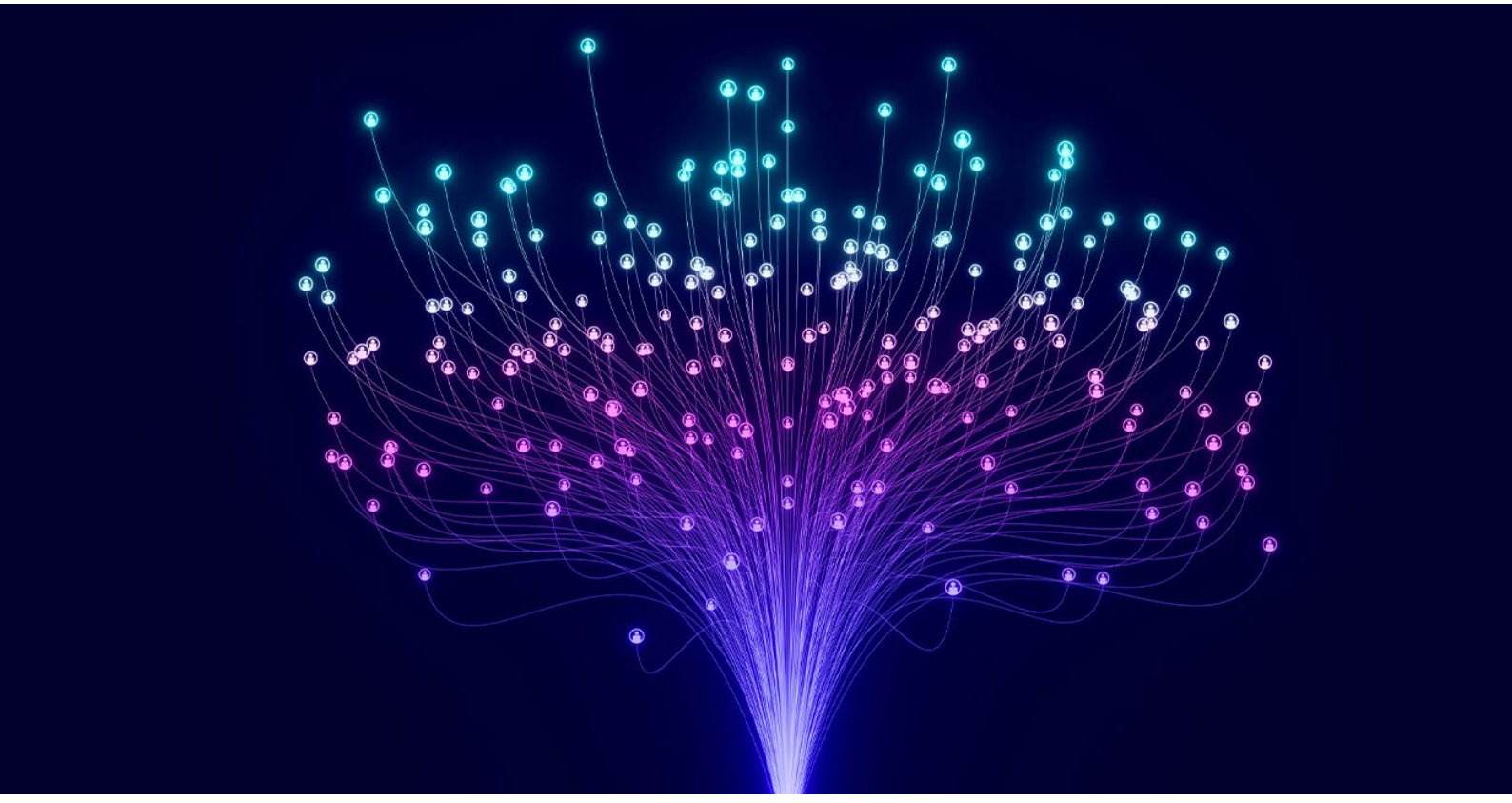
QuantumBlack
AI by McKinsey

Where AI agents pay off: A practical guide to the economics of agentic workflows

Early experience implementing agentic workflows highlights emerging trade-offs that frontline leaders need to understand.

