McKinsey & Company | QuantumBlack AI by McKinsey

Operations Practice

From promising to productive: Real results from gen AI in services

Service organizations that are early generative AI adopters are finding that to capture more value, they need to get more disciplined.

by Jorge Amar and Oana Cheta
with Ivan Huang and Stephen Xu

August 2024

Generative AI (gen AI) could provide the productivity boost operations leaders have hoped for, as well as a means to fight cost pressures—if only leaders could get going. McKinsey's latest tech trends research finds that only 11 percent of companies worldwide are using gen AI at scale.[1]

Operations is a major gap: in a February 2024 survey of 150 executives at large North American and European companies, only 3 percent of respondents said their organization has scaled a gen AI use case in an operations-related domain. A separate survey, conducted in April 2024 of more than 250 corporate-function leaders worldwide, found that service operations is faring only slightly better. In finance functions, for example, about 45 percent of organizations are now piloting gen AI solutions, compared with 11 percent in 2023—but only 6 percent have achieved scale.

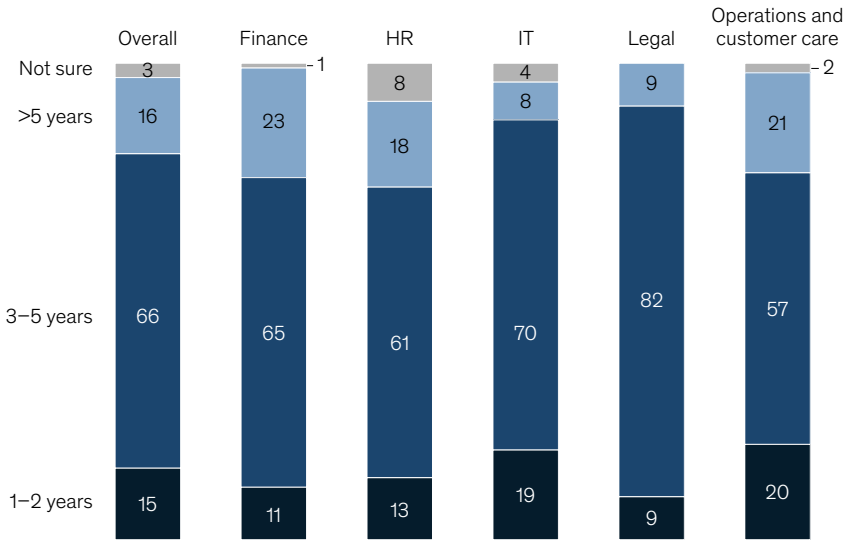
The results reflect uncertainty among operations leaders about which of the many use cases they have deployed will yield real competitive advantage. Executives understand that realizing full value from their gen AI investments won't be instantaneous: two-thirds of the April survey respondents set a three- to five-year timeline (Exhibit 1).

But many also said that they wanted to be more confident that their commitments would pay off. One CEO recently told us, "We've already spent about $100 million funding hundreds of gen AI

Exhibit 1

Most executives expect that it will take three to five years to capture the full value from their generative AI investments.

Expected time to realize value of current generative AI plans, by function, % of respondents



Note: Figures may not sum to 100%, because of rounding.
Source: 2024 McKinsey Corporate and Business Functions CXO Survey, conducted Apr 10–May 30, 2024, n = 276

McKinsey & Company

[1] "Moving past gen AI's honeymoon phase: Seven hard truths for CIOs to get from pilot to scale," McKinsey, May 13, 2024.

experiments; harvesting at least some of the value will help us see where additional investment will be worthwhile." Companies also cited unclear road maps, talent shortages, and immature governance as further impediments to scale.

A few companies, however, are capturing real value already, attributing more than 10 percent of their EBIT to their use of gen AI.[2] Early successes like these reveal three critical tasks in setting gen AI up for scaling across an organization. The first is to design a cohesive, disciplined operational strategy for deploying gen AI. That means prioritizing use cases for long-term value by focusing on their potential to not only transform specific process points or domains but also reimagine complete workflows.

Second, to sustain broad impact over time, companies will need to focus on the enablers supporting the humans who make gen AI work—providing the necessary governance and performance infrastructure while also investing in change management and a continuous innovation culture.

The third task is the culmination of the first two: thoughtfully integrating gen AI tools with human capabilities to create the most advanced solutions, such as autonomous gen AI agents or copilots. The most successful can tackle every step of a complex workflow. At one bank, for instance, a gen AI agent now drafts credit-risk memos, increasing revenue per relationship manager by 20 percent. And a copilot in the finance department of a consumer goods maker is reducing operating expenses relating to financial planning and analysis by between $6 million and $10 million.

