



Moving past gen AI's honeymoon phase: Seven hard truths for CIOs to get from pilot to scale

Getting to scale requires CIOs to focus on fewer things but do them better.

This article is a collaborative effort by Aamer Baig, Douglas Merrill, and Megha Sinha, with Danesha Mead and Stephen Xu, representing views from McKinsey Technology and QuantumBlack, AI by McKinsey.

The honeymoon phase of generative AI (gen AI) is over. As most organizations are learning, it is relatively easy to build gee-whiz gen AI pilots, but turning them into at-scale capabilities is another story. The difficulty in making that leap goes a long way to explaining why just 11 percent of companies have adopted gen AI at scale, according to our latest tech trends research.¹

This maturing phase is a welcome development because it gives CIOs an opportunity to turn gen AI's promise into business value. Yet while most CIOs know that pilots don't reflect real-world scenarios—that's not really the point of a pilot, after all—they often underestimate the amount of work that needs to be done to get gen AI production ready. Ultimately, getting the full value from gen AI requires companies to rewire how they work, and putting in place a scalable technology foundation is a key part of that process.

We explored many of the key initial technology issues in a previous article.² In this article, we want to explore seven truths about scaling gen AI for the “Shaper” approach, in which companies develop a competitive advantage by connecting large language models (LLMs) to internal applications and data sources (see sidebar “Three approaches to using gen AI” for more). Here are seven things that Shapers need to know and do:

1. **Eliminate the noise, and focus on the signal.** Be honest about what pilots have worked. Cut down on experiments. Direct your efforts toward solving important business problems.
2. **It's about how the pieces fit together, not the pieces themselves.** Too much time is spent assessing individual components of a gen AI engine. Much more consequential is figuring out how they work together securely.

¹ “McKinsey Technology Trends Outlook 2024,” forthcoming on McKinsey.com.

² “Technology's generational moment with generative AI: A CIO and CTO guide,” McKinsey, July 11, 2023.

Three approaches to using gen AI

There are three primary approaches to take in using gen AI:

- In “Taker” use cases, companies use off-the-shelf, gen AI–powered software from third-party vendors such as GitHub Copilot or Salesforce Einstein to achieve the goals of the use case.
- In “Shaper” use cases, companies integrate bespoke gen AI capabilities by engineering prompts, data sets, and connections to internal systems to achieve the goals of the use case.
- In “Maker” use cases, companies create their own LLMs by building large data sets to pre-train models from scratch. Examples include OpenAI, Anthropic, Cohere, and Mistral AI.

Most companies 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. The highest-value gen AI initiatives, however, generally rely on the Shaper approach.¹

¹ For more on the three approaches, see “Technology's generational moment with generative AI: A CIO and CTO guide,” McKinsey, July 11, 2023.

3. **Get a handle on costs before they sink you.** Models account for only about 15 percent of the overall cost of gen AI applications. Understand where the costs lurk, and apply the right tools and capabilities to rein them in.
4. **Tame the proliferation of tools and tech.** The proliferation of infrastructures, LLMs, and tools has made scaled rollouts unfeasible. Narrow down to those capabilities that best serve the business, and take advantage of available cloud services (while preserving your flexibility).
5. **Create teams that can build value, not just models.** Getting to scale requires a team with a broad cross-section of skills to not only build models but also make sure they generate the value they're supposed to, safely and securely.
6. **Go for the right data, not the perfect data.** Targeting which data matters most and investing in its management over time has a big impact on how quickly you can scale.
7. **Reuse it or lose it.** Reusable code can increase the development speed of generative AI use cases by 30 to 50 percent.

1. Eliminate the noise, and focus on the signal

Although many business leaders acknowledge the need to move past pilots and experiments, that isn't always reflected in what's happening on the ground. Even as gen AI adoption increases, examples of its real bottom-line impact are few and far between. Only 15 percent of companies in our latest AI survey say they are seeing use of gen AI have meaningful impact on their companies' EBIT.³

Exacerbating this issue is that leaders are drawing misleading lessons from their experiments. They try to take what is essentially a chat interface pilot and shift it to an application—the classic “tech looking for a solution” trap. Or a pilot might have

been deemed “successful,” but it was not applied to an important part of the business.

