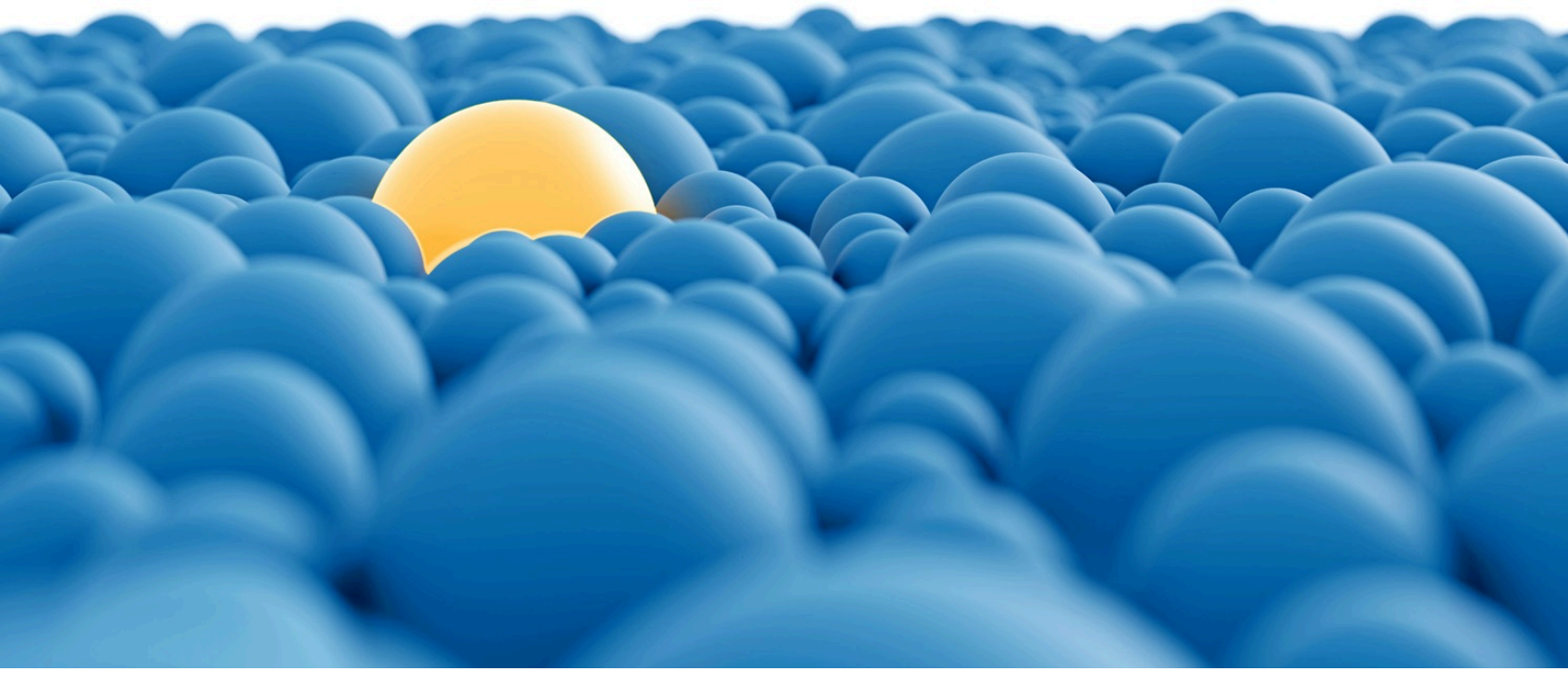
QuantumBlack
AI by McKinsey

One year of agentic AI: Six lessons from the people doing the work

Deploying agentic AI successfully isn't easy. Here's what we're learning about how to get it right.

by Lareina Yee, Michael Chui, and Roger Roberts
with Stephen Xu



September 2025

A year into the agentic AI revolution, one lesson is clear: It takes hard work to do it well.