Deploying operational gen AI strategically

As with earlier waves of technological change, gen AI raises the specter of pilot purgatory, in which dozens of experiments fail to amount to sustained impact. Organizations that have already built up their capabilities in deploying gen AI tend to see better returns on their gen AI investments over both the short and longer term. They especially excel at thinking through sequencing, with a focus on

Organizations that have already built up their capabilities in deploying gen AI tend to see better returns on their gen AI investments over both the short and longer term.

[2] "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value," McKinsey, May 30, 2024. Out of 876 survey respondents who estimated the proportion of their organization's EBIT that was attributable to gen AI, 46 gave a figure of more than 10 percent.

scalability and reusability so that they can reimagine entire chains of value creation.

Building this sort of maturity in gen AI transformation is now essential, leaving companies little time to waste. Ideally, lessons from lower-risk, earlier applications of gen AI build critical capabilities that help higher-risk (and higher-reward) later applications succeed.

Prioritizing use cases

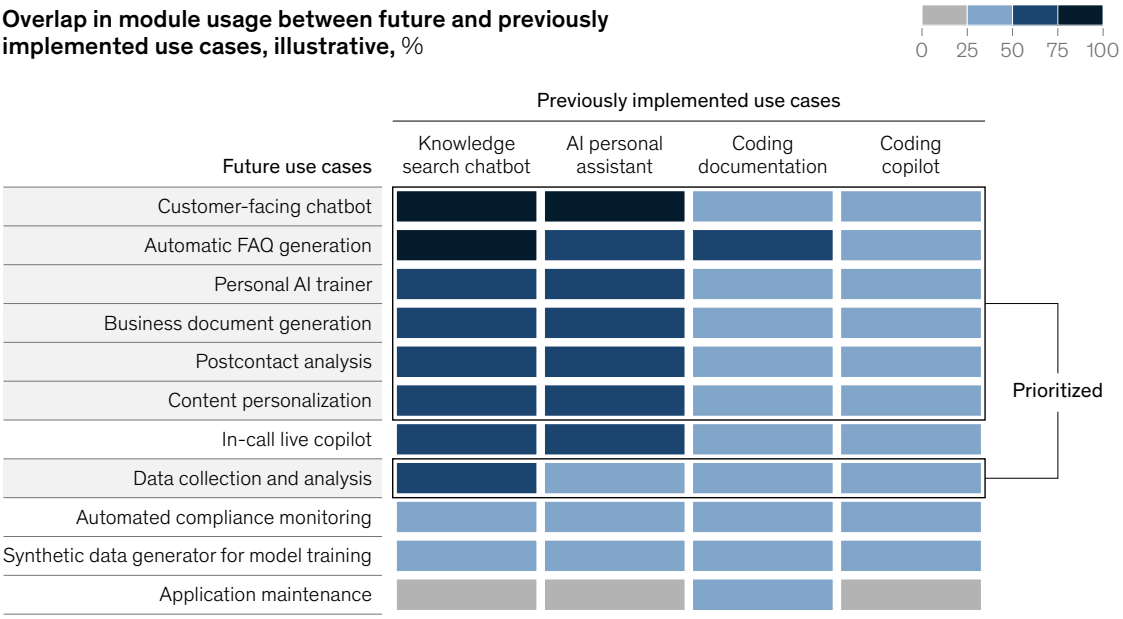
The experience of a global bank illustrates the benefits of deploying gen AI strategically. First, based on a detailed assessment of business impact and technical feasibility, it winnowed 23 potential domains for use of gen AI to just two: the contact center in its consumer banking unit and the know-your-customer (KYC) function for corporate and investment banking. Despite the apparent differences, the two domains not only showed high potential for gen AI impact but also shared a few commonalities, particularly for gen-AI-based knowledge extraction and synthesis. The same technologies could support customers looking for information and employees looking for internal documents, so that the underlying technology could be reused and scaled more effectively (Exhibit 2).

To determine which of the two finalists would go first, the company applied an additional screen: risk. The confidential nature of the KYC function's data made it a higher-risk target, so the bank started instead with the contact center. The final strategic decision concerned which use cases to deploy within customer care. Keeping "ability to scale" and "reusability" top of mind, the chatbot came out on

Exhibit 2

Companies can prioritize generative AI use cases based on potential synergies between underlying modules.

