McKinsey Quarterly

Is that AI agent worth it? Agentic economics and the modern operating model

As companies focus on spiraling AI costs, they need to make sure not to lose sight of the real prize.

This article is a collaborative effort by Lari Hämäläinen, Mark Patel, Sven Blumberg, Tanguy Catlin, and Wasim Lala, representing views from QuantumBlack, AI by McKinsey.

July 2026

For all the news about falling token prices, many enterprise leaders are experiencing sticker shock around AI agents. While the Stanford HAI 2025 AI Index reported that inference cost of GPT-3.5-level capability fell from $20 per million tokens to $0.07 through 2024,[1] enterprise large language model (LLM) spending tripled over a 12-month period by the end of 2025.[2] Some 93 percent of respondents to a McKinsey survey report exceeding their AI budgets,[3] while one-fifth of respondents to McKinsey's forthcoming global State of AI survey reported that their organizations have constrained use of AI because of AI-related operating costs.[4]

What gives? There are many reasons for the spike in AI costs: Enterprises are scaling up their AI efforts; LLM providers have pivoted from subscription to consumption, which has created new incentives (for example, answer length has increased to drive token usage); and expensive models are often used for simple tasks.

We go into more detail on these and other causes below, but it's worth waving a bright red flag before that to highlight a critical point: CEOs should not focus just on token cost reduction conversations. As David Tepper, CEO of Pay-i, frames it: Tokens are not value; tokens are the bill. The focus, instead, needs to be squarely on how to use AI agents to create value.

A lot goes into what determines that value, such as whether the agent output is correct, whether humans must supervise or repair it, how much compute it consumes while reasoning, and whether the value of the completed work exceeds the full operational cost of generating it. In essence, for CEOs this boils down to answering a basic question: Are the AI agent capabilities we're building and running worth the value we're getting from them?

Answering that question is becoming ever more pressing for CEOs as AI systems move from agentic coworkers to agents capable of more complex tasks like reasoning and orchestrating workflows across the enterprise. Furthermore, AI spending is expected to rise to roughly 25 percent of enterprise IT budgets within the next several years. Yet many organizations still cannot clearly explain which AI systems are generating value, what they truly cost to operate, or how those economics change as usage scales.

Getting clear on the supply and demand forces shaping AI's cost curve, and their implications on competitive dynamics, will determine how well companies manage this phase of the AI revolution. Our experience working with companies and on our own AI program demonstrates that leaders need to manage machine work differently, from allocating intelligence like capital to recalibrating the balance between insourcing and outsourcing (see sidebar "What McKinsey has learned from managing AI economics at scale"). That management discipline does not yet exist in most enterprises.

[1] Nestor Maslej et al., "The AI Index 2025 annual report," AI Index Steering Committee, Institute for Human-Centered AI (HAI), Stanford University, April 2025.
[2] Tim Tully, Joff Redfern, Deedy Das, and Derek Xiao, "2025: The state of generative AI in the enterprise," Menlo Ventures, December 9, 2025.
[3] Enterprise AI FinOps Survey, May 2026; 120 enterprise participants, 75 qualified respondents across five major industries.
[4] The forthcoming 2026 State of AI survey was in the field from May 4–June 8, 2026, and garnered responses from 1,719 participants representing the full range of regions, industries, company sizes, functional specialties, and tenures.

What McKinsey has learned from managing AI economics at scale

As AI adoption has expanded across McKinsey, we have experienced many of the same economic challenges our clients are beginning to face. While we are still learning, several key lessons have emerged:

— *Growth is exponential and shows little sign of abating.* As of May 2026, McKinsey is processing roughly five trillion AI tokens monthly. AI usage has grown in waves, expanding the user base, increasing consumption, and introducing new patterns of demand. In these initial phases, there were no limits placed on usage in order to accelerate AI skill development. We are now shifting to instituting guidelines and guardrails to better manage costs ahead of the next wave, where autonomous agents will consume intelligence on behalf of users rather than simply responding to prompts. Demand management has become a permanent capability rather than a one-time exercise.

— *Consumption follows a power law of distribution.* AI usage is highly concentrated. Roughly 10 percent of users account for about 65 percent of total token consumption. Software engineers and consultants are among the top users, generally driven by personal or working team interest, though there is a very long tail of users.

— *Pooling consumption spend is the best way to distribute costs.* LLM providers are transitioning from seat-based to consumption-based licensing models (combined with ad hoc user transparency about these consumption costs). Pooling consumption costs across the enterprise allows us to support legitimate power users while smoothing demand across the broader enterprise and avoiding unused capacity locked into individual licenses. Understanding where demand concentrates is often more valuable than understanding average usage.