An agentic enterprise transformation holds the promise of unmatched productivity. While some companies are enjoying early successes with such activities, many more are finding it challenging to see value from their investments. In some cases, they are even retrenching—rehiring people where agents have failed.

These stumbles are a natural evolution of any new technology, and we've seen this pattern before with other innovations. To understand the early lessons, we recently dug into more than 50 agentic AI builds we've led at McKinsey, as well as dozens of others in the marketplace. We've boiled down our analysis results to six lessons to help leaders successfully capture value from agentic AI (see sidebar "What is agentic AI?").

It's not about the agent; it's about the workflow

Achieving business value with agentic AI requires changing workflows. Often, however, organizations focus too much on the agent or the agentic tool. This inevitably leads to great-looking agents that don't actually end up improving the overall workflow, resulting in underwhelming value.

Agentic AI efforts that focus on fundamentally reimagining entire workflows—that is, the steps that involve people, processes, and technology—are more likely to deliver a positive outcome. Understanding how agents can help with each of these steps is the path to value. People will still be central to getting the work done, but now with different agents, tools, and automations to support them.

What is agentic AI?

Agentic AI is a system based on gen AI foundation models that can act in the real world and execute multistep processes. AI agents can automate and perform complex tasks, often using natural language processing, that would normally require human effort (for more, see "What is an AI agent?," from our *McKinsey Explainers* series).

An important starting point in redesigning workflows is mapping processes and identifying key user pain points. This step is critical in designing agentic systems that reduce unnecessary work and allow agents and people to collaborate and accomplish business goals more efficiently and effectively. That collaboration can happen through learning loops and feedback mechanisms, creating a self-reinforcing system. The more frequently agents are used, the smarter and more aligned they become.

Consider an alternative-legal-services provider that was working to modernize its contract review workflows. Legal reasoning in the company's domain was constantly evolving, with new case law, jurisdictional nuances, and policy interpretations, making it challenging to codify expertise.

To account for natural variance, the team designed its agentic systems to learn within the workflow. Every user edit in the document editor, for example, was logged and categorized. This provided the engineers and data scientists with a rich stream of feedback, which they could then use to teach the agents, adjust prompt logic, and enrich the knowledge base. Over time, the agents could codify new expertise.

Focusing on the workflow instead of the agent enabled teams to deploy the right technology at the right point, which is especially important when reengineering complex, multistep workflows (exhibit). For example, insurance companies often have big investigative workflows that span multiple steps (such as claims handling and underwriting), with each step requiring different types of activities and cognitive tasks. Companies can redesign these types of workflows by thoughtfully deploying a targeted mix of rule-based systems, analytical AI, gen AI, and agents, all underpinned by a common orchestration framework (such as open-source frameworks like AutoGen, CrewAI, and LangGraph). In these cases, the agents are the orchestrators and the integrators, accessing tools and integrating outputs of other systems into their context. They are the glue that unifies the workflow so it delivers real closure with less intervention needed.

Agents aren't always the answer

AI agents can do a lot, but they shouldn't necessarily be used for everything. Too often, leaders don't look closely enough at the work that needs to be done or ask whether an agent would be the best choice to perform that work.