There are many reasons for failing to scale, but the overarching one is that resources and executive focus are spread too thinly across dozens of ongoing gen AI initiatives. This is not a new development. We've seen a similar pattern when other technologies emerged, from cloud to advanced analytics. The lessons from those innovations, however, have not stuck.

The most important decision a CIO will need to make is to eliminate nonperforming pilots and scale up those that are both technically feasible and promise to address areas of the business that matter while minimizing risk (Exhibit 1). The CIO will need to work closely with business unit leaders on setting priorities and handling the technical implications of their choices.

2. It's about how the pieces fit together, not the pieces themselves

In many discussions, we hear technology leaders belaboring decisions around the component parts required to deliver gen AI solutions—LLMs, APIs, and so on. What we are learning, however, is that solving for these individual pieces is relatively easy and integrating them is anything but. This creates a massive roadblock to scaling gen AI.

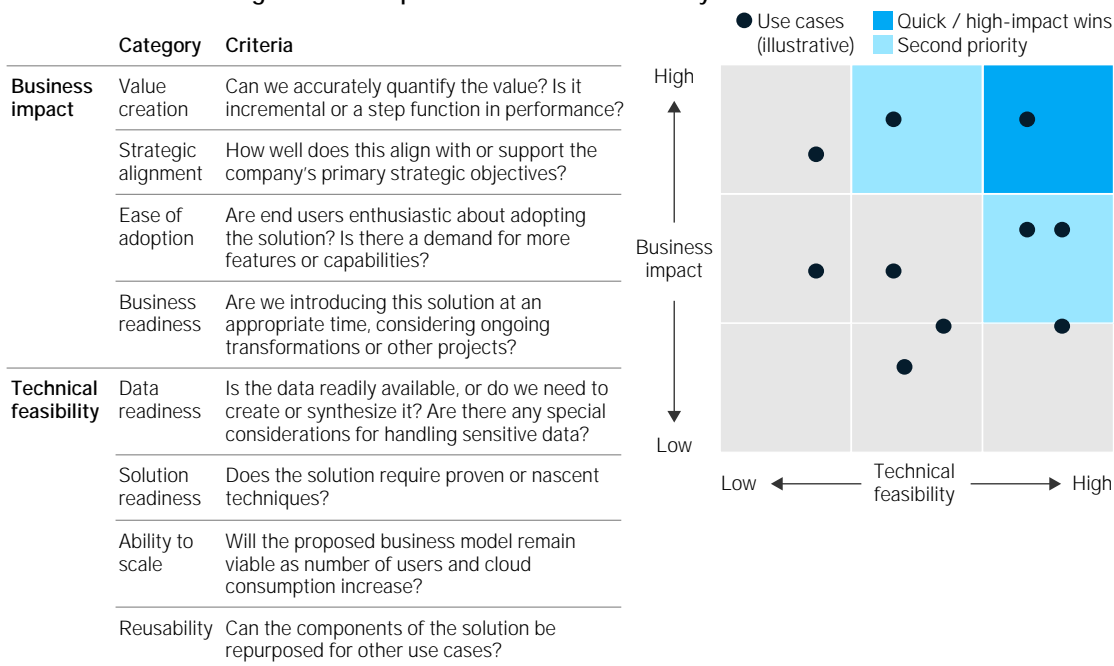
The challenge lies in orchestrating the range of interactions and integrations at scale. Each use case often needs to access multiple models, vector databases, prompt libraries, and applications (Exhibit 2). Companies have to manage a variety of sources (such as applications or databases in the cloud, on-premises, with a vendor, or a combination), the degree of fidelity (including latency and resilience), and existing protocols (for example, access rights). As a new component is added to deliver a solution, it creates a ripple effect on all the other components in the system, adding exponential complexity to the overall solution.

³ That is, they attribute 5 percent or more of their organizations' EBIT to gen AI use. McKinsey Global Survey on the state of AI in early 2024, February 22 to March 5, 2024, forthcoming on McKinsey.com.

Exhibit 1

Focus on use cases that are feasible and where business impact is clear.

