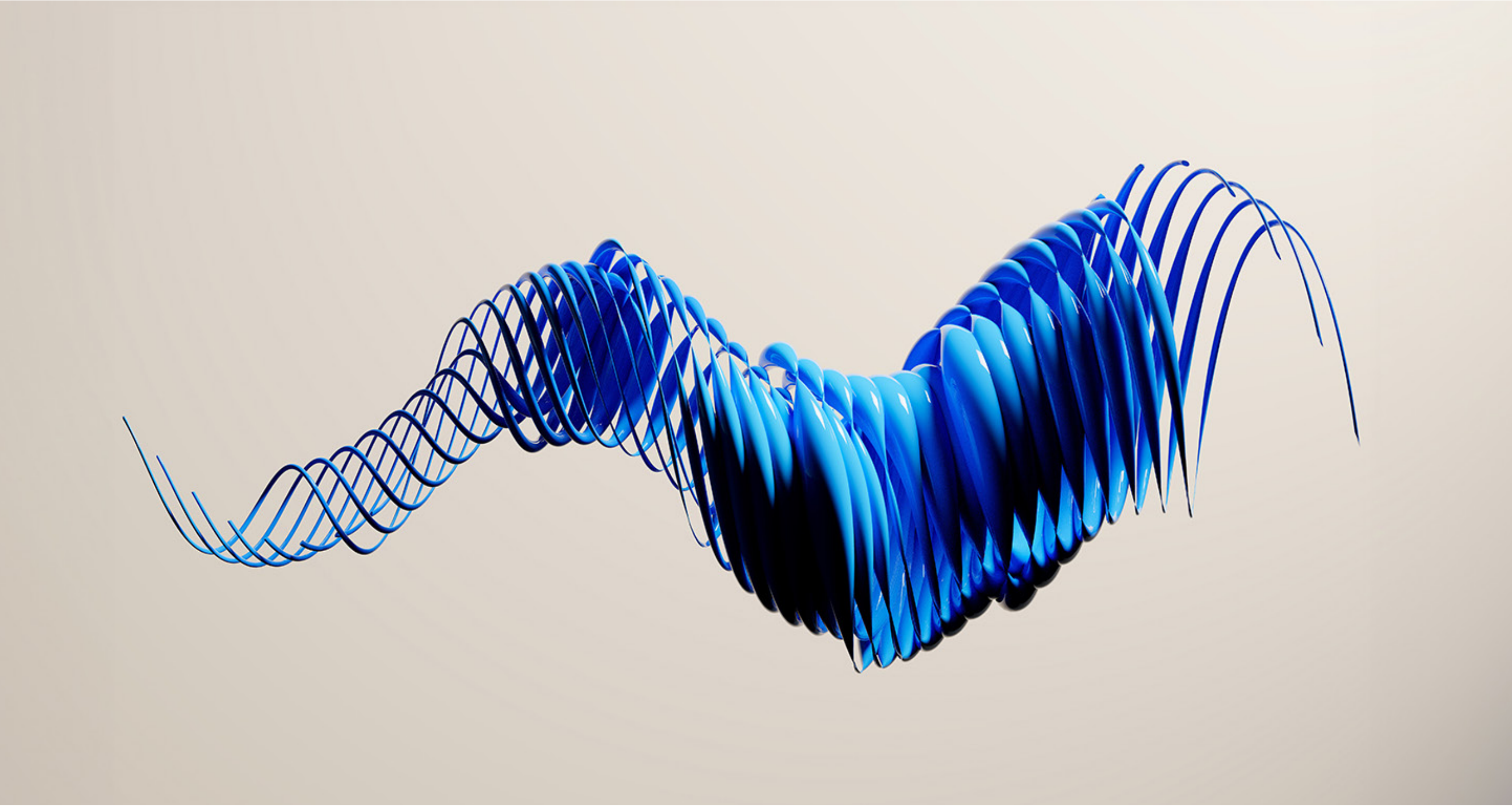
McKinsey
Digital

Technology's generational moment with generative AI: A CIO and CTO guide

CIOs and CTOs can take nine actions to reimagine business and technology with generative AI.

This article is a collaborative effort by Aamer Baig, Sven Blumberg, Eva Li, Douglas Merrill, Adi Pradhan, Megha Sinha, Alexander Sukharevsky, and Stephen Xu, representing views from McKinsey Digital.



© Getty Images

July 2023

Hardly a day goes by without some new business-busting development related to generative AI surfacing in the media. The excitement is well deserved—McKinsey research estimates that generative AI could add the equivalent of $2.6 trillion to $4.4 trillion of value annually.[1]

CIOs and chief technology officers (CTOs) have a critical role in capturing that value, but it's worth remembering we've seen this movie before. New technologies emerged—the internet, mobile, social media—that set off a melee of experiments and pilots, though significant business value often proved harder to come by. Many of the lessons learned from those developments still apply, especially when it comes to getting past the pilot stage to reach scale. For the CIO and CTO, the generative AI boom presents a unique opportunity to apply those lessons to guide the C-suite in turning the promise of generative AI into sustainable value for the business.

