that weigh expected benefits and both primary and secondary costs—as well as how each cost may evolve. Moreover, the evolving technological innovations underscore the need to be flexible in decision-making. Executives may be better off making decisions with a medium-term horizon and being prepared to pivot quickly.

Putting the four questions together—how to win with generative AI

There are four strategic control points you need to consider to be on the winning side of generative AI:

Focus on economic viability at scale when making core technology choices

Building and running generative AI applications can be very expensive. Customers must evaluate the cost-efficiency and performance expectations of different model and deployment options vis-a-vis the use case they are solving. Factors such as data privacy, customization requirements, latency, and volume must be considered to make the right technology bet at the optimal price/performance ratio. Moreover, once companies understand their priority use cases, they have to invest in scale to avoid the cycles of repeated demos. Generative AI can only be strategic if it is done at the right cost at scale.

Invest early in closing the talent gaps

Talent remains a top-of-mind concern and a major stumbling block for most companies that are implementing generative AI. Companies must invest in up-skilling their current workforce and hiring to close the talent gap over time. Consider other options like acquisitions or partnerships to bridge the gap in the near term.

Look ahead to solve the data and day-two operational equation

No technology can scale unless you solve day-two operational challenges. In our research, customers are starting to run into operational challenges across the stack: data quality and data pipeline management, machine learning operations, model observability and governance, etc. Companies need to plan for and get ahead of these challenges to deploy AI at scale.

Assess feasibility and impact before investing for scale, and be clear about when to buy versus build

Like any transformative technology early in the adoption cycle, generative AI is rife with hype, partly driven by news media, and marketing messages from some software and tooling vendors. Companies must take these claims with a grain of salt, and instead choose to experiment aggressively to understand generative AI's capabilities and impact in their contexts. They must continue to make investments in talent and dive deep into the right use cases, while carefully assessing "build versus buy" decisions in these early days. Sometimes, saying "no" and waiting to get the timing right is more strategic.

Conclusion

Generative AI is here to stay and needs to be a strategic priority for businesses today. There is no doubt that it will disrupt industries and will create a step change in productivity. However, it is also complex, rapidly evolving, and can be very expensive at scale. While it is imperative to move quickly, it is equally important to carefully prioritize where to place your bets.

To avoid missteps and future-proof their generative AI investments, companies ought to start with a clear strategy and understanding of use cases. They need to assess if generative AI is a sufficiently mature technology for their use cases and test within their organizational contexts. They also need to fully understand the technology options, trends, and tradeoffs to make the right technology bet, from Models-as-a-Service to building custom models. They need to know when to say "no" and wait, and fully think through when to build versus buy. Most importantly, companies must understand the economics and full scope of costs—primary and secondary—to realistically estimate ROI. And if they do decide to take action, they must be prepared to solve basic but fundamental problems: data processes, talent, and day-two operations. Making strategic and focused generative AI bets—without rushing hastily—can save your company from costly missteps at the very least, or, at best, dramatically accelerate your company's position in the market.

About the authors

Pranay Ahlawat is a partner and associate director in the firm's Washington D.C. office.
You may contact him at ahlawat.pranay@bcg.com.

Drake Watten is managing director and partner in firm's San Francisco office
You may contact him at Watten.Drake@bcg.com

Matt Kropp is managing director and senior partner in firm's San Francisco office.
You may contact him at Kropp.Matthew@bcg.com

Vlad Lukic is managing director and senior partner in firm's Boston office.
You may contact him at Lukic.Vladimir@bcg.com

For further contact

If you would like to discuss this report, please contact the authors.

Acknowledgments

The authors thank the following for their contributions to the development of this report: Aakash Joshi and Sai Masipeddi from BCG and Phil Le-Brun, Archana Vemulapalli, Tom Adams, Ahmad Tawil, Susane Seitinger, Adil Soofi, Ritesh Vajariya, Kamran Khan, Joe Senerchia, Nitin Nagarkatte, Salman Taherian, Jake Burns, Ross Richards and Priya Arora from AWS.

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.