

QuantumBlack, AI by McKinsey

The cost of intelligence: How CIOs can manage AI demand at scale

As AI costs spiral, CIOs need to manage enterprise AI demand to optimize for outcomes, not just cost.

This article is a collaborative effort by Pankaj Sachdeva and Wasim Lala, with Avinash Javaji, Kaavini Takkar, and Purva Arora, representing views from QuantumBlack, AI by McKinsey, and McKinsey's Technology and AI group.



At a recent tech leadership meeting for a tech services company, a team enthusiastically approved a new AI capability. As the discussion turned to implementation, one executive asked a simple question: “How much is this going to cost?”

The room went quiet. No one knew.

That lack of clarity is not unusual, and it is becoming one of the defining risks of enterprise AI. Across industries, organizations are discovering that AI costs can spiral far faster than traditional technology spending. In some cases, companies have exhausted annual AI budgets in a matter of months, forcing emergency contract renegotiations and unexpected funding requests.

Much of this spend remains invisible. Business units purchase AI capabilities independently. Employees build AI-powered workflows outside central IT. Citizen developers and “vibe coders” can unintentionally create autonomous agents or applications that consume millions of tokens every day.

As organizations move from isolated use cases to enterprise-wide adoption, AI spend increases nearly fourfold, according to a McKinsey survey.¹ While 62 percent of organizations have moved beyond experimentation into active deployment of AI, 93 percent of respondents report exceeding their AI budgets (Exhibit 1).²

This problem is poised to get more pronounced, with a majority of firms in our survey expecting their AI spend to increase by at least 25 percent over the next 12 months.³ Many organizations are realizing they lack the visibility and controls needed to manage that spend. “Tokenmaxxing” is becoming a dirty word, with some companies taking down their AI user leaderboards. The “go, go, go” energy driving AI over the past two years is yielding to a “wait a sec” moment.

While some companies may be tempted to halt their AI programs, thoughtful chief information officers (CIOs) are looking to scale AI while managing demand by mastering the emerging field of enterprise AI tokenomics (see sidebar “What is tokenomics?”). Think of it as FinOps for AI—a way to actively and continuously predict and manage AI model usage to maximize ROI.

What is tokenomics?

Tokenomics extends beyond prompt and inference costs. It encompasses the broader economics of AI consumption: model selection, routing decisions, orchestration patterns, agent behavior, workflow design, infrastructure utilization, and the elimination of waste throughout AI-enabled processes.

¹ McKinsey Enterprise AI FinOps survey, May 2026; 120 enterprise participants, 75 qualified respondents across five major industries.

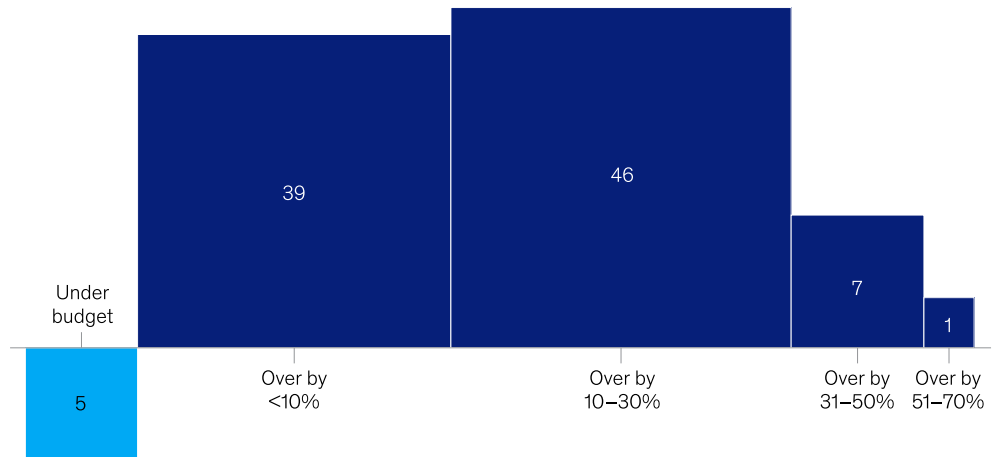
² McKinsey Enterprise AI FinOps survey, May 2026; 120 enterprise participants, 75 qualified respondents across five major industries.

³ McKinsey Enterprise AI FinOps survey, May 2026; 120 enterprise participants, 75 qualified respondents across five major industries.

Exhibit 1

Most companies are exceeding their AI budgets.

AI budget adherence, % of respondents



Note: Figures do not sum to 100%, because of rounding.
Source: McKinsey Enterprise AI FinOps survey, May 2026

McKinsey & Company

For CIOs, this is an important opportunity to shape how technology drives economic benefit for the business. In our experience, companies that are thoughtful in their AI consumption can save 20–30 percent on their AI costs. They’re finding these savings by developing the capabilities to optimize spend, improve accountability, and redirect savings toward the highest-value opportunities.

Why AI costs are so difficult to manage

Many CIOs instinctively compare the challenge of managing AI spend to the early days of cloud computing. The comparison is useful on one level because many of the governance issues are similar. But the unique issues regarding AI tied to tokens, API calls, and infrastructure usage add a new array of challenges for CIOs to manage:

- *AI usage patterns are not predictable.* The same task can generate dramatically different token volumes, invoke different models, trigger different agent chains, and produce significantly different costs. Agentic workflows multiply model calls per interaction, rapidly

outpacing budget assumptions built on prior usage baselines.