Through conversations with dozens of tech leaders and an analysis of generative AI initiatives at more than 50 companies (including our own), we have identified nine actions all technology leaders can take to create value, orchestrate technology and data, scale solutions, and manage risk for generative AI (see sidebar, "A quick primer on key terms"):

1. Move quickly to **determine the company's posture for the adoption of generative AI**, and develop practical communications to, and appropriate access for, employees.

A quick primer on key terms

Generative AI is a type of AI that can create new content (text, code, images, video) using patterns it has learned by training on extensive (public) data with machine learning (ML) techniques.

Foundation models (FMs) 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 2, DALL·E 2, and Stable Diffusion.

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. Cohere Command is one type of LLM; LaMDA is the LLM behind Bard.

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.

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.

Learn more about generative AI in our explainer "What is generative AI" on McKinsey.com.

[1] "The economic potential of generative AI: The next productivity frontier," McKinsey, June 14, 2023.

2. Reimagine the business and **identify use cases that build value through improved productivity, growth, and new business models.** Develop a "financial AI" (FinAI) capability that can estimate the true costs and returns of generative AI.

3. **Reimagine the technology function,** and focus on quickly building generative AI capabilities in software development, accelerating technical debt reduction, and dramatically reducing manual effort in IT operations.

4. **Take advantage of existing services or adapt open-source generative AI models** to develop proprietary capabilities (building and operating your own generative AI models can cost tens to hundreds of millions of dollars, at least in the near term).

5. **Upgrade your enterprise technology architecture to integrate and manage generative AI models** and orchestrate how they operate with each other and existing AI and machine learning (ML) models, applications, and data sources.

6. **Develop a data architecture to enable access to quality data** by processing both structured and unstructured data sources.

7. **Create a centralized, cross-functional generative AI platform team t**o provide approved models to product and application teams on demand.

8. Invest in upskilling key roles—software developers, data engineers, MLOps engineers, and security experts—as well as the broader nontech workforce. But you need to **tailor the training programs by roles and proficiency levels** due to the varying impact of generative AI.

9. **Evaluate the new risk landscape and establish ongoing mitigation practices** to address models, data, and policies.

1. Determine the company's posture for the adoption of generative AI

As use of generative AI becomes increasingly widespread, we have seen CIOs and CTOs respond by blocking employee access to publicly available applications to limit risk. In doing so, these companies risk missing out on opportunities for innovation, with some employees even perceiving these moves as limiting their ability to build important new skills.

Instead, CIOs and CTOs should work with risk leaders to balance the real need for risk mitigation with the importance of building generative AI skills in the business. This requires establishing the company's posture regarding generative AI by building consensus around the levels of risk with which the business is comfortable and how generative AI fits into the business's overall strategy. This step allows the business to quickly determine company-wide policies and guidelines.

Once policies are clearly defined, leaders should communicate them to the business, with the CIO and CTO providing the organization with appropriate access and user-friendly guidelines. Some companies have rolled out firmwide communications about generative AI, provided broad access to generative AI for specific user groups, created pop-ups that warn users any time they input internal data into a model, and built a guidelines page that appears each time users access a publicly available generative AI service.

2. Identify use cases that build value through improved productivity, growth, and new business models

CIOs and CTOs should be the antidote to the "death by use case" frenzy that we already see in many companies. They can be most helpful by working with the CEO, CFO, and other business leaders to think through how generative AI challenges existing business models, opens doors to new ones, and creates new sources of value. With a deep understanding of the technical possibilities, the

CIO and CTO should identify the most valuable opportunities and issues across the company that can benefit from generative AI—and those that can't. In some cases, generative AI is *not* the best option.

McKinsey research, for example, shows generative AI can lift productivity for certain marketing use cases (for example, by analyzing unstructured and abstract data for customer preference) by roughly 10 percent and customer support (for example, through intelligent bots) by up to 40 percent.[2] The CIO and CTO can be particularly helpful in developing a perspective on how best to cluster use cases either by domain (such as customer journey or business process) or use case type (such as creative content creation or virtual agents) so that generative AI will have the most value. Identifying opportunities won't be the most strategic task—there are many generative AI use cases out there—but, given initial limitations of talent and capabilities, the CIO and CTO will need to provide feasibility and resource estimates to help the business sequence generative AI priorities.

Providing this level of counsel requires tech leaders to work with the business to develop a FinAI capability to estimate the true costs and returns on generative AI initiatives. Cost calculations can be particularly complex because the unit economics must account for multiple model and vendor costs, model interactions (where a query might require input from multiple models, each with its own fee), ongoing usage fees, and human oversight costs.

