March 2024
McKinsey Quarterly

A generative AI reset: Rewiring to turn potential into value in 2024

The generative AI payoff may only come when companies do deeper organizational surgery on their business.

by Eric Lamarre, Alex Singla, Alexander Sukharevsky, and Rodney Zemmel

It's time for a generative AI (gen AI) reset. The initial enthusiasm and flurry of activity in 2023 is giving way to second thoughts and recalibrations as companies realize that capturing gen AI's enormous potential value is harder than expected.

With 2024 shaping up to be the year for gen AI to prove its value, companies should keep in mind the hard lessons learned with digital and AI transformations: competitive advantage comes from building organizational and technological capabilities to broadly innovate, deploy, and improve solutions at scale—in effect, rewiring the business for distributed digital and AI innovation.

Companies looking to score early wins with gen AI should move quickly. But those hoping that gen AI offers a shortcut past the tough—and necessary—organizational surgery are likely to meet with disappointing results. Launching pilots is (relatively) easy; getting pilots to scale and create meaningful value is hard because they require a broad set of changes to the way work actually gets done.

Let's briefly look at what this has meant for one Pacific region telecommunications company. The company hired a chief data and AI officer with a mandate to "enable the organization to create value with data and AI." The chief data and AI officer worked with the business to develop the strategic vision and implement the road map for the use cases. After a scan of domains (that is, customer journeys or functions) and use case opportunities across the enterprise, leadership prioritized the home-servicing/maintenance domain to pilot and then scale as part of a larger sequencing of initiatives. They targeted, in particular, the development of a gen AI tool to help dispatchers and service operators better predict the types of calls and parts needed when servicing homes.

Leadership put in place cross-functional product teams with shared objectives and incentives to build the gen AI tool. As part of an effort to upskill the entire enterprise to better work with data and gen AI tools, they also set up a data and AI academy, which the dispatchers and service operators enrolled in as part of their training. To provide the technology and data underpinnings for gen AI, the chief data and AI officer also selected a large language model (LLM) and cloud provider that could meet the needs of the domain as well as serve other parts of the enterprise. The chief data and AI officer also oversaw the implementation of a data architecture so that the clean and reliable data (including service histories and inventory databases) needed to build the gen AI tool could be delivered quickly and responsibly.

Our book *Rewired: The McKinsey Guide to Outcompeting in the Age of Digital and AI* (Wiley, June 2023) provides a detailed manual on the six capabilities needed to deliver the kind of broad change that harnesses digital and AI technology. In this article, we will explore how to extend each of those capabilities to implement a successful gen AI program at scale. While recognizing that these are still early days and that there is much more to learn, our experience has shown that breaking open the gen AI opportunity requires companies to rewire how they work in the following ways.

Figure out where gen AI copilots can give you a real competitive advantage

The broad excitement around gen AI and its relative ease of use has led to a burst of experimentation across organizations. Most of these initiatives, however, won't generate a competitive advantage. One bank, for example, bought tens of thousands of GitHub Copilot licenses, but since it didn't have a clear sense of how to work with the technology, progress was slow. Another unfocused effort we often see is when companies move to incorporate gen AI into their customer service capabilities. Customer service is a commodity capability, not part of the core business, for most companies. While gen AI might help with productivity in such cases, it won't create a competitive advantage.

To create competitive advantage, companies should first understand the difference between being a "taker" (a user of available tools, often via APIs and subscription services), a "shaper" (an integrator of available models with proprietary data), and a "maker" (a builder of LLMs). For now, the maker approach is too expensive for most companies, so the sweet spot for businesses is implementing a taker model for productivity improvements while building shaper applications for competitive advantage.

Much of gen AI's near-term value is closely tied to its ability to help people do their current jobs better. In this way, gen AI tools act as copilots that work side by side with an employee, creating an initial block of code that a developer can adapt, for example, or drafting a requisition order for a new part that a maintenance worker in the field can review and submit (see sidebar "Copilot examples across three generative AI archetypes"). This means companies should be focusing on where copilot technology can have the biggest impact on their priority programs.