This article is a collaborative effort by Chandana Asif, Dieter Kiewell, Lari Hämäläinen, Tom Kolaja, and Tunde Olanrewaju, representing views from QuantumBlack, AI by McKinsey.



August 2026

As companies rush to adopt agentic AI, they are learning that the cost dynamics are both complex and, at times, unexpectedly expensive. While the cost of individual tokens continues to fall, enterprise spending on AI is spiking.

At the same time, the pricing structures continue to evolve and shift, injecting a high degree of uncertainty into "tokenomics"—or AI economics, the discipline of managing agentic AI spend at scale. As of early July 2026, using a frontier model such as OpenAI's GPT-5.5 costs about $5 per one million input tokens and $30 per one million output tokens, compared with just $0.20 and $1.25, respectively, for an earlier lightweight model.[1]

Clearly, these models offer different capability levels, but not all tasks require the most expensive model option. To date, however, many leaders don't operate based on that understanding, leading to significant cost implications.

Even as the C-suite navigates the broader economic implications of AI, business-line leaders and domain owners charged with implementing agentic programs need to make practical decisions, such as figuring out which workflows to target and how to allocate both work and budgets across them. At the same time, they face pressures to act fast, show progress, and deliver value.

These are high-stakes challenges. Customer-facing AI agents in some banks, for example, can cost as much as $20,000 to $30,000 to run a single-agent workflow[2] and $100,000 to $200,000 to run a multiagent team, our analysis shows. Multiply that by the number of runs a multiagent team typically makes and the many workflows that can be "agentified"—and throw in the changing dynamics of AI pricing—and it's easy to see both how implementation leaders can lose control of costs and how AI can fail to create value.

These are still early days, but it's clear that business leaders need to understand the evolving economic trade-offs when implementing agentic workflows. In our work with leaders across sectors, we see that those who navigate these trade-offs most effectively ask themselves three questions:

— Where are the most significant economic opportunities emerging?

— What drives the unit economics of AI-based workflows?

— How do I evolve my choices in line with changing AI economics and capabilities?

Where are the most significant economic opportunities emerging?

Historically, technology prioritization was relatively straightforward. Leaders developed a business case and estimated implementation costs against expected gains.

Agentic systems are different. Economics evolve continuously as model capabilities improve, costs decline, and oversight requirements change. A workflow that fails the economics test today may become attractive months from now. Another may look compelling in a pilot but disappoint once deployed at scale.

[1] OpenAI Developers, Flagship models, prices per 1M tokens, accessed July 2026.
[2] McKinsey analysis based on public research and public pricing information.

Our work in this area has led us to identify two important tests in identifying the business value of agentic deployments that are unique to AI:

— *How can AI agents expand the business model or unlock new offerings?* Historically, organizations have had to limit their economic aspirations because of constraints on what was possible. In some cases, for example, top experts were a limited resource, so it only made economic sense for them to serve high-value customers. With AI, cognition is scalable and cheap, making it possible to open formerly uneconomic markets or tap long-tail opportunities.

  Agents can change the unit economics of personalized service in areas such as mass-affluent banking, for example, where tailored relationship support has historically been difficult to justify at scale.

— *Are the unit economics of these new agentic workflows attractive at scale?* It is easy to become excited by what agents can accomplish. But technically impressive capabilities do not necessarily create meaningful enterprise value. For AI to generate attractive economics, workflows need both sufficient scale and enough automation potential for value to compound over time. We've found in early work with companies that some of the strongest opportunities sit in high-value workflows with large pools of repeatable work where small improvements compound over thousands or millions of transactions.

What drives the unit economics of AI-based workflows?

Determining where agents have taken on work previously done by people is crucial, but that is just the beginning. Our experience so far suggests that when you unpack the economics of designing AI-driven workflows, several observations are particularly important.

Tokens sometimes aren't the highest cost

To manage agents, leaders need a "total cost of ownership" mindset. That includes understanding both fixed and variable costs (Exhibit 1):

— *Variable costs.* Token costs can vary significantly based on the task. High-compute tasks, like generating code, have significant token costs. But for an agent performing a customer service task in banking, token costs frequently represent just 20 to 25 percent of the variable run costs of an AI agent. Human oversight, on the other hand, accounts for 70 to 75 percent of the variable costs. Functional and risk experts often tend to perform these oversight tasks. In banking customer onboarding, for example, we would expect 10 to 20 percent of agentic runs to be reviewed by risk and functional experts.