Criteria for determining business impact and technical feasibility



McKinsey & Company

The key to effective orchestration is embedding the organization's domain and workflow expertise into the management of the step-by-step flow and sequencing of the model, data, and system interactions of an application running on a cloud foundation. The core component of an effective orchestration engine is an API gateway, which authenticates users, ensures compliance, logs request-and-response pairs (for example, to help bill teams for their usage), and routes requests to the best models, including those offered by third parties. The gateway also enables cost tracking and provides risk and compliance teams a way to monitor usage in a scalable way. This gateway capability is crucial for scale because it allows teams to operate independently while ensuring that they follow best practices (see sidebar "Main components for gen AI model orchestration").

The orchestration of the many interactions required to deliver gen AI capabilities, however, is impossible without effective end-to-end automation. "End-to-end" is the key phrase here. Companies will often automate elements of the workflow, but the value comes only by automating the entire solution, from data wrangling (cleaning and integration) and data pipeline construction to model monitoring and risk review through "policy as code." Our latest research has shown that gen AI high performers are more than three times as likely as their peers to have testing and validation embedded in the release process for each model.⁴ A modern MLOps platform is critical in helping to manage this automated flow and, according to McKinsey analysis, can accelerate production by ten times as well as enable more efficient use of cloud resources.

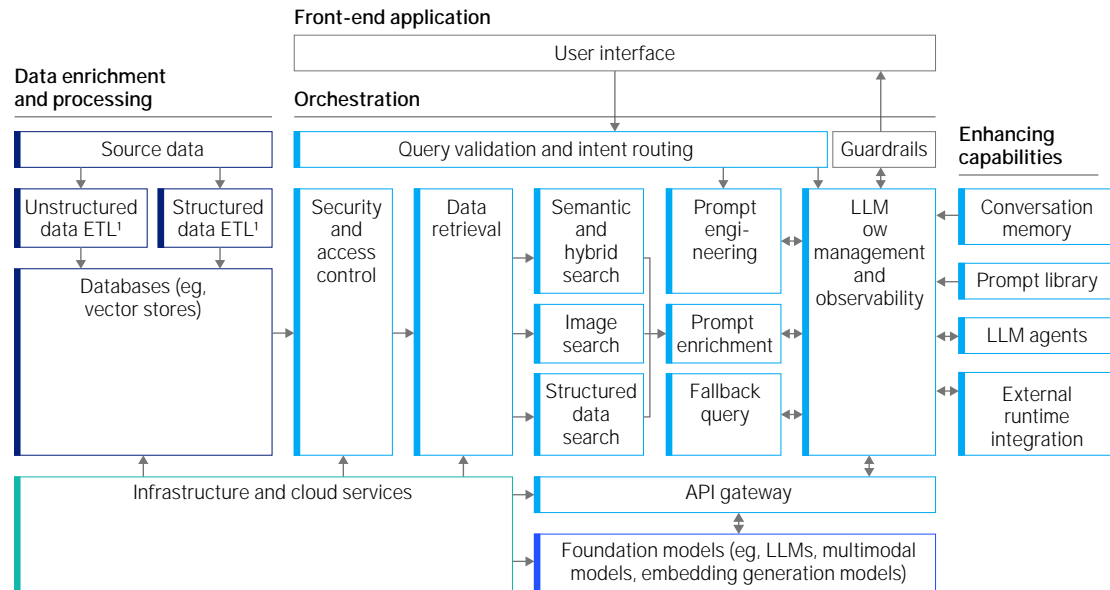
⁴ We define gen AI high performers as those who attribute more than 10 percent of their organizations' EBIT to their use of gen AI. McKinsey Global Survey on the state of AI in early 2024, February 22 to March 5, 2024, forthcoming on McKinsey.com.

Exhibit 2

A gen AI solution needs to accommodate a complex set of integrations across the entire tech stack.

Illustrative tech stack with end-to-end automation

■ Data ■ Gen AI capabilities ■ Cloud ■ Models



¹Extract, transform, load.

McKinsey & Company

Gen AI models can produce inconsistent results, due to their probabilistic nature or the frequent changes to underlying models. Model versions can be updated as often as every week, which means companies can't afford to set up their orchestration capability and let it run in the background. They need to develop hyperattentive observing and triaging capabilities to implement gen AI with speed and safety. Observability tools monitor the gen AI application's interactions with users in real time, tracking metrics such as response time, accuracy, and user satisfaction scores. If an application begins to generate inaccurate or inappropriate responses, the tool alerts the development team to investigate and make any necessary adjustments to the model parameters, prompt templates, or orchestration flow.