Overlap in module usage between future and previously implemented use cases, illustrative, %



Source: Expert interviews

McKinsey & Company

The core question is: "How could gen AI help me rethink my operations?" Answering it means reexamining each process as part of a larger workflow.

top: it's comparatively easy to implement, generates measurable outcomes, and helps build a foundation for similar use cases that extract and synthesize complex data.

Within just a few weeks, the center's fully designed use cases included a customer-facing chatbot. In just seven weeks of use, the new chatbot offered an improved customer experience, eliminating wait times for about 20 percent of contact center requests.

Moreover, lessons from the contact center have formed a reusable foundation that the bank can adapt for the KYC function. Chatbots are now a component of a "smart, virtual agent" that guides relationship managers through a far more automated KYC process. The virtual agent can prepopulate client information into forms, determine which documents are required, validate data uploads, and follow up on any missing information.

From point solutions to complete workflows

As the bank example illustrates, however, the core operational question for generative AI isn't "How could gen AI help me improve my current processes?" To "improve" a process often means addressing only a symptom rather than the underlying condition—for example, using gen AI to automate note-taking and action-item generation for meetings without considering why there were so many meetings in the first place.

The core question is therefore much broader: "How could gen AI help me rethink my operations?" Answering it means reexamining each process as part of a larger workflow—and, in many cases, as part of a user or customer journey.

Breaking barriers to better service. To illustrate the difference, consider the case of a leading North American telecommunications provider, whose use case prioritization exercise led it to focus on customer care. Rather than start by exploring how gen AI tools could improve particular process steps in care, the company stepped back, asking instead how gen AI could combine with traditional process improvement techniques and new talent to raise productivity within the customer care function overall.

That shift in perspective led the company to reevaluate its customer journeys, starting with a traditional mapping of every touchpoint, from initial contact to final resolution. With the resulting flowcharts in hand, company leaders questioned each process step to see if it was overengineered or unnecessary. It considered the step's effect on customer experience (such as increasing complexity or wait times) versus the potential risks from its elimination (such as increased fraud or security lapses).

For example, after the company mapped out the journey of changing a phone number, one particular

step surfaced as so complex and painful that the company gave customers the option to delegate it to staff in exchange for a fee. But because customers were reluctant to pay, staff would often guide them through the step—a costly alternative for the contact center. Once the company understood the reasons customers got stuck, it could design a self-service solution. In combination with other technologies, gen AI's capabilities to provide detailed, automated guidance meant that the company could reduce average call length (and cost) while eliminating the fee entirely, improving customer experience (Exhibit 3).

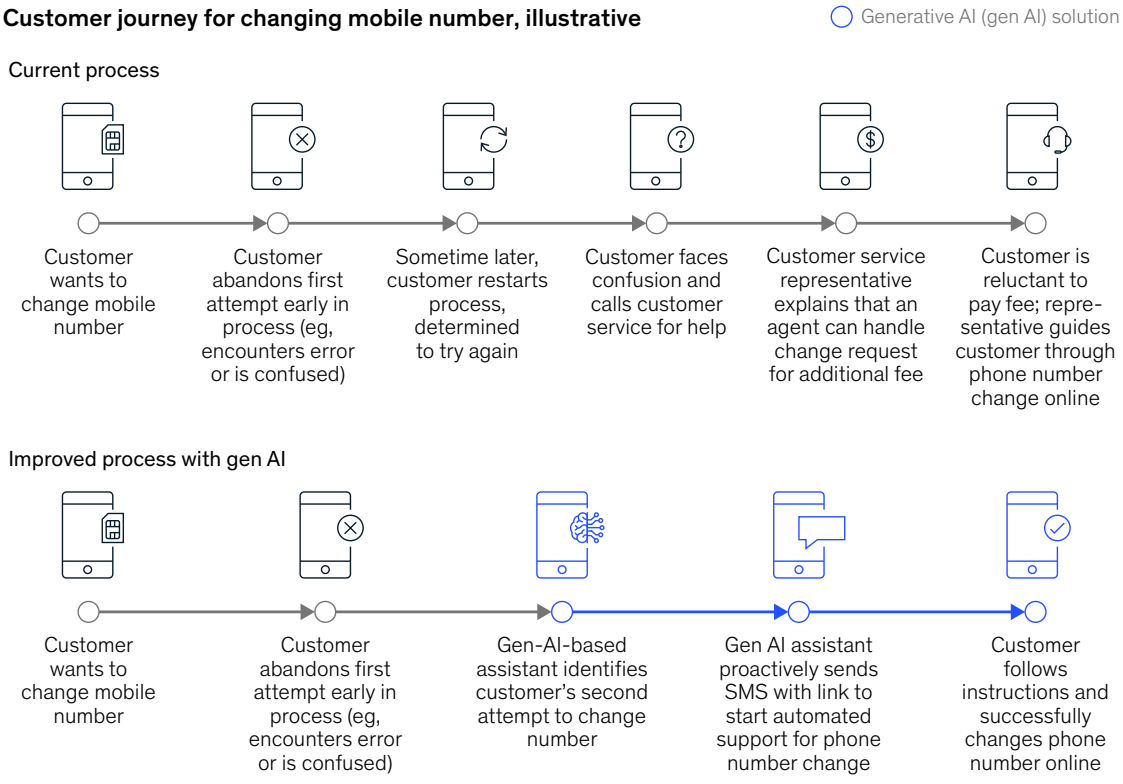
Deeper analysis into the root causes of customer pain points also revealed internal misalignments that the company needed to address before gen AI could provide a solution—such as when price changes set by the marketing team led to surges in customer calls that the care team couldn't handle. Unaware of the changes, agents would transfer customers to other departments, often in multiple loops, leading the care team to provide deep discounts in hopes of retaining the frustrated customers.