3. Reimagine the technology function

Generative AI has the potential to completely remake how the tech function works. CIOs and CTOs need to make a comprehensive review of the potential impact of generative AI on all areas of tech, but it's important to take action quickly to build experience and expertise. There are three areas where they can focus their initial energies:

— **Software development:** McKinsey research shows generative AI coding support can help software engineers develop code 35 to 45 percent faster, refactor code 20 to 30 percent faster, and perform code documentation 45 to 50 percent faster.[3] Generative AI can also automate the testing process and simulate edge cases, allowing teams to develop more-resilient software prior to release, and accelerate the onboarding of new developers (for example, by asking generative AI questions about a code base). Capturing these benefits will require extensive training (see more in action 8) and automation of integration and deployment pipelines through DevSecOps practices to manage the surge in code volume.

— **Technical debt:** Technical debt can account for 20 to 40 percent of technology budgets and significantly slow the pace of development.[4] CIOs and CTOs should review their tech-debt balance sheets to determine how generative AI capabilities such as code refactoring, code translation, and automated test-case generation can accelerate the reduction of technical debt.

— **IT operations (ITOps):** CIOs and CTOs will need to review their ITOps productivity efforts to determine how generative AI can accelerate processes. Generative AI's capabilities are particularly helpful in automating such tasks as password resets, status requests, or basic diagnostics through self-serve agents; accelerating triage and resolution through improved routing; surfacing useful context, such as topic or priority, and generating suggested responses; improving observability through analysis of vast streams of logs to identify events that truly require attention; and developing documentation, such as standard operating procedures, incident postmortems, or performance reports.

[2] Ibid.

[3] Begum Karaci Deniz, Martin Harrysson, Alharith Hussin, and Shivam Srivastava, "Unleashing developer productivity with generative AI," McKinsey, June 27, 2023.

[4] Vishal Dalal, Krish Krishnakanthan, Björn Münstermann, and Rob Patenge, "Tech debt: Reclaiming tech equity," McKinsey, October 6, 2020.

4. Take advantage of existing services or adapt open-source generative AI models

A variation of the classic "rent, buy, or build" decision exists when it comes to strategies for developing generative AI capabilities. The basic rule holds true: a company should invest in a generative AI capability where it can create a proprietary advantage for the business and access existing services for those that are more like commodities.

The CIO and CTO can think through the implications of these options as three archetypes:

- **Taker**—uses publicly available models through a chat interface or an API, with little or no customization. Good examples include off-the-shelf solutions to generate code (such as GitHub Copilot) or to assist designers with image generation and editing (such as Adobe Firefly). This is the simplest archetype in terms of both engineering and infrastructure needs and is generally the fastest to get up and running. These models are essentially commodities that rely on feeding data in the form of prompts to the public model.

- **Shaper**—integrates models with internal data and systems to generate more customized results. One example is a model that supports sales deals by connecting generative AI tools to customer relationship management (CRM) and financial systems to incorporate customers' prior sales and engagement history. Another is fine-tuning the model with internal company documents and chat history to act as an assistant to a customer support agent. For companies that are looking to scale generative AI capabilities, develop more proprietary capabilities, or meet higher security or compliance needs, the Shaper archetype is appropriate.

  There are two common approaches for integrating data with generative AI models in this archetype. One is to "bring the model to the data," where the model is hosted on the organization's infrastructure, either on-premises or in the cloud environment. Cohere, for example, deploys foundation models on clients' cloud infrastructure, reducing the need for data transfers. The other approach is to "bring data to the model," where an organization can aggregate its data and deploy a copy of the large model on cloud infrastructure. Both approaches achieve the goal of providing access to the foundation models, and choosing between them will come down to the organization's workload footprint.

- **Maker**—builds a foundation model to address a discrete business case. Building a foundation model is expensive and complex, requiring huge volumes of data, deep expertise, and massive compute power. This option requires a substantial one-off investment—tens or even hundreds of millions of dollars—to build the model and train it. The cost depends on various factors, such as training infrastructure, model architecture choice, number of model parameters, data size, and expert resources.

Each archetype has its own costs that tech leaders will need to consider (Exhibit 1). While new developments, such as efficient model training approaches and lower graphics processing unit (GPU) compute costs over time, are driving costs down, the inherent complexity of the Maker archetype means that few organizations will adopt it in the short term. Instead, most will turn to some combination of Taker, to quickly access a commodity service, and Shaper, to build a proprietary capability on top of foundation models.

5. Upgrade your enterprise technology architecture to integrate and manage generative AI models

Organizations will use many generative AI models of varying size, complexity, and capability. To generate value, these models need to be able to work both together and with the business's existing systems or applications. For this reason, building a separate tech stack for generative AI creates more complexities than it solves. As an example, we can look at a consumer querying customer service at a travel company to resolve a booking issue (Exhibit 2). In interacting with the customer, the generative AI model needs to access multiple applications and data sources.

Exhibit 1

Each archetype has its own costs.