— *Older models are better (and cheaper) for most use cases.* Users tend to gravitate toward the best (and most expensive) model, even though many of their tasks can be done with much less sophisticated models. We have found meaningful savings through intelligent model routing (the right intelligence at the right time) and matching different models to different stages of a workflow—for example, using a frontier model to generate an initial hypothesis before handing subsequent refinement to a smaller model.

— *Pricing flexibility is increasing.* The same model may be available through multiple providers, each offering different pricing. Enterprise discounts can often be stacked.

— *Automate cost decisions as much as possible.* McKinsey performs prompt optimization and has invested in an internal AI gateway to route LLM vendors (with plans to perform model routing in the future) and other optimization options transparently before requests reach model providers.

— *Simple tips and updates can have dramatic effects.* Providing transparency to each user about the cost of a prompt helps people make more economical decisions. Giving users specific guidelines and tips at the point of usage can have a significant effect. Encouraging more concise prompts and outputs ("caveman language"), for example, can reduce token consumption for some workflows by 30 to 40 percent without materially affecting quality.

— *There are significant costs beyond the model itself.* Some of the fastest-growing AI costs sit outside the LLM. Security gateways, orchestration platforms, monitoring tools, and even software vendors are increasingly pricing services based on token consumption. When you consider that the same tokens may be inspected or processed multiple times as they move through the enterprise architecture, that can have big cost implications.

Why the operating expenditure bill keeps rising: The six drivers of agentic economics

Six patterns explain why agentic work negates token-price deflation:

— *Long-lived context.* LLMs are stateless, which means agents often resend prior context as work progresses, compounding the total cost. Agentic tasks can consume roughly 1,000 times more tokens than code reasoning (single-turn problem-solving without tool interaction) or chat tasks (multi-turn dialogue about a coding problem).[5] Context becomes an operating asset and accounts for a significant portion of agentic costs. Poorly managed context turns every step of a workflow into a recurring cost.

— *Refinement is the sink.* In agentic workflows, the expensive part is not the first answer generated but the checking, repairing, and reverifying that follows. About 60 percent of an agentic task's costs, in fact, are tied to refining answers.[6] Leaders need to address and factor in quality control, exception handling, and rework as part of process costs across all workflows.

— *Autonomy creates variance.* Agentic systems can produce materially different costs for the same task because they may take different paths, call different tools, or retry in different ways. In programming, for example, the same task can have a factor-of-30 variation between completions.[7] Cost behaves as a distribution, not a fixed unit price. Average-cost budgets will not be enough.

— *Expensive reasoning is used for basic tasks.* Extended thinking pays for itself many times over on hard tasks. For easy ones, it is expensive overhead. Model routing is critical to ensure that expensive reasoning is reserved for work where it can change the outcome, and that cheaper models are used for basic tasks.

— *Agent choice orchestration can compound costs. How* an agent calls a tool matters as much as *what* it calls. The way work is decomposed, coordinated, and handed off across agents, tools, and models can change costs dramatically without changing the business outcome.

— *Information structure can drive inefficiencies.* How efficiently information is presented to or by a model (for example, prompt design, context length, language, formatting, and data structure) affects token consumption. Non-English text, for example, gets fragmented into more tokens per meaning, so the same conversation costs more in some languages than others.

These six drivers help explain why only making cheaper model calls does not automatically produce lower AI bills (read more on this topic in our upcoming article "The cost of intelligence: How CIOs can manage AI demand at scale").

The implications for competitive dynamics

While AI is creating complexities in terms of cost management, it is also forcing significant recalibrations around competitive dynamics. As CEOs determine where AI will create value and how to develop their business's strategic advantages, they need to pay attention to four implications in particular.

[5] Longju Bai et al., "How do AI agents spend your money? Analyzing and predicting token consumption in agentic coding tasks," *arXiv* 2604.22750, April 2026; "How we built our multi-agent research system," Anthropic, June 13, 2025.
[6] Mohamad Salim et al., "Tokenomics: Quantifying where tokens are used in agentic software engineering," *arXiv* 2601.14470, January 2026.
[7] Longju Bai et al., "How do AI agents spend your money? Analyzing and predicting token consumption in agentic coding tasks," *arXiv* 2604.22750, April 2026.

The first implication is that process advantage becomes harder to defend. For decades, companies created advantage through superior execution, such as better underwriting, faster claims handling, more efficient operations, stronger customer service. Agentic systems compress those advantages because the same capability can now be scaled across thousands of workflows through software platforms. What once required years of process redesign and implementation can increasingly be done in months or weeks.