Copilot examples across three generative AI archetypes

- "Taker" copilots help real estate customers sift through property options and find the most promising one, write code for a developer, and summarize investor transcripts.
- "Shaper" copilots provide recommendations to sales reps for upselling customers by connecting generative AI tools to customer relationship management systems, financial systems, and customer behavior histories; create virtual assistants to personalize treatments for patients; and recommend solutions for maintenance workers based on historical data.
- "Maker" copilots are foundation models that lab scientists at pharmaceutical companies can use to find and test new and better drugs more quickly.

Some industrial companies, for example, have identified maintenance as a critical domain for their business. Reviewing maintenance reports and spending time with workers on the front lines can help determine where a gen AI copilot could make a big difference, such as in identifying issues with equipment failures quickly and early on. A gen AI copilot can also help identify root causes of truck breakdowns and recommend resolutions much more quickly than usual, as well as act as an ongoing source for best practices or standard operating procedures.

The challenge with copilots is figuring out how to generate revenue from increased productivity. In the case of customer service centers, for example, companies can stop recruiting new agents and use attrition to potentially achieve real financial gains. Defining the plans for how to generate revenue from the increased productivity up front, therefore, is crucial to capturing the value.

Upskill the talent you have but be clear about the gen-AI-specific skills you need

By now, most companies have a decent understanding of the technical gen AI skills they need, such as model fine-tuning, vector database administration, prompt engineering, and context engineering. In many cases, these are skills that you can train your existing workforce to develop. Those with existing AI and machine learning (ML) capabilities have a strong head start. Data engineers, for example, can learn multimodal processing and vector database management, MLOps (ML operations) engineers can extend their skills to LLMOps (LLM operations), and data scientists can develop prompt engineering, bias detection, and fine-tuning skills.

The learning process can take two to three months to get to a decent level of competence because of the complexities in learning what various LLMs can and can't do and how best to use them. The coders need to gain experience building software, testing, and validating

answers, for example. It took one financial-services company three months to train its best data scientists to a high level of competence. While courses and documentation are available—many LLM providers have boot camps for developers—we have found that the most effective way to build capabilities at scale is through apprenticeship, training people to then train others, and building communities of practitioners. Rotating experts through teams to train others, scheduling regular sessions for people to share learnings, and hosting biweekly documentation review sessions are practices that have proven successful in building communities of practitioners (see sidebar "A sample of new generative AI skills needed").

It's important to bear in mind that successful gen AI skills are about more than coding proficiency. Our experience in developing our own gen AI platform, Lilli, showed us that the best gen AI technical talent has design skills to uncover where to focus solutions, contextual understanding to ensure the most relevant and high-quality answers are generated, collaboration skills to work well with knowledge experts (to test and validate answers and develop an appropriate curation approach), strong forensic skills to figure out causes of breakdowns (is the issue the data, the interpretation of the user's intent, the quality of metadata on embeddings, or something else?), and anticipation skills to conceive of and plan for possible outcomes and to put the right kind of tracking into their code. A pure coder who doesn't intrinsically have these skills may not be as useful a team member.

While current upskilling is largely based on a "learn on the job" approach, we see a rapid market emerging for people who have learned these skills over the past year. That skill growth is moving quickly. GitHub reported that developers were working on gen AI projects "in big numbers," and that 65,000 public gen AI projects were created on its platform in 2023—a jump of almost 250 percent over the previous year. If your company is just starting its gen AI journey, you could consider hiring two or three senior engineers who have built a gen AI shaper product for their companies. This could greatly accelerate your efforts.

Form a centralized team to establish standards that enable responsible scaling

To ensure that all parts of the business can scale gen AI capabilities, centralizing competencies is a natural first move. The critical focus for this central team will be to develop and put in place protocols and standards to support scale, ensuring that teams can access models while also minimizing risk and containing costs. The team's work could include, for example, procuring models and prescribing ways to access them, developing standards for data readiness, setting up approved prompt libraries, and allocating resources.