  For workflows that are highly deterministic, have limited variables, or have few compliance needs—and as AI agents become increasingly reliable—human oversight might be less necessary, thereby reducing variable costs. Other needs, however, such as cybersecurity and training, are likely to increase the expense of agentic programs, creating ongoing uncertainty around variable agentic costs.

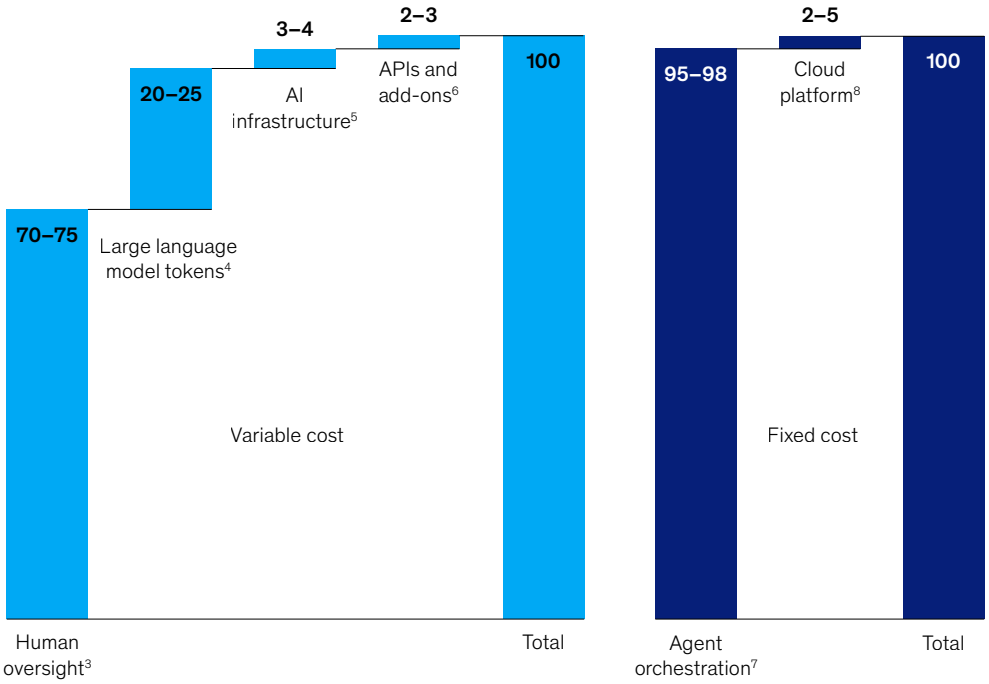
— *Fixed costs.* AI infrastructure and agent orchestration costs are fixed per agent and make up the rest of the annual run costs of an agent. AI infrastructure costs include public cloud containers, memory, management, and analytics, while agent orchestration includes the data scientist capacity required to maintain and enhance the agent in production.

Exhibit 1

In the banking sector, human oversight is a significant cost category in some agentified workflows with current tech.

Total annual run cost per agent, banking sector, illustrative,[1] % share (10,000 runs)

Cost distribution per agent run at 10,000 runs,[2] %



[1]Customer-facing agent that drives activities with direct profit-and-loss impact. [2]Midpoint values used. [3]10–20% of runs reviewed by human full-time employees (FTEs) for ~5 minutes each, at ~$100,000 annual cost per FTE. [4]Approximately 70% of runs use frontier or higher-cost models at published token pricing; each run averages 12 turns, 4,000 input words, and 500 output words per turn, with ~1.3 words per token. [5]Orchestration, model gateway, observability, and vector database reads priced per turn, and vector database writes priced per run, using Langfuse, LangGraph, and Portkey benchmarks. [6]Approximately 10–15 tool and data API calls per run. [7]Approximately 500 agents per enterprise, managed by agent operations FTEs at ~$150,000 per year and 20 agents per FTE. [8]Approximately $80,000 per year platform cost based on Azure benchmarks, covering a 500-agent fleet with no corporate discount.