Archetype	Example use cases	Estimated total cost of ownership
Taker	— Off-the-shelf coding assistant for software developers — General-purpose customer service chatbot with prompt engineering only and text chat only	**~ $0.5 million to $2.0 million, one-time** — Off-the-shelf coding assistant: ~$0.5 million for integration. Costs include a team of 6 working for 3 to 4 months. — General-purpose customer service chatbot: ~$2.0 million for building plug-in layer on top of 3rd-party model API. Costs include a team of 8 working for 9 months. **~ $0.5 million, recurring annually** — Model inference: • Off-the-shelf coding assistant: ~$0.2 million annually per 1,000 daily users • General-purpose customer service chatbot: ~$0.2 million annually, assuming 1,000 customer chats per day and 10,000 tokens per chat — Plug-in-layer maintenance: up to ~$0.2 million annually, assuming 10% of development cost.
Shaper	— Customer service chatbot fine-tuned with sector-specific knowledge and chat history	**~ $2.0 million to $10.0 million, one-time unless model is fine-tuned further** — Data and model pipeline building: ~$0.5 million. Costs include 5 to 6 machine learning engineers and data engineers working for 16 to 20 weeks to collect and label data and perform data ETL.[1] — Model fine-tuning[2]: ~$0.1 million to $6.0 million per training run[3] • Lower end: costs include compute and 2 data scientists working for 2 months • Upper end: compute based on public closed-source model fine-tuning cost — Plug-in-layer building: ~$1.0 million to $3.0 million. Costs include a team of 6 to 8 working for 6 to 12 months. **~ 0.5 million to $1.0 million, recurring annually** — Model inference: up to ~$0.5 million recurring annually. Assume 1,000 chats daily with both audio and texts. — Model maintenance: ~$0.5 million. Assume $100,000 to $250,000 annually for MLOps platform[4] and 1 machine learning engineer spending 50% to 100% of their time monitoring model performance. — Plug-in-layer maintenance: up to ~$0.3 million recurring annually, assuming 10% of development cost.
Maker	— Foundation model trained for assisting in patient diagnosis	**~ $5.0 million to $200.0 million, one-time unless model is fine-tuned or retrained** — Model development: ~$0.5 million. Costs include 4 data scientists spending 3 to 4 months on model design, development, and evaluation leveraging existing research. — Data and model pipeline: ~$0.5 million to $1.0 million. Costs include 6 to 8 machine learning engineers and data engineers working for ~12 weeks to collect data and perform data ETL.[1] — Model training[5]: ~$4.0 million to $200.0 million per training run.[3] Costs include compute and labor cost of 4 to 6 data scientists working for 3 to 6 months. — Plug-in-layer building: ~$1.0 million to $3.0 million. Costs include a team of 6 to 8 working 6 to 12 months. **~ $1.0 million to $5.0 million, recurring annually** — Model inference: ~$0.1 million to $1.0 million annually per 1,000 users. Assume each physician sees 20 to 25 patients per day and patient speaks for 6 to 25 minutes per visit. — Model maintenance: ~$1.0 million to $4.0 million recurring annually. Assume $250,000 annually for MLOps platform[4] and 3 to 5 machine learning engineers to monitor model performance. — Plug-in-layer maintenance: up to ~$0.3 million recurring annually, assuming 10% of development cost.

Note: Through engineering optimizations, the economics of generative AI are evolving rapidly, and these are high-level estimates based on total cost of ownership (resources, model training, etc) as of mid-2023.

[1] Extract, transform, and load.

[2] Model is fine-tuned on data set consisting of ~100,000 pages of sector-specific documents and 5 years of chat history from ~1,000 customer representatives, which is ~48 billion tokens. Lower end cost consists of 1% parameters retrained on open-source models (eg, LLaMA) and upper end on closed-source models. Chatbot can be accessed via both text and audio.