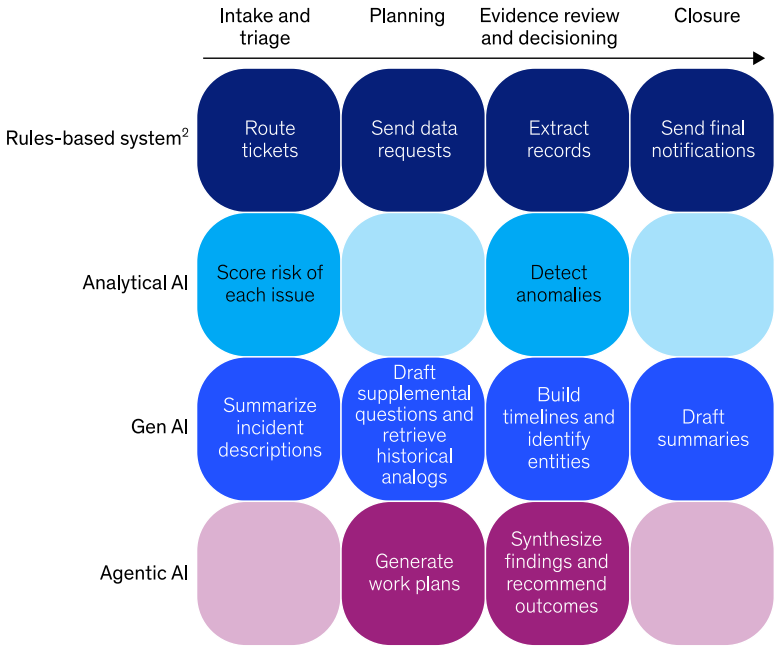
To help avoid wasted investments or unwanted complexity, business leaders can approach the role of agents much like they do when evaluating people for a high-performing team. The key question to ask is, "What is the work to be done and what are the relative talents of each potential team member—or agent—to work together to achieve those goals?" Business problems can often be addressed with simpler automation approaches, such as rules-based automation, predictive analytics, or large language model (LLM) prompting, which can be more reliable than agents out of the box.

Before rushing into an agentic solution, business leaders should take stock of the task's demands. In practice, that means getting clear on how standardized the process should be, how

Exhibit 1

Complex workflows should incorporate the best tool for each task.

Example of investigative workflow[1] by tool used for each task in the workflow



[1]For example, claims handling.
[2]For example, robotic process automation.

McKinsey & Company

much variance the process needs to handle, and what portions of the work agents are best suited to do.

On one level, these issues are straightforward. For example, low-variance, high-standardization workflows, such as investor onboarding or regulatory disclosures, tend to be tightly governed and follow predictable logic. In these cases, agents based on nondeterministic LLMs could add more complexity and uncertainty than value.

By contrast, high-variance, low-standardization workflows could benefit significantly from agents. For example, agents were deployed at a financial-services company to extract complex financial information, reducing the amount of human validation required and streamlining workflows. These tasks demanded information aggregation, verification checks, and compliance analysis—tasks where agents can be effective.

The important thing to remember is not to get trapped in a binary "agent/no agent" mindset. Some agents can do specific tasks well, others can help people do their work better, and in many cases, different technologies altogether might be more appropriate. The key is to figure out which tool or agent is best suited to the task, how people can work with them most effectively, and how agents and workers should be combined to deliver the greatest output. How well people, agents, and tools work together is the secret sauce for value (see sidebar "High-level rules of thumb when considering what AI tools to use").

Stop 'AI slop': Invest in evaluations and build trust with users

One of the most common pitfalls teams encounter when deploying AI agents is agentic systems that seem impressive in demos but frustrate users who are actually responsible for the work. It's common to hear users complain about "AI slop" or low-quality outputs. Users quickly lose trust in the agents, and adoption levels are poor. Any efficiency gains achieved through automation can easily be offset by a loss in trust or a decline in quality.

A hard-won lesson of this recurring problem is that companies should invest heavily in agent development, just like they do for employee development. As one business leader told us, "Onboarding agents is more like hiring a new employee versus deploying software." Agents should be given clear job descriptions, onboarded, and given continual feedback so they become more effective and improve regularly.

High-level rules of thumb when considering what AI tools to use

When deciding which AI tool to use for different tasks, the following guidelines can help:

- If the task is rule based and repetitive, with structured input (say, data entry), use rule-based automation.
- If the input is unstructured (for example, lengthy documents), but the task is still extractive or generative, use gen AI, natural language processing, or predictive analytics.
- If the task involves classification or forecasting from past data, use predictive analytics or gen AI.
- If the output requires synthesis, judgment, or creative interpretation, use gen AI.
- If the task involves multistep decision-making and has a long tail of highly variable inputs and contexts, use AI agents.