Accordingly, the company revamped its cross-functional workflows so that the customer care team could work with marketing to anticipate potential customer concerns and develop appropriate responses in advance. Leaders also reexamined the skills that service teams would need, developing new talent profiles (and associated capability building modules) that

Exhibit 3

Generative-AI-based assistants can proactively message and help customers who get stuck in administrative tasks.

Customer journey for changing mobile number, illustrative



McKinsey & Company

could evolve with the workflows. Changing the internal collaboration model set the foundation for a later, gen-AI-based self-service option, while analytic AI tools could optimize staff allocation to provide additional call center coverage.

Freeing up employees' capacity. Employee journeys were the final piece of the puzzle. The company analyzed every step of the agent experience, from logging in to resolving customer inquiries and completing tasks. This analysis involved streamlining processes and reducing the complexity of technology systems that agents had to interact with. The telecommunications provider also identified potential misalignments between agent incentives and customer needs, ensuring that agents were given incentives to prioritize customer satisfaction and resolution rather than simply handle a high volume of calls.

By taking an integrated approach to workflow optimization as a part of critical journeys, the telecommunications provider achieved significant and lasting improvement in its customer care function, with gen AI building on a range of analog and tech-based improvements. Total call volume fell by about 30 percent, and average handle time by more than one-quarter, even as service quality improved: first-call resolution rates rose by ten to 20 percentage points.

The (human) secret to scale

As with previous technologies, gen AI's full potential depends on its reaching scale throughout an organization. Few companies have reached this point. Their experience underscores the importance of four elements, all of which center on humans rather than technology. The first two elements provide critical guidance; the second two more directly change the way people work, with a particular focus on change management.

Governance

Successful deployment of gen AI can't be ad hoc. This is due to not only gen AI's well-publicized risks—from inaccurate training data for gen-AI-based tools to "hallucinations" that produce incorrect results—but also the tendency of the most advanced organizations (the ones generating more than 10 percent of EBIT from gen AI) to centralize their gen AI initiatives. Almost half of these high performers report centralizing compared with only 35 percent of other companies.