While developing Lilli, our team had its mind on scale when it created an open plug-in architecture and setting standards for how APIs should function and be built. They developed standardized tooling and infrastructure where teams could securely experiment and access a GPT LLM, a gateway with preapproved APIs that teams could access, and a self-serve developer portal. Our goal is that this approach, over time, can

A sample of new generative AI skills needed

The following are examples of new skills needed for the successful deployment of generative AI tools:

- data scientist:
  - prompt engineering
  - in-context learning
  - bias detection
  - pattern identification
  - reinforcement learning from human feedback
  - hyperparameter/large language model fine-tuning; transfer learning

- data engineer:
  - data wrangling and data warehousing
  - data pipeline construction
  - multimodal processing
  - vector database management

help shift "Lilli as a product" (that a handful of teams use to build specific solutions) to "Lilli as a platform" (that teams across the enterprise can access to build other products).

For teams developing gen AI solutions, squad composition will be similar to AI teams but with data engineers and data scientists with gen AI experience and more contributors from risk management, compliance, and legal functions. The general idea of staffing squads with resources that are federated from the different expertise areas will not change, but the skill composition of a gen-AI-intensive squad will.

Set up the technology architecture to scale

Building a gen AI model is often relatively straightforward, but making it fully operational at scale is a different matter entirely. We've seen engineers build a basic chatbot in a week, but releasing a stable, accurate, and compliant version that scales can take four months. That's why, our experience shows, the actual model costs may be less than 10 to 15 percent of the total costs of the solution.

Building for scale doesn't mean building a new technology architecture. But it does mean focusing on a few core decisions that simplify and speed up processes without breaking the bank. Three such decisions stand out:

- *Focus on reusing your technology.* Reusing code can increase the development speed of gen AI use cases by 30 to 50 percent. One good approach is simply creating a source for approved tools, code, and components. A financial-services company, for example, created a library of production-grade tools, which had been approved by both the security and legal teams, and made them available in a library for teams to use. More important is taking the time to identify and build those capabilities that are common across the most priority use cases. The same financial-services company, for example, identified three components that could be reused for more than 100 identified use cases. By building those first, they were able to generate a significant portion of the code base for all the identified use cases—essentially giving every application a big head start.

- *Focus the architecture on enabling efficient connections between gen AI models and internal systems.* For gen AI models to work effectively in the shaper archetype, they need access to a business's data and applications. Advances in integration and orchestration frameworks have significantly reduced the effort required to make those connections. But laying out what those integrations are and how to enable them is critical to ensure these models work efficiently and to avoid the complexity that creates technical debt (the "tax" a company pays in terms of time and resources needed to redress existing technology issues). Chief information officers and chief technology officers can define reference architectures and integration standards for their organizations. Key elements should include a model hub, which contains trained and approved models that can be provisioned on demand; standard APIs that act as bridges connecting gen AI models to applications or data; and context management and caching, which speed up processing by providing models with relevant information from enterprise data sources.

- *Build up your testing and quality assurance capabilities.* Our own experience building Lilli taught us to prioritize testing over development. Our team invested in not only developing testing protocols for each stage of development but also aligning the entire team so that, for example, it was clear who specifically needed to sign off on each stage of the process. This slowed down initial development but sped up the overall delivery pace and quality by cutting back on errors and the time needed to fix mistakes.

Ensure data quality and focus on unstructured data to fuel your models

The ability of a business to generate and scale value from gen AI models will depend on how well it takes advantage of its own data. As with technology, targeted upgrades to existing data architecture are needed to maximize the future strategic benefits of gen AI:

- *Be targeted in ramping up your data quality and data augmentation efforts.* While data quality has always been an important issue, the scale and scope of data that gen AI models can use—especially unstructured data—has made this issue much more consequential. For this reason, it's critical to get the data foundations right, from clarifying decision rights to defining clear data processes to establishing taxonomies so models can access the data they need. The companies that do this well tie their data quality and augmentation efforts to the specific AI/gen AI application and use case—you don't need this data foundation to extend to every corner of the enterprise. This could mean, for example, developing a new data repository for all equipment specifications and reported issues to better support maintenance copilot applications.