3. Get a handle on costs before they sink you

The sheer scale of gen AI data usage and model interactions means costs can quickly spiral out of control. Managing these costs will have a huge impact on whether CIOs can manage gen AI programs at scale. But understanding what drives costs is crucial to gen AI programs. The models themselves, for example, account for only about 15 percent of a typical project effort.⁵ LLM costs have dropped significantly over time and continue to decline.

CIOs should focus their energies on four realities:

- **Change management is the biggest cost.** Our experience has shown that a good rule of thumb for managing gen AI costs is that for every \$1

⁵"Generative AI in the pharmaceutical industry: Moving from hype to reality," McKinsey, January 9, 2024.

Main components for gen AI model orchestration

Orchestration is the process of coordinating various data, transformation, and AI components to manage complex AI workflows. The API (or LLM) gateway layer serves as a secure and efficient interface between users or applications and underlying gen AI models. The orchestration engine itself is made up of the following components:

- **Prompt engineering and prompt library:** Prompt engineering is the process of crafting input prompts or queries that guide the behavior and output of AI models. A prompt library is a collection of predefined prompts that users can leverage as best practices/shortcuts when they invoke a gen AI model.
- **Context management and caching:** Context management highlights background information relevant to a specific task or interaction. Caching relates to storing previously computed results or intermediate data to accelerate future computations.
- **Information retrieval (semantic search and hybrid search):** Information-retrieval logic allows gen AI models to search for and retrieve relevant information from a collection of documents or data sources.
- **Evaluation and guardrails:** Evaluation and guardrail tools help assess the performance, reliability, and ethical considerations of AI models. They also provide input to governance and LLMOps. This encompasses tools and processes for evaluating model accuracy, robustness, fairness, and safety.

spent on developing a model, you need to spend about \$3 for change management. (By way of comparison, for digital solutions, the ratio has tended to be closer to \$1 for development to \$1 for change management.⁶) Discipline in managing the range of change actions, from training your people to role modeling to active performance tracking, is crucial for gen AI. Our analysis has shown that high performers are nearly three times more likely than others to have a strong performance-management infrastructure, such as key performance indicators (KPIs), to measure and track value of gen AI. They are also twice as likely to have trained nontechnical people well enough to understand the potential value and risks associated with using gen AI at work.⁷

Companies have been particularly successful in handling the costs of change management by focusing on two areas: first, involving end users in solution development from day one (too often,

companies default to simply creating a chat interface for a gen AI application), and second, involving their best employees in training models to ensure the models learn correctly and quickly.

- **Run costs are greater than build costs for gen AI applications.** Our analysis shows that it's much more expensive to run models than to build them. Foundation model usage and labor are the biggest drivers of that cost. Most of the labor costs are for model and data pipeline maintenance. In Europe, we are finding that significant costs are also incurred by risk and compliance management.
- **Driving down model costs is an ongoing process.** Decisions related to how to engineer the architecture for gen AI, for example, can lead to cost variances of 10 to 20 times, and sometimes more than that. An array of cost-reduction tools and capabilities are available,

⁶ Eric Lamarre, Kate Smaje, and Rodney Zimmel, "Rewired to outcompete," McKinsey, June 20, 2023.

⁷ McKinsey Global Survey on the state of AI in early 2024, February 22 to March 5, 2024, forthcoming on McKinsey.com.

such as preloading embeddings. This is not a one-off exercise. The process of cost optimization takes time and requires multiple tools, but done well, it can reduce costs from a dollar a query to less than a penny (Exhibit 3).

- **Investments should be tied to ROI.** Not all gen AI interactions need to be treated the same, and they therefore shouldn't all cost the same. A gen AI tool that responds to live questions from customers, for example, is critical to customer experience and requires low-latency rates, which are more expensive. But code documentation tools don't have to be so responsive, so they can be run more cheaply. Cloud plays a crucial role in driving ROI because its prime source of value lies in supporting business growth, especially supporting scaled

analytics solutions. The goal here is to develop a modeling discipline that instills an ROI focus on every gen AI use case without getting lost in endless rounds of analysis.