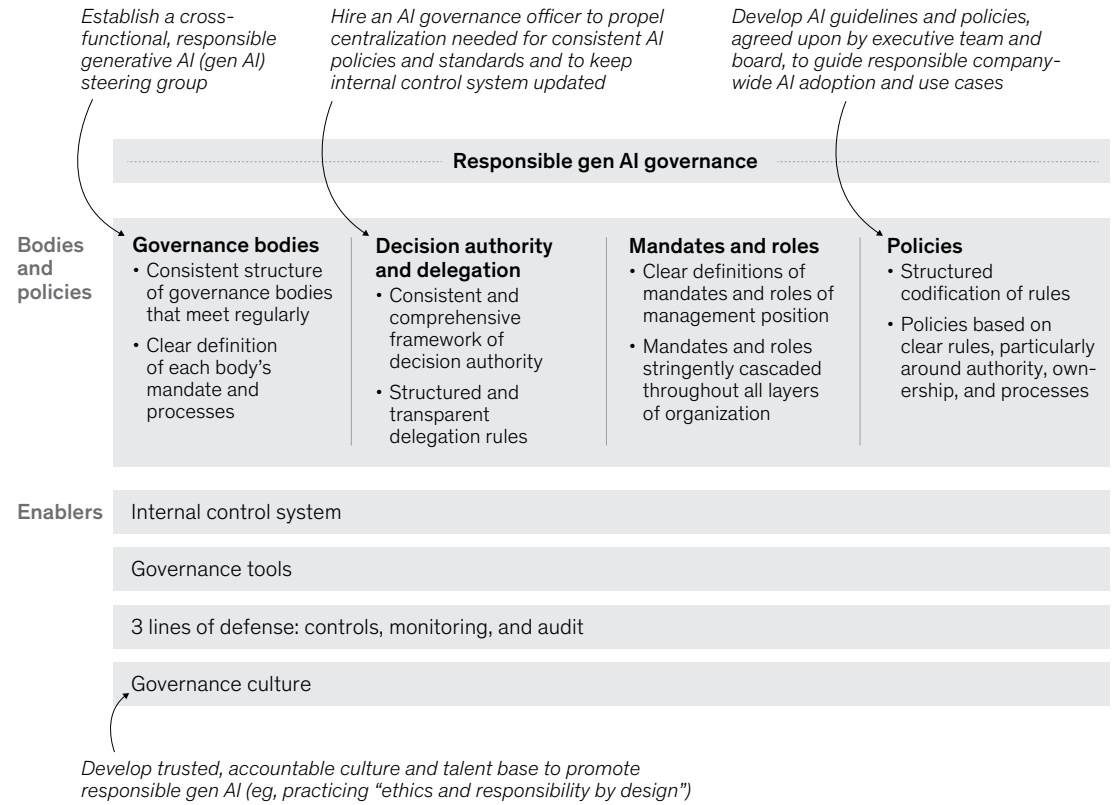
As with previous technologies, gen AI's full potential depends on its reaching scale throughout an organization. Few companies have reached this point.

The components of the governance structure help support rapid implementation and common standards (Exhibit 4). Clear decision rights are especially important for assessing gen AI proposals, supported by a transparent vetting process with well-articulated standards for each stage gate.

Performance infrastructure, data, and analytics
Modernizing performance infrastructure is crucial to accommodate gen AI's changes to the work landscape. The first step is redefining metrics to reflect the company's new operational strategy—and to allow leaders to see how gen AI itself is progressing across the organization. Such metrics can help the organization generate and sustain positive results. Next, a disciplined, stage-gated review process with clear go/no-go criteria separates the merely promising deployments from the ones most likely to be productive. Finally, with better measurement of productivity gains, customer experience improvements, and related outputs, companies can tailor coaching and training programs for human workers and interventions when gen AI's performance lags.

Exhibit 4

A revised governance approach to generative AI can help companies move at speed while mitigating risk.



McKinsey & Company

Change management

It's a truism that changing technology isn't the hard part of transforming an organization—it's changing how people work that's hard. Early experience seems to show that this is even more true for gen AI, for which a good rule of thumb is "for every dollar spent on model development, a company should plan to spend three dollars on change management."

Communication is the starting point. By providing updates on what to expect and addressing potential anxieties, organizations can promote future adoption and create a culture of understanding among employees. But even better than just speaking to staff is to listen: the expertise and knowledge they contribute can make the difference between robust, cost-effective gen-AI-based solutions and gen AI gimmicks with little impact. In parallel, upskilling and reskilling initiatives can help smooth the transition.

Continuous innovation culture

Celebrating successes and sharing best practices is especially vital with a new technology such as gen AI, where innovation cycles are short. Simply keeping abreast of the latest opportunity requires both effort and openness: it's not a question of "buy versus build" but "buy *and* build," continually reviewing what the market offers.

Organizations can foster an environment where frontline workers feel empowered to contribute ideas, whatever their source—and where they feel free to reexamine assumptions about the potential role of partners and vendors in sourcing innovation. By encouraging continuous improvement through feedback and innovation, organizations can optimize the agent and customer experience while maximizing the value of gen AI.

To illustrate, consider the case of a leading European media and telecommunications company. This organization embarked on a mission to industrialize and scale gen AI by 2024, with tangible benefits expected within another year. The company's approach was not merely about chasing the latest tech trend; it was about empowering its workforce and transforming the customer experience. To bring its vision to reality, the company identified a high-impact use case: a gen-AI-powered copilot designed to equip customer service agents with faster and more effective knowledge retrieval during calls.