Developing effective agents is challenging work that requires harnessing individual expertise to create evaluations (or "evals") and codifying best practices with sufficient granularity for given tasks. This codification serves as both the training manual and performance test for the agent, ensuring that it performs as expected.

These practices may exist in standard operating procedures or as tacit knowledge in people's heads. When codifying practices, it's important to focus on what separates top performers from the rest. For sales reps, this might include how they drive the conversation, handle objections, and match the customer's style (see sidebar "Eval types").

Eval types

These are some typical evaluations used to assess agent performance:

- *Task success rate (end to end).* The task success rate measures the percentage of workflows completed correctly without escalation or human intervention, reflecting real-world utility.
- *F1 score/precision and recall.* This metric balances false positives and false negatives, making it useful for classification, extraction, and decision accuracy tasks where there is a clear measurable outcome (that is, yes or no).
- *Retrieval accuracy.* Retrieval accuracy is the percentage of correct documents, facts, or evidence retrieved relative to the ground truth set, which is critical for retrieval-augmented workflows.
- *Semantic similarity.* Semantic similarity is measured using embedding-based cosine similarity between generated output and reference output, capturing meaning alignment beyond exact word matching.
- *LLM as judge.* Using a large language model (LLM) as a judge involves evaluating outputs against gold standards or human preferences. This metric scales well for subjective judgments such as clarity, helpfulness, and reasoning soundness.
- *Bias detection (via confusion matrices).* Bias detection measures systematic differences in outcomes across user groups using confusion matrices, which highlight where bias manifests (for example, false negatives disproportionately affecting one group).
- *Hallucination rate.* This metric tracks the frequency of factually incorrect or unsupported claims, ensuring the trustworthiness of agent outputs.
- *Calibration error (confidence versus accuracy).* Calibration error measures whether the agent's confidence scores align with actual correctness, which is important for risk-sensitive workflows.

Crucially, experts should stay involved to test agents' performance over time; there can be no "launch and leave" in this arena. This commitment to evaluation requires, for example, experts to literally write down or label desired (and perhaps undesired) outputs for given inputs, which can sometimes number in the thousands for more complex agents. In this way, teams can evaluate how much an agent got right or wrong and make the necessary corrections.

A global bank took this approach to heart when transforming its know-your-customer and credit-risk-analysis processes. Whenever the agent's recommendation on compliance with intake guidelines differed from human judgment, the team identified the logic gaps, refined the decision criteria, and reran tests.

In one case, for example, the agents' initial analysis was too general. The team provided that feedback, then developed and deployed additional agents to ensure that the depth of analysis provided useful insights at the right level of granularity. One way they did this was by asking the agents "why" in multiple succession. This approach ensured the agents performed well, making it much more likely for people to accept their outputs.

Make it easy to track and verify every step

When working with only a few AI agents, reviewing their work and spotting errors can be mostly straightforward. But as companies roll out hundreds, or even thousands, of agents, the task becomes challenging. Exacerbating the issue is that many companies track only outcomes. So when there's a mistake—and there will always be mistakes as companies scale agents—it's hard to figure out precisely what went wrong.

Agent performance should be verified at each step of the workflow. Building monitoring and evaluation into the workflow can enable teams to catch mistakes early, refine the logic, and continually improve performance, even after the agents are deployed.

In one document review workflow, for instance, an alternative-legal-services provider's product team observed a sudden drop in accuracy when the system encountered a new set of cases. But since they'd built the agentic workflow with observability tools to track every step of the process, the team quickly identified the issue: Certain user segments were submitting lower-quality data, leading to incorrect interpretations and poor downstream recommendations.

With that insight, the team improved its data collection practices, provided document formatting guidelines to upstream stakeholders, and adjusted the system's parsing logic. Agent performance quickly rebounded.

The best use case is the reuse case

In the rush to make progress with agentic AI, companies often create a unique agent for each identified task. This can lead to significant redundancy and waste because the same agent can

often accomplish different tasks that share many of the same actions (such as ingesting, extracting, searching, and analyzing).