4. Tame the proliferation of tools and tech

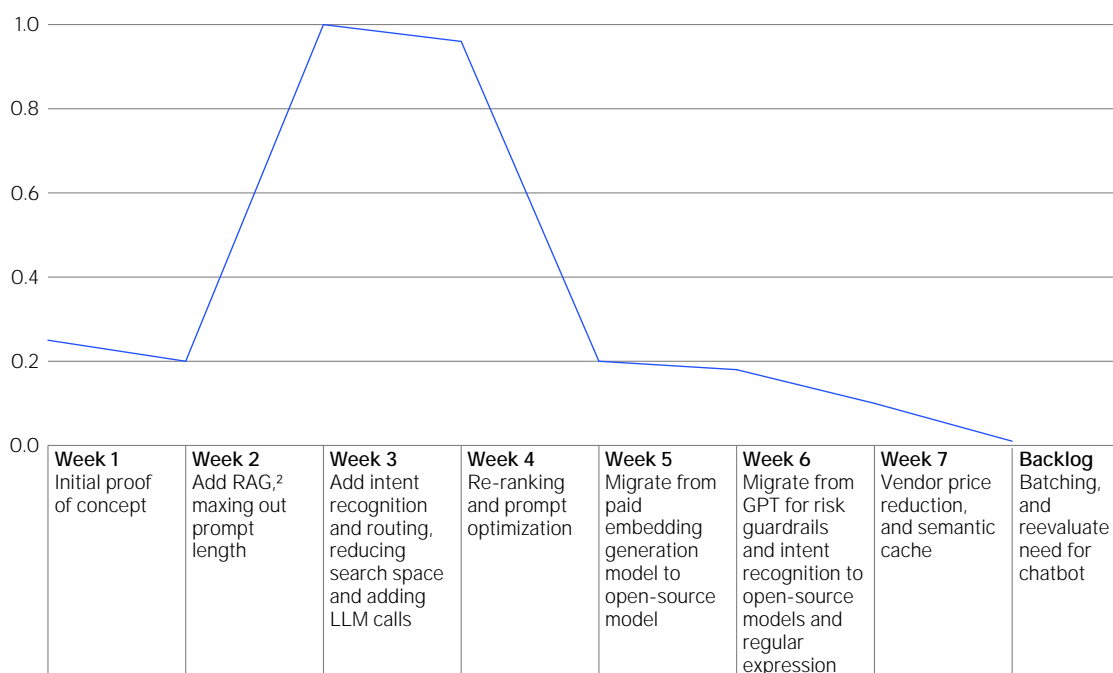
Many teams are still pushing their own use cases and have often set up their own environments, resulting in companies having to support multiple infrastructures, LLMs, tools, and approaches to scaling. In a recent McKinsey survey, in fact, respondents cited “too many platforms” as the top technology obstacle to implementing gen AI at scale.⁸ The more infrastructures and tools, the higher the complexity and cost of operations, which in turn makes scaled rollouts unfeasible. This state

⁸ McKinsey survey on generative AI in operations, November 2023.

Exhibit 3

As solutions scale, organizations can optimize costs.

Cost per query by week,¹ \$



¹ Illustrative example pulling from multiple case studies.

² Retrieval-augmented generation.

of affairs is similar to the early days of cloud and software as a service (SaaS), when accessing the tech was so easy—often requiring no more than a credit card—that a “wild west” of proliferating tools created confusion and risk.

To get to scale, companies need a manageable set of tools and infrastructures. Fair enough—but how do you know which providers, hosts, tools, and models to choose? The key is to not waste time on endless rounds of analysis on decisions that don’t matter much (for example, the choice of LLMs is less critical as they increasingly become a commodity) or where there isn’t much of a choice in the first place—for example, if you have a primary cloud service provider (CSP) that has most of your data and your talent knows how to work with the CSP, you should probably choose that CSP’s gen AI offering. Major CSPs, in fact, are rolling out new gen AI services that can help companies improve the economics of some use cases and open access to new ones. How well companies take advantage of these services depends on many variables, including their own cloud maturity and the strength of their cloud foundations.