The second implication is that data and context become more valuable than models. As foundation models converge, the differentiator increasingly shifts toward proprietary sources of advantage. The same commercial agent will produce radically different outcomes depending on the customer signals, operational data, decision history, and workflow context it receives. In an agentic world, the contextual data the enterprise captures systematically, such as call transcripts, decision precedent, and process telemetry, is becoming the layer that creates a competitive advantage.

The third one is that AI governance becomes a new basis of competition. The economics of machine work exhibit enormous variation. The same task can cost 30 times more depending on how it is executed. Similarly, the same infrastructure can produce dramatically different levels of output depending on how intelligence is allocated and routed. The organizations that learn to govern machine work effectively—measuring it, allocating it, improving it, and integrating it into operations—will create more value than their competitors from the same intelligence spend.

The fourth implication is that agentic economics are redefining firm boundaries in terms of what should be in- or outsourced. For years, sourcing decisions were largely driven by labor costs, scale economies, and coordination overhead. But the effectiveness of agentic systems depends less on the model and more on access to context, as we mentioned above. Enterprises need to treat this context capability like a crown jewel and keep it inside the business, while other tasks that are not core to its competitive footprint can be outsourced.

The CEO to-do list for building the agentic operating model

The CEO agenda is to design the enterprise's agentic operating model to maximize value, not to optimize token usage. It requires a management discipline for machine work: what outcomes it is measured against, where and what type of intelligence is allocated against those outcomes, how much autonomy agents are granted, and which workloads justify different deployment models. The following imperatives are what CEOs should require before agentic systems scale:

— *Allocate intelligence like capital.* One of the most expensive mistakes organizations make is assuming every problem requires frontier intelligence. In reality, many enterprise workflows can be performed effectively using smaller models, open-weight models, deterministic systems, or traditional software. The CEO should require that each major workload have a clear answer for why it runs where it runs, what would cause that decision to change, and who reviews it as prices, regulation, usage volume, and model quality move.

— *Set the competitive strategy before the technology strategy.* AI should force every company to ask itself: "What is the essence of our business?" and "Where are our sources of advantage?" The highest-return AI investments are rarely spread evenly across the enterprise. They tend to concentrate in a handful of economically critical capabilities. In insurance, claims and underwriting may dominate. In pharma, the largest returns may come from R&D and clinical development. In software companies, they may come from engineering. The first executive task is therefore identifying the handful of domains where machine work can most materially change the economics of the business.

As CEOs develop their strategy to build competitive advantage, it will be important to create a "glide path" that combines bold aspirations with practical intermediate goals. Clear milestones tied to outcomes, budget development and allocation requirements, and sufficient flexibility for teams to meet goals are core elements of an effective plan.

— *Build agentic operations as a managed enterprise discipline.* Enterprises have spent decades building disciplines for financial control, lean operations, supply chain productivity, cyber risk, and cloud cost management. They now need the equivalent discipline for agentic work. Tool architecture, model routing, evaluation, and autonomy should be governed as management decisions, not merely technical ones. No autonomous system should operate without a defined mandate, budget, and stopping rule. Fund this as a multiyear capability build, not a project line item. The capability should not be outsourced to vendors by default.

— *Answer the "who owns the capability" question.* The technical and operational levers required to govern machine work (such as context management, model routing, evaluation, autonomy, and workload placement) don't sit cleanly within the mandates of today's technology, finance, operations, or human resources leaders. That lack of clarity tends to lead to mistakes, redundant activities, and waste. The CEO will need to determine who leads AI operations (see sidebar "What machine-work economics means for key leaders").

What machine-work economics means for key leaders

Chief information officer (CIO): Build the system to manage intelligence. The CIO's role extends well beyond managing AI infrastructure. It is to ensure that every agent has access to the right enterprise context at the right time by connecting systems of record, operational data, unstructured content, and enterprise semantics into a coherent information layer. That requires modernizing data governance, managing both structured and unstructured information, and building the enterprise ontology, metadata, and context services that allow agents to reason reliably across the organization.

Chief technology officer (CTO): Improve the economics of machine work. The CTO increasingly owns the productivity of intelligence itself. Decisions about architecture, model selection, context management, routing, caching, evaluation, and autonomy directly affect the cost and performance of machine work. The key question is how to deliver the same business outcome with technologies that deliver less intelligence consumption, lower latency, higher quality, or greater reliability.

Chief financial officer (CFO): Turn AI into an economic asset class. For the CFO, AI is becoming a new category of enterprise spend that requires its own measurement, forecasting, and capital-allocation frameworks. The focus should shift from technology budgets to economics—for example, cost per outcome, return on intelligence, and the productivity of machine work relative to human work and traditional automation. The objective is to ensure that spending is directed toward the highest-value opportunities.