Deciding how much to invest in building reusable agents (versus an agent that executes one specific task) is analogous to the classic IT architecture problem where companies need to build fast but not lock in choices that constrain future capabilities. How to strike that balance often requires a lot of judgment and analysis.

Identifying recurring tasks is a good starting point. Companies can develop agents and agent components that can easily be reused across different workflows, and make it simple for developers to access them. That includes developing a centralized set of validated services (such as LLM observability or preapproved prompts) and assets (for example, application patterns, reusable code, and training materials) that are easy to locate and use. Integrating these capabilities into a single platform is critical. In our experience, this helps to virtually eliminate 30 to 50 percent of the nonessential work typically required.

Humans remain essential, but their roles and numbers will change

As AI agents continue to proliferate, the question of what role humans will play has generated much anxiety—about job security, on the one hand, and about high expectations for productivity increases, on the other. This has led to wildly diverging views on the role of humans in many present-day jobs.

To be clear: Agents will be able to accomplish a lot, but humans will remain an essential part of the workforce equation even as the type of work that both agents and humans do changes over time. People will need to oversee model accuracy, ensure compliance, use judgment, and handle edge cases, for example. And as we discussed earlier, agents will not always be the best answer, so people working with other tools such as machine learning models will be needed. The number of people working in a particular workflow, however, will likely change and often will be lower once the workflow is transformed using agents. Business leaders will crucially have to manage these transitions as they would for any change program and thoughtfully allocate the work necessary to train and evaluate agents.

Another big lesson from our experience is that companies should be deliberate in redesigning work so that people and agents can collaborate well together. Without that focus, even the most advanced agentic programs risk silent failures, compounding errors, and user rejection.

Take the example of the alternative-legal-services provider mentioned earlier that wanted to use agents for a legal-analysis workflow. In designing the workflow, the team took the time to identify where, when, and how to integrate human input. In one case, agents were able to organize core claims and dollar amounts with high levels of accuracy, but it was important for lawyers to double-check and approve them, given how central the claims were to the entire case.

Similarly, agents were able to recommend workplan approaches to a case, but given the importance of the decision, it was critical for people to not just review but also adjust the recommendation. The agents were also programmed to highlight edge cases and anomalies, helping lawyers develop more comprehensive views. Someone still had to sign the document at the end of the process, underwriting the legal decision with the person's license and credentials.

An important part of this human–agent collaborative design is developing simple visual user interfaces that make it easy for people to interact with agents. For example, one property and casualty insurance company developed interactive visual elements (such as bounding boxes, highlights, and automated scrolling) to help reviewers quickly validate AI-generated summaries. When people clicked on an insight, for example, the application would scroll directly to the correct page and highlight the appropriate text. This focus on the user experience saved time, reduced second-guessing, and built confidence in the system, leading to user acceptance levels near 95 percent.

The world of AI agents is moving quickly, so we can expect to learn many more lessons. But unless companies approach their agentic programs with learning in mind (and in practice), they're likely to repeat mistakes and slow their progress.

Lareina Yee is a McKinsey Global Institute director and a senior partner in McKinsey's Bay Area office, where **Michael Chui** is a senior fellow and **Roger Roberts** is a partner; **Stephen Xu** is a senior director of product management in the Toronto office.

The authors wish to thank Alex Singla, Alexander Sukharevsky, Alberto Mario Pirovano, Allen Chen, Ani Aghababyan, Antonio Castro, Carlo Giovine, Medha Bankhwal, Rickard Ström, and the entire product team at QuantumBlack Labs, McKinsey's center dedicated to driving innovation and experimentation in AI, for their contributions to this article.

This article was edited by Barr Seitz, an editorial director in the New York office.

To request a demo or follow-up with an expert in QuantumBlack Labs, our software development and R&D hub, please reach out to helloqb@mckinsey.com.