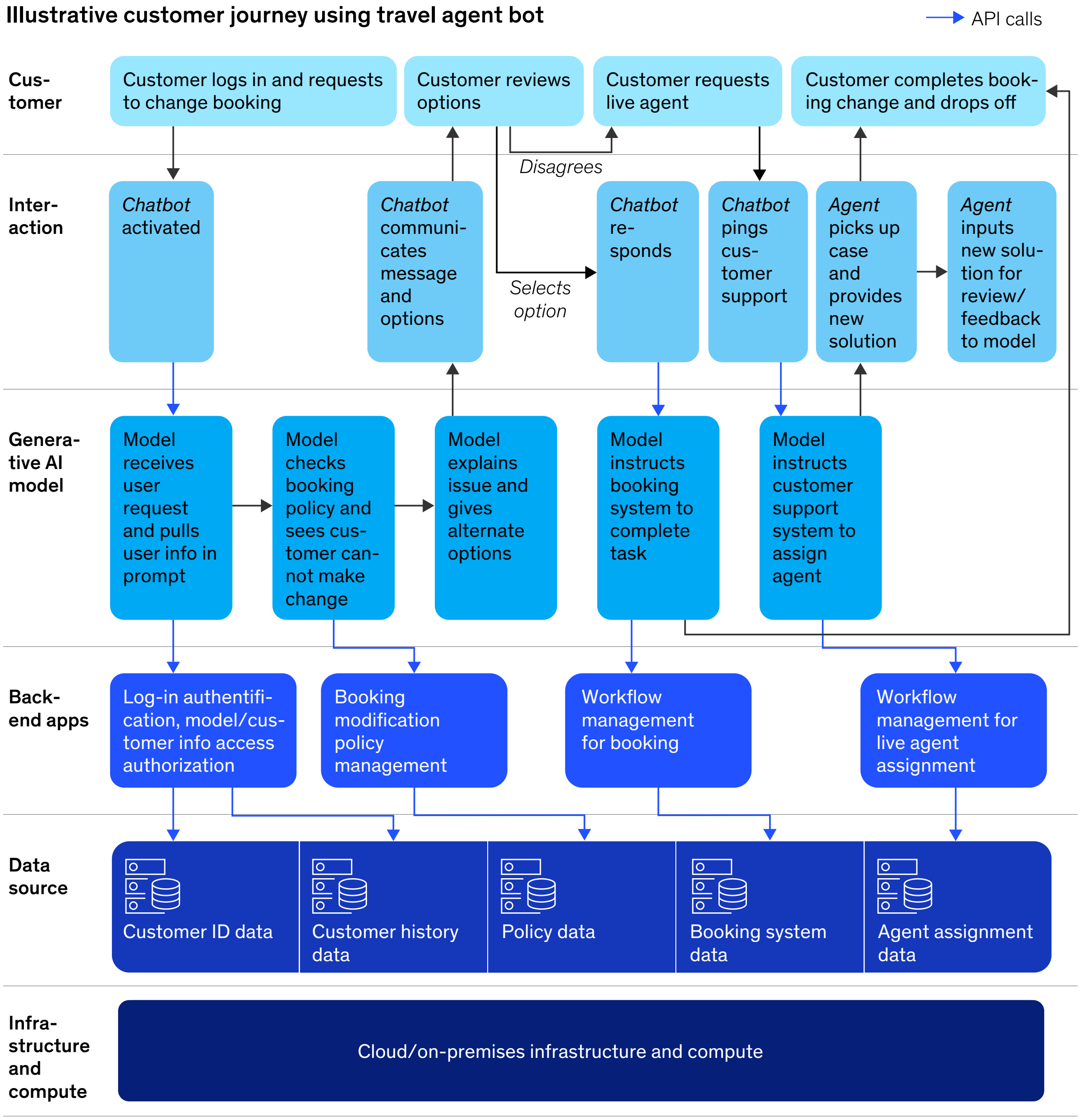
[3] Model is optimized after each training run based on use of hyperparameters, data set, and model architecture. Model may be refreshed periodically when needed (eg, with fresh data).

[4] Gilad Shaham, "Build or buy your MLOps platform: Main considerations," LinkedIn, November 3, 2021.

[5] Model is trained on 65 billion to 1 trillion parameters and data set of 1.2 to 2.4 trillion tokens. The tool can be accessed via both text and audio.

Exhibit 2

Generative AI is integrated at key touchpoints to enable a tailored customer journey.



Illustrative customer journey using travel agent bot
API calls
Cus-tomer
Customer logs in and requests to change booking
Customer reviews options
Customer requests live agent
Customer completes booking change and drops off
Disagrees
Inter-action
Chatbot activated
Chatbot communicates message and options
Selects option
Chatbot responds
Chatbot pings customer support
Agent picks up case and provides new solution
Agent inputs new solution for review/ feedback to model
Generative AI model
Model receives user request and pulls user info in prompt
Model checks booking policy and sees customer cannot make change
Model explains issue and gives alternate options
Model instructs booking system to complete task
Model instructs customer support system to assign agent
Back-end apps
Log-in authentification, model/customer info access authorization
Booking modification policy management
Workflow management for booking
Workflow management for live agent assignment
Data source
Customer ID data
Customer history data
Policy data
Booking system data
Agent assignment data
Infra-structure and compute
Cloud/on-premises infrastructure and compute

For the Taker archetype, this level of coordination isn't necessary. But for companies looking to scale the advantages of generative AI as Shapers or Makers, CIOs and CTOs need to upgrade their technology architecture. The prime goal is to integrate generative AI models into internal systems and enterprise applications and to build pipelines to various data sources. Ultimately, it's the maturity of the business's enterprise technology architecture that allows it to integrate and scale its generative AI capabilities.

Recent advances in integration and orchestration frameworks, such as LangChain and LlamaIndex, have significantly reduced the effort required to connect different generative AI models with other applications and data sources. Several integration patterns are also emerging, including those that enable models to call APIs when responding to a user query—GPT-4, for example, can invoke functions—and provide contextual data from an external data set as part of a user query, a technique known as retrieval augmented generation. Tech leaders will need to define reference architectures and standard integration patterns for their organization (such as standard API formats and parameters that identify the user and the model invoking the API).