What *does* require detailed thinking is how to build your infrastructure and applications in a way that gives you the flexibility to switch providers or models relatively easily. Consider adopting standards widely used by providers (such as KFServing, a serverless solution for deploying gen AI models), Terraform for infrastructure as code, and open-source LLMs.

It’s worth emphasizing that overengineering for flexibility eventually leads to diminishing returns. A plethora of solutions becomes expensive to maintain, making it difficult to take full advantage of the services providers offer.

5. Create teams that can build value, not just models

One of the biggest issues companies are facing is that they’re still treating gen AI as a technology program rather than as a broad business priority. Past technology efforts demonstrate, however, that creating value is never a matter of “just tech.” For gen AI to have real impact, companies have to build teams that can take it beyond the IT function and embed it into the business. Past lessons are applicable here, too. Agile practices sped up technical development,

for example. But greater impact came only when other parts of the organization—such as risk and business experts—were integrated into the teams along with product management and leadership.

There are multiple archetypes for ensuring this broader organizational integration. Some companies have built a center of excellence to act as a clearinghouse to prioritize use cases, allocate resources, and monitor performance. Other companies split strategic and tactical duties among teams. Which archetype makes sense for any given business will depend on its available talent and local realities. But what’s crucial is that this centralized function enables close collaboration between technology, business, and risk leads, and is disciplined in following proven protocols for driving successful programs. Those might include, for example, quarterly business reviews to track initiatives against specific objectives and key results (OKRs), and interventions to resolve issues, reallocate resources, or shut down poor-performing initiatives.

A critical role for this governing structure is to ensure that effective risk protocols are implemented and followed. Build teams, for example, need to map the potential risks associated with each use case; technical and “human-in-the-loop” protocols need to be implemented throughout the use-case life cycle. This oversight body also needs a mandate to manage gen AI risk by assessing exposures and implementing mitigating strategies.

One issue to guard against is simply managing the flow of tactical use cases, especially where the volume is large. This central organization needs a mandate to cluster related use cases to ensure large-scale impact and drive large ideas. This team needs to act as the guardians for value, not just managers of work.

One financial services company put in place clearly defined governance protocols for senior management. A steering group, sponsored by the CIO and chief strategy officer, focused on enterprise governance, strategy, and communication, driving use-case identification and approvals. An enablement group, sponsored by the CTO, focused on decisions around data architecture, data science, data engineering, and building core enabling

capabilities. The CTO also mandated that at least one experienced architect join a use-case team early in their process to ensure the team used the established standards and tool sets. This oversight and governance clarity was crucial in helping the business go from managing just five to more than 50 use cases in its pipeline.

6. Go for the right data, not the perfect data

Misconceptions that gen AI can simply sweep up the necessary data and make sense of it are still widely held. But high-performing gen AI solutions are simply not possible without clean and accurate data, which requires real work and focus. The companies that invest in the data foundations to generate good data aim their efforts carefully.

Take the process of labeling, which often oscillates between seeking perfection for all data and complete neglect. We have found that investing in targeted labeling—particularly for the data used for retrieval-augmented generation (RAG)—can have a significant impact on the quality of answers to gen AI queries. Similarly, it's critical to invest the time to grade the importance of content sources ("authority weighting"), which helps the model understand the relative value of different sources. Getting this right requires significant human oversight from people with relevant expertise.

Because gen AI models are so unstable, companies need to maintain their platforms as new data is added, which happens often and can affect how models perform. This is made vastly more difficult at most companies because related data lives in so many different places. Companies that have invested in creating data products are ahead of the game because they have a well-organized data source to use in training models over time.

At a materials science product company, for example, various teams accessed product information, but each one had a different version. R&D had materials safety sheets, application

engineering teams (tech sales/support teams) developed their own version to find solutions for unique client calls, commercialization teams had product descriptions, and customer support teams had a set of specific product details to answer queries. As each team updated its version of the product information, conflicts emerged, making it difficult for gen AI models to use the data. To address this issue, the company is putting all relevant product information in one place.