Keeping agents informed and engaged was a top priority for the company, which hosted weekly working groups to gather qualitative feedback on usability and design. Additionally, quantitative feedback was collected through agent ratings of the AI-generated responses. "Office hours" provided a forum for questions and project updates, fostering a sense of ownership among agents. This transparency helped mitigate potential frustrations and ensured that agents felt invested in the success of the copilot—and led to substantial changes in design.

The user-centric approach proved instrumental not only in refining the copilot but also in encouraging successful scaling. By including frontline agents early in the process, the company made sure that the solution solved real problems in current processes and improved customer service and agent experience. The end result was a 65 percent reduction in average handle time for agents in finding relevant knowledge.

Honing gen AI's cutting edge

The most advanced companies are already combining tools, which can help overcome some of the limitations of the large language models (LLMs) and retrieval-augmented-generation (RAG) technologies at the core of gen AI's initial wave. In particular, LLMs and RAG struggle with complex processes—but automating only part of a process, even at a high level of reliability, often doesn't free enough worker time to create much benefit. Lack of cross-verification can leave LLMs and RAG prone to error. LLMs are also limited to text applications, while both LLMs and RAG using multiple data sources are expensive to build and scale. Finally, LLMs have only limited capabilities with quantitative analysis, making entire streams of value unreachable.

Find more content like this on the
McKinsey Insights App
Scan • Download • Personalize

By contrast, a multiagent system helps humans coach machines to perform complex workflows, often by augmenting models with human expertise. By recursively breaking down a process into smaller tasks until each task can be executed reliably, the multiagent corrects itself, improving the quality of the outcome.

This approach has allowed a North American bank to transform its workflows for writing credit risk memos, a tedious and time-consuming task with little room for error. Relationship managers (RMs) would spend one to three days gathering data from a dozen sources (or more), analyzing multiple interdependencies, and, finally, writing a 20-page memo providing nuanced reasoning to support a lending decision.

A multiagent system now automatically identifies the correct data sources, ingests up-to-date data, and integrates qualitative and quantitative insights that reflect the latest business rules and products. It cites the data sources for each assumption and provides the key rationale behind quantitative trends, while also generating insightful comments based on the data it has integrated, content from prior memos, the RM's feedback, and human expertise.

Credit decisions are now 30 percent faster, while RM productivity has more than doubled. Most important, revenue per RM has risen by 20 percent.

Renewing gen AI potential

With so many companies already investing in gen AI, the question isn't where to start but how to find gen AI's potential quickly—and start earning the rewards. For the 97 percent of companies that haven't yet scaled gen AI in their operations, there's an urgent need for focus. At the same time, gen AI technology is moving so quickly that companies can often find opportunity simply by reassessing where they stand on a two- to three-month cadence. Issues to consider include the following:

— What is our current backlog of gen AI ideas? The first step is to review the ideas that haven't yet moved forward.

— What's feasible now? New combinations of gen AI tools mean that problems too difficult to solve a quarter ago may now be within reach.

— Which pilots aren't showing value? Rationalizing gen AI efforts means letting go of gen AI ideas that aren't meeting their promise, no matter how tantalizing. They might come back later, though, as tools evolve.

— Is our gen AI transformation approach still working? As the organization gains maturity in deploying gen AI, its transformation approach will likely need to evolve—becoming more agile as technologies become more familiar or more robust as risks come into focus. Applying a continuous-improvement mindset to the transformation itself helps ensure that it keeps finding more value from gen AI.

With better performance, greater specialization, and increased accessibility, gen AI can revolutionize service operations across industries. A few judicious decisions now could help a company leapfrog its competitors in creating new and sustained value.

Jorge Amar is a senior partner in McKinsey's Miami office, where **Ivan Huang** is a consultant; **Oana Cheta** is a partner in the Chicago office; and **Stephen Xu** is a senior director of product management in the Toronto office.

The authors wish to thank Abhishek Shirali, Carlo Giovine, Ed Woodcock, Jamie Vickers, Jan Svoboda, Julian Raabe, Max Gemeinhardt, Maximilian Haug, Roderick Lamb, Simon Barres, Sohrab Rahimi, Stefan Moritz, and Victória Lei for their contributions to this article.

Designed by McKinsey Global Publishing
Copyright © 2024 McKinsey & Company. All rights reserved.