There are five key elements that need to be incorporated into the technology architecture to integrate generative AI effectively (Exhibit 3):

— **Context management and caching** to provide models with relevant information from enterprise data sources. Access to relevant data at the right time is what allows the model to understand the context and produce compelling outputs. Caching stores results to frequently asked questions to enable faster and cheaper responses.

— **Policy management** to ensure appropriate access to enterprise data assets. This control ensures that HR's generative AI models that include employee compensation details, for example, cannot be accessed by the rest of the organization.

— **Model hub,** which contains trained and approved models that can be provisioned on demand and acts as a repository for model checkpoints, weights, and parameters.

— **Prompt library,** which contains optimized instructions for the generative AI models, including prompt versioning as models are updated.

— **MLOps platform,** including upgraded MLOps capabilities, to account for the complexity of generative AI models. MLOps pipelines, for example, will need to include instrumentation to measure task-specific performance, such as measuring a model's ability to retrieve the right knowledge.

In evolving the architecture, CIOs and CTOs will need to navigate a rapidly growing ecosystem of generative AI providers and tooling. Cloud providers provide extensive access to at-scale hardware and foundation models, as well as a proliferating set of services. MLOps and model hub providers, meanwhile, offer the tools, technologies, and practices to adapt a foundation model and deploy it into production, while other companies provide applications directly accessed by users built on top of foundation models to perform specific tasks. CIOs and CTOs will need to assess how these various capabilities are assembled and integrated to deploy and operate generative AI models.

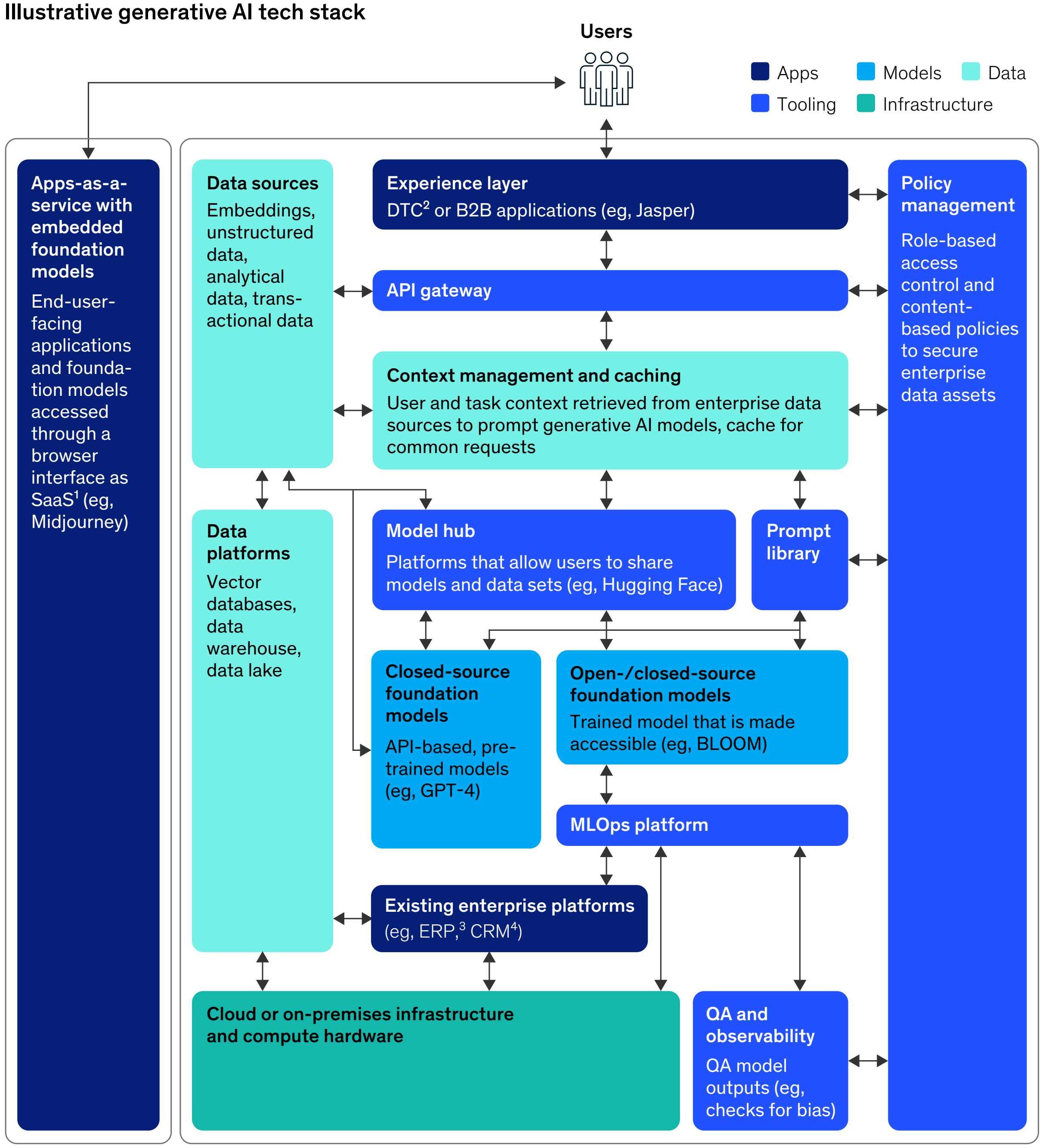
6. Develop a data architecture to enable access to quality data