Whether the answer is a chief AI operations officer, a new operating committee, or another model matters less than being explicit about where ownership and accountability lie. They should have explicit responsibility for critical KPIs, such as cost per outcome, vendor performance, learning cadence, and the operating effectiveness. They also need to stay on top of market dynamics to anticipate changes, and build sufficient flexibility into the operating model to adjust.

— *Measure machine work with metrics that matter to the business.* The unit of governance is the completed business outcome, not the token, model call, or technology line item. CEOs should require that every material agentic workflow be measured against business outcomes rather than solely technical metrics. The specific metric matters less than the principle. Leaders will need visibility into tail costs, escalation thresholds, and acceptable overall cost levels, rather than single execution costs. The objective is to create a common economic unit that allows management to compare human work, machine work, and hybrid work.

— *Rethink your sourcing boundaries.* Understanding the value dynamics of AI should force a major recalibration of the existing outsourcing/insourcing model. Some activities that once made economic sense to outsource may become more valuable to keep close to the enterprise because machine work performs best when tightly integrated with proprietary context. At the same time, other activities that previously required significant internal capability (such as legal research, financial analysis, or content generation) may increasingly be delivered through agentic services with relatively light internal oversight.

  Perform an activity-by-activity review to determine which should be managed inside the enterprise or outside based on sources of value. The default direction is no longer "outsource more." At the same time, review contracts and renegotiate terms to privilege sources of competitive advantage and provide more benefits based on economic dynamics. Build the muscle to move activities back in-house when the case calls for it.

— *Reshape the buy versus build versus partner portfolio.* Treat workload placement as a portfolio review, not a single moment of build versus buy. The right decision for any given workload depends on volume, predictability, sensitivity, quality threshold, and operating maturity (table). Over time, most AI capabilities will likely be purchased through commercial software, models, and services. Internal development increasingly belongs in only three areas: proprietary "glue layers" that connect enterprise-specific data and workflows, differentiated capabilities where the agent itself is a source of advantage, and strategic learning investments that help the organization build expertise and shape future decisions.

  Commercial agent-native software in many categories will likely take two to four years to mature. Enterprises that build selectively during this window will learn things that transfer across the rest of their portfolio and that buy-only competitors cannot replicate.

Table

Five enterprise scenarios and their likely workload placements

Scenario	Volume	Predictability	Sensitivity	Quality threshold	Likely placement
Expert reasoning (legal, strategy, research)	Low	Volatile	Medium	Frontier required	API flagship tier (capability beats unit cost)
High-volume classification (support triage, fraud signals)	High	Stable	Low-medium	Validatable	Managed open or private (unit cost dominates)
Regulated workload (banking, healthcare claims)	Medium-high	Stable	High	Medium-high	Private cloud/ sovereign (residency and control needed)
Heterogeneous agent platform (many use cases, mixed needs)	Mixed	Volatile	Mixed	Mixed	API + routing layer (optionality has real value)
Very high-volume internal task (summarization, document extraction)	Very high	Predictable	Low-medium	Validatable	Private/on-premises candidate (scale economics win)

Note: Scenarios drawn from observed enterprise patterns across financial services, healthcare, technology, and consumer industries. "Likely placement" is the typical answer; specific cases vary with regulatory posture, data sensitivity, and existing infrastructure. The placement decision is workload by workload, not platform by platform. Most enterprises will run on three or four placements simultaneously, with the routing layer making the choice transparent to the agent.

— *Invest in contextual data capture as a strategic asset.* Start the systematic capture of operational, customer-interaction, and process-telemetry context now. This context layer should be managed as a strategic asset, distinct from traditional systems of record. Treat data acquisition as a procurement category, with consent and retention discipline as part of the capability. Consider a three-year horizon where "compounded context" produces differentiated performance on otherwise identical commercial software.

The next wave of advantage will come from governing machine work with the same discipline that companies have learned to apply to human labor, capital, risk, and operations. Leaders that build that capability now can turn agentic AI into an operating system for the company. Those that do not may scale a new cost pool faster than they learn to manage it.

Lari Hämäläinen is a senior partner in McKinsey's Seattle office, **Mark Patel** is a senior partner in the Bay Area office, **Sven Blumberg** is a senior partner in the Istanbul office, **Tanguy Catlin** is a senior partner in the Boston office, and **Wasim Lala** is a partner in the Washington, DC, office.

The authors wish to thank Uwe Agsten for his contributions to this article.

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