McKinsey & Company

Many organizations focus their optimization efforts on model selection and token costs because those are the most visible expenses. The more productive approach is to redesign workflows to reduce exception rates and simplify review processes that require costly human interventions.

Three elements primarily affect the token cost of a run

Three main factors determine token costs: the quality of the model selected (a lightweight general model versus a specialist reasoning model), the latency required, and the task frequency (a one-step lookup versus an open-ended job that loops thousands of times).

Model routing and rightsizing, caching and reuse, workflow redesign that reduces inferences, batching, token budget discipline, and committed use pricing can all reduce token consumption.

Scale depends on amortizing fixed costs and agent reuse

Agent economics favor workflows with two particular characteristics. One is high volumes (Exhibit 2). Once fixed costs are set, for example, increasing agentic run volumes amortize that spending.

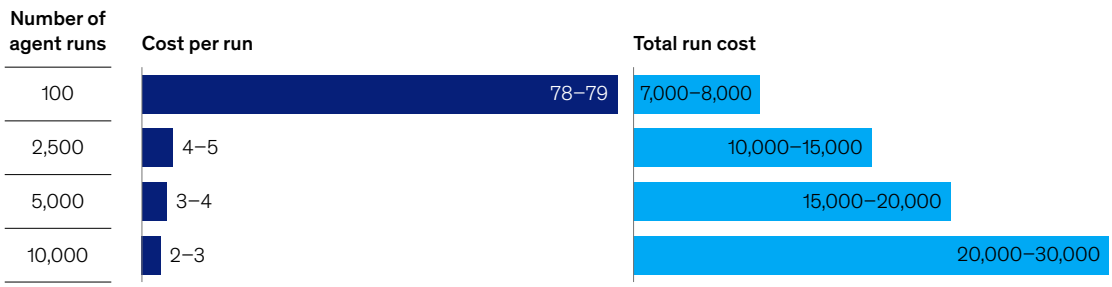
In one particular example, using a conversational agent to onboard 2,500 new customers per year costs $10,000 to $15,000. Doubling the volume of new customers increased costs to just $15,000 to $20,000.

The other characteristic is reuse. Leaders should prioritize building agents once and then reusing them across multiple high-value workflows. That requires a disciplined evaluation to identify tasks that are common across priority workflows.

Exhibit 2

The cost per run of an agent can decrease with volume.

Cost per run and total run cost, by number of agent runs, $



McKinsey & Company

Completed work ROI is the metric that matters most

A single agent can look cheap or expensive on its own and still say nothing about whether the workflow pays off. What matters is tracking impact: the "fully loaded" cost to finish the job (for example, a completed onboarding, a resolved claim, a closed sale) with humans, agents, and deterministic systems, relative to the value generated.

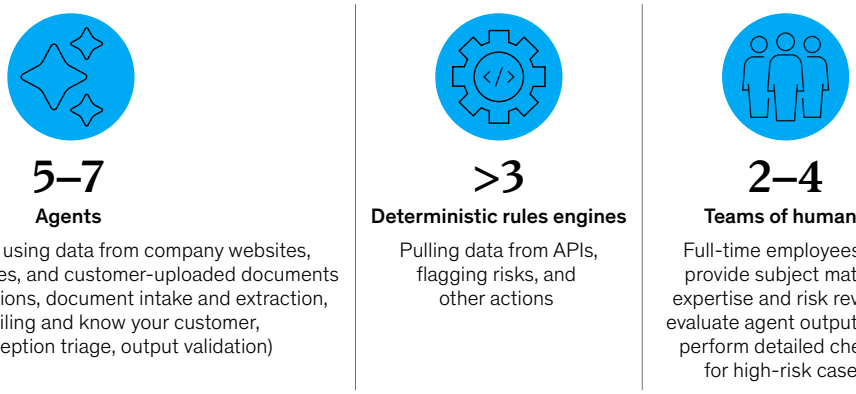
That fully loaded workflow is often more complex than leaders expect. Take the illustrative example of a customer opening a bank account (Exhibit 3). The completed task requires five to seven agents, multiple deterministic systems, and two to four teams of people providing oversight.