7. Reuse it or lose it

Reusable code can increase the development speed of generative AI use cases by 30 to 50 percent.⁹ But in their haste to make meaningful breakthroughs, teams often focus on individual use cases, which sinks any hope for scale. CIOs need to shift the business's energies to building transversal solutions that can serve many use cases. In fact, we have found that gen AI high performers are almost three times as likely as their peers to have gen AI foundations built strategically to enable reuse across solutions.¹⁰

In committing to reusability, however, it is easy to get caught in building abstract gen AI capabilities that don't get used, even though, technically, it would be easy to do so. A more effective way to build up reusable assets is to do a disciplined review of a set of use cases, typically three to five, to ascertain their common needs or functions. Teams can then build these common elements as assets or modules that can be easily reused or strung together to create a new capability. Data preprocessing and ingestion, for example, could include a data-chunking mechanism, a structured data-and-metadata loader, and a data transformer as distinct modules. One European bank reviewed which of its capabilities could be used in a wide array of cases and invested in developing a synthesizer module, a translator module, and a sentiment analysis module.

CIOs can't expect this to happen organically. They need to assign a role, such as the platform owner,

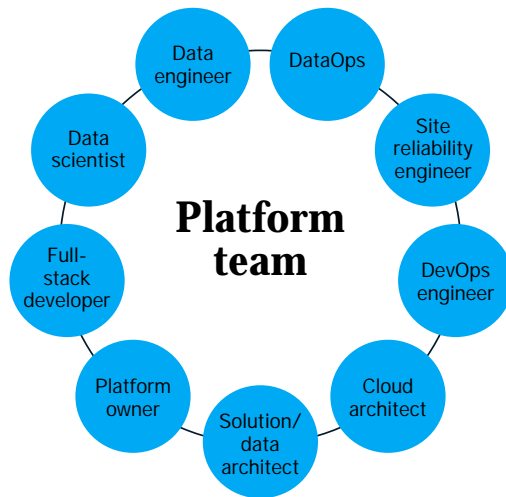
⁹ Eric Lamarre, Alex Singla, Alexander Sukharevsky, and Rodney Zimmel, "A generative AI reset: Rewiring to turn potential into value in 2024," McKinsey, March 4, 2024.

¹⁰ McKinsey Global Survey on the state of AI in early 2024, February 22 to March 5, 2024, forthcoming on McKinsey.com.

Exhibit 4

A gen AI platform team needs an array of skills.

Cross-functional platform team roles and skills



DataOps: Manages and optimizes the data pipeline, ensuring the availability and quality of data; supports training and deployment of gen AI models

Site reliability engineer: Ensures reliability, availability, and performance of software systems and applications

DevOps engineer: Establishes the CI/CD¹ pipeline and other automation needed for teams to rapidly develop and deploy code (eg, chatbot, APIs) to production

Cloud architect: Ensures scalability, security, and cost optimization of the cloud infrastructure; designs data storage and management systems; facilitates integration and deployment of the AI models

Solution/data architect: Develops creative and efficient solutions using engineering practices and software/web development technologies

Platform owner: Acts like a product owner, oversees the build of a gen AI platform

Full-stack developer: Writes clean and quality scalable code (eg, front-end/back-end APIs) that can be easily deployed with CI/CD¹ pipelines

Data scientist: Fine-tunes foundational models to help RAG²-based approach, ensures alignment of LLM outputs with responsible AI guidelines

Data engineer: Architects data models to ingest data into vector databases, creates and maintains automated pipelines, performs closed-loop testing to validate responses and improve performance

¹Continuous integration (CI) and continuous delivery (CD).

²Retrieval-augmented generation.

McKinsey & Company

and a cross-functional team with a mandate to develop reusable assets for product teams (Exhibit 4), which can include approved tools, code, and frameworks.

The value gen AI could generate is transformational. But capturing the full extent of that value will come only when companies harness gen AI at scale. That requires CIOs to not just acknowledge hard truths but be ready to act on them to lead their business forward.

Aamer Baig is a senior partner in McKinsey's Chicago office, **Douglas Merrill** is a partner in the Southern California office, **Megha Sinha** is a partner in the Bay Area office, **Danesh Mead** is a consultant in the Denver office, and **Stephen Xu** is director of product management in the Toronto office.

The authors wish to thank Mani Gopalakrishnan, Mark Gu, Ankur Jain, Rahil Jogani, and Asin Tavakoli for their contributions to this article.