The ability of a business to generate and scale value, including cost reductions and improved data and knowledge protections, from generative AI models will depend on how well it takes advantage of its own data. Creating that advantage relies on a data architecture that connects generative AI models to internal data sources, which provide context or help fine-tune the models to create more relevant outputs.

Exhibit 3

The tech stack for generative AI is emerging.

Illustrative generative AI tech stack



Apps-as-a-service with embedded foundation models
End-user-facing applications and foundation models accessed through a browser interface as SaaS[1] (eg, Midjourney)

Data sources
Embeddings, unstructured data, analytical data, transactional data

Experience layer
DTC[2] or B2B applications (eg, Jasper)

API gateway

Context management and caching
User and task context retrieved from enterprise data sources to prompt generative AI models, cache for common requests

Policy management
Role-based access control and content-based policies to secure enterprise data assets

Data platforms
Vector databases, data warehouse, data lake

Model hub
Platforms that allow users to share models and data sets (eg, Hugging Face)

Prompt library

Closed-source foundation models
API-based, pre-trained models (eg, GPT-4)

Open-/closed-source foundation models
Trained model that is made accessible (eg, BLOOM)

MLOps platform

Existing enterprise platforms
(eg, ERP,[3] CRM[4])

Cloud or on-premises infrastructure and compute hardware

QA and observability
QA model outputs (eg, checks for bias)

[1] Software as a service.
[2] Direct to consumer.
[3] Enterprise resource planning.
[4] Customer relationship management.

In this context, CIOs, CTOs, and chief data officers need to work closely together to do the following:

- Categorize and organize data so it can be used by generative AI models. Tech leaders will need to develop a comprehensive data architecture that encompasses both structured and unstructured data sources. This requires putting in place standards and guidelines to optimize data for generative AI use—for example, by augmenting training data with synthetic samples to improve diversity and size; converting media types into standardized data formats; adding metadata to improve traceability and data quality; and updating data.

- Ensure existing infrastructure or cloud services can support the storage and handling of the vast volumes of data needed for generative AI applications.

- Prioritize the development of data pipelines to connect generative AI models to relevant data sources that provide "contextual understanding." Emerging approaches include the use of vector databases to store and retrieve embeddings (specially formatted knowledge) as input for generative AI models as well as in-context learning approaches, such as "few shot prompting," where models are provided with examples of good answers.

7. Create a centralized, cross-functional generative AI platform team

Most tech organizations are on a journey to a product and platform operating model. CIOs and CTOs need to integrate generative AI capabilities into this operating model to build on the existing infrastructure and help to rapidly scale adoption of generative AI. The first step is setting up a generative AI platform team whose core focus is developing and maintaining a platform service where approved generative AI models can be provisioned on demand for use by product and application teams. The platform team also defines protocols for how generative AI models integrate with internal systems, enterprise applications, and tools, and also develops and implements standardized approaches to manage risk, such as responsible AI frameworks.

CIOs and CTOs need to ensure that the platform team is staffed with people who have the right skills. This team requires a senior technical leader who acts as the general manager. Key roles include software engineers to integrate generative AI models into existing systems, applications, and tools; data engineers to build pipelines that connect models to various systems of record and data sources; data scientists to select models and engineer prompts; MLOps engineers to manage deployment and monitoring of multiple models and model versions; ML engineers to fine-tune models with new data sources; and risk experts to manage security issues such as data leakage, access controls, output accuracy, and bias. The exact composition of the platform team will depend on the use cases being served across the enterprise. In some instances, such as creating a customer-facing chatbot, strong product management and user experience (UX) resources will be required.

Realistically, the platform team will need to work initially on a narrow set of priority use cases, gradually expanding the scope of their work as they build reusable capabilities and learn what works best. Technology leaders should work closely with business leads to evaluate which business cases to fund and support.

8. Tailor upskilling programs by roles and proficiency levels

Generative AI has the potential to massively lift employees' productivity and augment their capabilities. But the benefits are unevenly distributed depending on roles and skill levels, requiring leaders to rethink how to build the actual skills people need.