This overview is a simplification but highlights the support needs (and cost drivers) that leaders need to form a more accurate ROI picture. In this case, using standard benchmarks, the total cost to complete the onboarding workflow can still fall from about $50 to $150 per customer to about $10 to $30.

Exhibit 3

An illustrative customer onboarding journey highlights the range of elements needed to support agentic workflows.

Agentic workflow for a retail customer, illustrative



McKinsey & Company

The economics are dynamic

Perhaps the most important point leaders should focus on is that agent economics continue to evolve. Workflows that are uneconomic today may become attractive (or less attractive) tomorrow as model capabilities improve and economic dynamics change. This means that leaders should establish a regular review process to reassess workflows as model capabilities, costs, and operating practices evolve.

How do I evolve my choices in line with changing AI economics and capabilities?

Unlike traditional software projects, agent economics change much more often and more significantly. Business leaders, therefore, need to focus on a few key disciplines to manage agentic work over time.

Budget for evolution, not just the build

Agentic workflows should be funded with the expectation that their economics and capabilities will change over time. Production agents, for example, often need to be tweaked every couple of days as new foundation models, model context protocols, enterprise systems, and business requirements emerge. In other cases, it will make more sense to sunset an agent and build another one.

As the AI tech estate grows, leaders also need to budget for AI gateways, agent operations (AgentOps), monitoring, compliance checks, model approvals, and remediation required to keep the companies secure and in compliance. A change in regulation or internal risk appetite (for example, a restriction on which frontier models are permitted) can force agents to be reevaluated, modified, or paused.

Stay close to central management to access AI capabilities and provider relationships

Business leaders should not build everything themselves. Many AI capabilities are proven and easy to use. As leaders navigate what to build, partner with, or buy, they should maintain strong connections to a central team, which can provide two key benefits. One is that the central team can continuously evaluate models, tools, and providers; negotiate commercial terms; identify lower-cost alternatives as they emerge; and oversee provider and partner risk to ensure common standards for security, auditability, and observability.

Second, the central team can maintain reusable capabilities and establish helpful guardrails. By operating within these enterprise standards, domain teams can reduce tool and agent fragmentation while tailoring or developing assets that meet their unique workflow needs.

Build an AgentOps capability and management cadence

Just as organizations built a FinOps capability to manage cloud spend, business units now need an equivalent discipline for continually managing agents. That discipline requires a cross-functional team that can continuously manage spend and reallocate tasks.

To manage agentic programs over time, leading organizations should incorporate AI unit economics into quarterly business reviews (QBRs). QBRs provide leaders with the transparency to see where AI is creating the most value, which workflows still justify spend, and which can be redesigned, consolidated, or shifted to lower-cost models or enterprise capabilities.

A new wave of implications and questions

We are only in the foothills of the AI economics climb. While frontline leaders should rightly focus on today's economic trade-offs when implementing agentic workflows, they need to soon confront a set of second-order questions that are emerging. While it is too early to offer definitive answers, some important questions include the following:

- When should AI costs be passed through to customers or absorbed by the organization?
- How much pricing flexibility do business leaders need to build into their budgets?
- How do current budgeting, governance, and operating disciplines (internal chargebacks, cost coding) need to evolve to account for AI economics?
- How should cost and pricing decisions be made as AI agents work increasingly across functional boundaries?

Preparing for and answering these questions will require leaders to build stronger AI economics muscles so they don't inadvertently make choices that damage their business model over time.

The business-line leaders and domain owners creating the greatest value from agentic AI will be those who understand how to use unit economics to drive better decisions. While many dynamics of AI economics are still evolving, what is clear is that those who invest in the necessary capabilities are best positioned to win an enormous prize.

Chandana Asif is a partner in McKinsey's London office, where **Dieter Kiewell** and **Tunde Olanrewaju** are senior partners; **Lari Hämäläinen** is a senior partner in the Seattle office; and **Tom Kolaja** is a senior partner in the Columbus office.

The authors wish to thank Brandon Shi, Keisha Rastogi, and Varsha Seshadri for their contributions to this article.

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