- *Understand what value is locked into your unstructured data.* Most organizations have traditionally focused their data efforts on structured data (values that can be organized in tables, such as prices and features). But the real value from LLMs comes from their ability to work with unstructured data (for example, PowerPoint slides, videos, and text). Companies can map out which unstructured data sources are most valuable and establish metadata tagging standards so models can process the data and teams can

find what they need (tagging is particularly important to help companies remove data from models as well, if necessary). Be creative in thinking about data opportunities. Some companies, for example, are interviewing senior employees as they retire and feeding that captured institutional knowledge into an LLM to help improve their copilot performance.

- *Optimize to lower costs at scale.* There is often as much as a tenfold difference between what companies pay for data and what they could be paying if they optimized their data infrastructure and underlying costs. This issue often stems from companies scaling their proofs of concept without optimizing their data approach. Two costs generally stand out. One is storage costs arising from companies uploading terabytes of data into the cloud and wanting that data available 24/7. In practice, companies rarely need more than 10 percent of their data to have that level of availability, and accessing the rest over a 24- or 48-hour period is a much cheaper option. The other costs relate to computation with models that require on-call access to thousands of processors to run. This is especially the case when companies are building their own models (the maker archetype) but also when they are using pretrained models and running them with their own data and use cases (the shaper archetype). Companies could take a close look at how they can optimize computation costs on cloud platforms—for instance, putting some models in a queue to run when processors aren't being used (such as when Americans go to bed and consumption of computing services like Netflix decreases) is a much cheaper option.

Build trust and reusability to drive adoption and scale

Because many people have concerns about gen AI, the bar on explaining how these tools work is much higher than for most solutions. People who use the tools want to know how they work, not just what they do. So it's important to invest extra time and money to build trust by ensuring model accuracy and making it easy to check answers.

One insurance company, for example, created a gen AI tool to help manage claims. As part of the tool, it listed all the guardrails that had been put in place, and for each answer provided a link to the sentence or page of the relevant policy documents. The company also used an LLM to generate many variations of the same question to ensure answer consistency. These steps, among others, were critical to helping end users build trust in the tool.

Part of the training for maintenance teams using a gen AI tool should be to help them understand the limitations of models and how best to get the right answers. That includes teaching workers strategies to get to the best answer as fast as possible by starting with broad questions then narrowing them down. This provides the model with more context, and it also helps remove any bias of the people who might think they know the answer already. Having model interfaces that look and feel the same as existing tools also helps users feel less pressured to learn something new each time a new application is introduced.

Getting to scale means that businesses will need to stop building one-off solutions that are hard to use for other similar use cases. One global energy and materials company, for

example, has established ease of reuse as a key requirement for all gen AI models, and has found in early iterations that 50 to 60 percent of its components can be reused. This means setting standards for developing gen AI assets (for example, prompts and context) that can be easily reused for other cases.

While many of the risk issues relating to gen AI are evolutions of discussions that were already brewing—for instance, data privacy, security, bias risk, job displacement, and intellectual property protection—gen AI has greatly expanded that risk landscape. Just 21 percent of companies reporting AI adoption say they have established policies governing employees' use of gen AI technologies.

Similarly, a set of tests for AI/gen AI solutions should be established to demonstrate that data privacy, debiasing, and intellectual property protection are respected. Some organizations, in fact, are proposing to release models accompanied with documentation that details their performance characteristics. Documenting your decisions and rationales can be particularly helpful in conversations with regulators.

In some ways, this article is premature—so much is changing that we'll likely have a profoundly different understanding of gen AI and its capabilities in a year's time. But the core truths of finding value and driving change will still apply. How well companies have learned those lessons may largely determine how successful they'll be in capturing that value. Q

Eric Lamarre is a senior partner in McKinsey's Boston office, **Alex Singla** is a senior partner in the Chicago office, **Alexander Sukharevsky** is a senior partner in the London office, and **Rodney Zemmel** is a senior partner in the New York office.

The authors wish to thank Michael Chui, Juan Couto, Ben Ellencweig, Josh Gartner, Bryce Hall, Holger Harreis, Phil Hudelson, Suzana Iacob, Sid Kamath, Neerav Kingsland, Kitti Lakner, Robert Levin, Matej Macak, Lapo Mori, Alex Peluffo, Aldo Rosales, Erik Roth, Abdul Wahab Shaikh, and Stephen Xu for their contributions to this article.

Designed by McKinsey Global Publishing