Our latest empirical research using the generative AI tool GitHub Copilot, for example, helped software engineers write code 35 to 45 percent faster.[5] The

[5] "Unleashing developer productivity with generative AI," June 27, 2023.

benefits, however, varied. Highly skilled developers saw gains of up to 50 to 80 percent, while junior developers experienced a 7 to 10 percent *decline* in speed. That's because the output of the generative AI tools requires engineers to critique, validate, and improve the code, which inexperienced software engineers struggle to do. Conversely, in less technical roles, such as customer service, generative AI helps low-skill workers significantly, with productivity increasing by 14 percent and staff turnover dropping as well, according to one study.[6]

These disparities underscore the need for technology leaders, working with the chief human resources officer (CHRO), to rethink their talent management strategy to build the workforce of the future. Hiring a core set of top generative AI talent will be important, and, given the increasing scarcity and strategic importance of that talent, tech leaders should put in place retention mechanisms, such as competitive salaries and opportunities to be involved in important strategic work for the business.

Tech leaders, however, cannot stop at hiring. Because nearly every existing role will be affected by generative AI, a crucial focus should be on upskilling people based on a clear view of what skills are needed by role, proficiency level, and business goals. Let's look at software developers as an example. Training for novices needs to emphasize accelerating their path to become top code reviewers in addition to code generators. Similar to the difference between writing and editing, code review requires a different skill set. Software engineers will need to understand what good code looks like; review the code created by generative AI for functionality, complexity, quality, and readability; and scan for vulnerabilities while ensuring they do not themselves introduce quality or security issues in the code. Furthermore, software developers will need to learn to *think* differently when it comes to coding, by better understanding user intent so they can create prompts and define contextual data that help generative AI tools provide better answers.

Beyond training up tech talent, the CIO and CTO can play an important role in building generative AI skills among nontech talent as well. Besides understanding how to use generative AI tools for such basic tasks as email generation and task management, people across the business will need to become comfortable using an array of capabilities to improve performance and outputs. The CIO and CTO can help adapt academy models to provide this training and corresponding certifications.

The decreasing value of inexperienced engineers should accelerate the move away from a classic talent pyramid, where the greatest number of people are at a junior level, to a structure more like a diamond, where the bulk of the technical workforce is made up of experienced people. Practically speaking, that will mean building the skills of junior employees as quickly as possible while reducing roles dedicated to low-complexity manual tasks (such as writing unit tests).

9. Evaluate the new risk landscape and establish ongoing mitigation practices

Generative AI presents a fresh set of ethical questions and risks, including "hallucinations," whereby the generative AI model presents an incorrect response based on the highest-probability response; the accidental release of confidential personally identifiable information; inherent bias in the large data sets the models use; and high degrees of uncertainty related to intellectual property (IP). CIOs and CTOs will need to become fluent in ethics, humanitarian, and compliance issues to adhere not just to the letter of the law (which will vary by country) but also to the spirit of responsibly managing their business's reputation.

Addressing this new landscape requires a significant review of cyber practices and updating the software development process to evaluate risk and identify mitigation actions before model

[6] Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond, *Generative AI at work*, National Bureau of Economic Research (NBER) working paper, number 31161, April 2023.

development begins, which will both reduce issues and ensure the process doesn't slow down. Proven risk-mitigation actions for hallucinations can include adjusting the level of creativity (known as the "temperature") of a model when it generates responses; augmenting the model with relevant internal data to provide more context; using libraries that impose guardrails on what can be generated; using "moderation" models to check outputs; and adding clear disclaimers. Early generative AI use cases should focus on areas where the cost of error is low, to allow the organization to work through inevitable setbacks and incorporate learnings.

To protect data privacy, it will be critical to establish and enforce sensitive data tagging protocols, set up data access controls in different domains (such as HR compensation data), add extra protection when data is used externally, and include privacy safeguards. For example, to mitigate access control risk, some organizations have set up a policy-management layer that restricts access by role once a prompt is given to the model. To mitigate risk to intellectual property, CIOs and CTOs should insist that providers of foundation models maintain transparency regarding the IP (data sources, licensing, and ownership rights) of the data sets used.

Generative AI is poised to be one of the fastest-growing technology categories we've ever seen. Tech leaders cannot afford unnecessary delays in defining and shaping a generative AI strategy. While the space will continue to evolve rapidly, these nine actions can help CIOs and CTOs responsibly and effectively harness the power of generative AI at scale.

Aamer Baig is a senior partner in McKinsey's Chicago office; **Sven Blumberg** is a senior partner in the Düsseldorf office; **Eva Li** is a consultant in the Bay Area office, where **Megha Sinha** is a partner; **Douglas Merrill** is a partner in the Southern California office; **Adi Pradhan** and **Stephen Xu** are associate partners in the Toronto office; and **Alexander Sukharevsky** is a senior partner in the London office.

The authors wish to thank Stephanie Brauckmann, Anusha Dhasarathy, Martin Harrysson, Klemens Hjartar, Alharith Hussin, Naufal Khan, Sam Nie, Chandrasekhar Panda, Henning Soller, Nikhil Srinidhi, Asin Tavakoli, Niels Van der Wildt, and Anna Wiesinger for their contributions to this article.