- *Consumption pricing exposes companies to true usage costs.* Bundled seat-based pricing acted as a “safety net” because it hid the usage costs. With large language model (LLM) providers shifting to consumption pricing, companies now have to pay directly for high-usage scenarios.
- *Governance levels are not mature.* There is no playbook, no tagging, no sourcing strategy. Without mature FinOps, tagging, and sourcing, organizations cannot forecast, allocate, or control AI-driven consumption at scale.
- *Teams do not know which model, tool, or pricing tier to use.* Users often default to premium models or familiar tools because trade-offs across quality, latency, risk, and cost are unclear.
- *Citizen developers and AI-assisted application development are accelerating consumption.* Employees can now rapidly build AI-powered applications, workflows, and autonomous agents with little engineering effort. While this democratizes innovation, it also dramatically increases unmanaged token consumption and makes forecasting significantly more difficult.

How to manage AI consumption

Our analysis and experience suggest four actions to better manage AI demand.

Know what you’re spending on AI to manage and forecast better

Access to an expanding ecosystem of off-the-shelf AI capabilities has led to AI sprawl with limited visibility into their cumulative cost. Organizations cannot optimize what they cannot see. Across many enterprises, AI spending remains fragmented across cloud providers, foundation model vendors, software platforms, experimentation environments, and business units. This makes it hard to have a view of enterprise-wide consumption, forecast future demand, or understand cost drivers. In fact, our experience has shown that 20–30 percent of AI spend is often unaccounted for because AI investments are fragmented across vendors, tools, and commercial models.

In one organization, an effort to establish a consolidated view of AI spending uncovered costs spread across enterprise copilots, foundation-model contracts, AI-enabled software features, API-based services, experimentation environments, and business-unit purchases. What appeared to be a straightforward technology budget exercise had evolved into a fragmented portfolio of AI expenditures with no single source of truth.

This lack of transparency often means that cost overruns frequently become visible only after consumption has already occurred. Furthermore, it hampers the development of reliable forecasts, a particular challenge given the nondeterministic nature of AI agents. Token usage can vary by up to 30 times when executing the same task.⁴

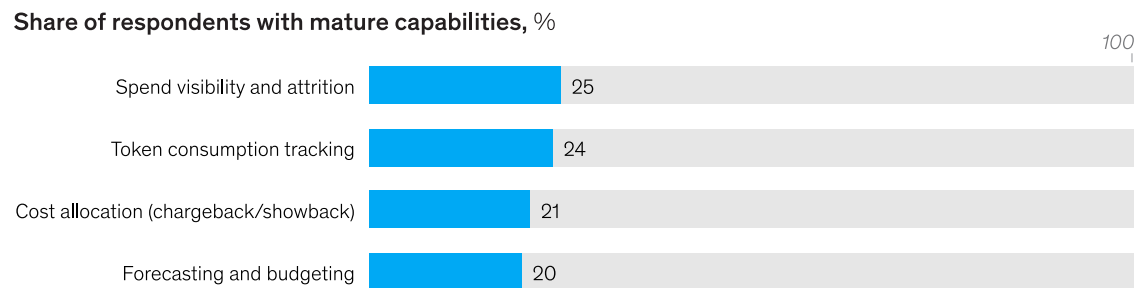
⁴Longju Bai et al., “How do AI agents spend your money? Analyzing and predicting token consumption in agentic coding tasks,” Stanford Digital Economy Lab, April 14, 2026; Matty Smith, “How are AI agents spending your tokens?,” Stanford Digital Economy Lab, May 5, 2026.

As organizations move from copilots to agentic workflows, CIOs will increasingly need demand forecasts that model multiple scenarios, including adoption growth, workflow expansion, routing changes, and shifts between frontier and open-weight models. In fact, organizations with high forecasting maturity save 10 percent more on AI spend than their peers on average.⁵

Meeting that demand requires strong AI FinOps capabilities. To date, however, only 20–25 percent of companies have mature AI FinOps practices in place (Exhibit 2).

Exhibit 2

AI FinOps maturity is generally low.



Source: McKinsey Enterprise AI FinOps survey, May 2026

McKinsey & Company

As CIOs build up their FinOps muscles for AI, they need to prioritize establishing a centralized AI control plane that provides a single source of truth for spend, usage, performance, and business outcomes as an important source of insight for scenario planning (see sidebar “What is an AI control plane?”).

This transparency creates accountability through showback and chargeback mechanisms that connect AI consumption directly to the business activities generating demand. This capability is critical, since the unit of governance should be the completed business outcome, not the token cost.

AI is emerging as a powerful tool for managing spend. Organizations are turning to AI for a widening range of tasks, including analyzing consumption patterns, recommending lower-cost models, forecasting demand, detecting inefficient prompt designs, identifying unnecessary

⁵ McKinsey Enterprise AI FinOps survey, May 2026; 120 enterprise participants, 75 qualified respondents across five major industries.

What is an AI control plane?

An AI control plane is the enterprise management layer that sits between users, applications, agents, and the AI models or platforms they consume. If enterprise resource planning (ERP) systems became the system of record for financial transactions, AI control planes may become the system of record for intelligence consumption. Its purpose is to make AI usage observable, attributable, and governable at scale across three vectors:

- *Visibility and attribution.* A control plane captures request-level telemetry across users, applications, workflows, models, prompts, agents, tokens, API calls, latency, and cost. By tagging usage to a business unit, product, use case, workflow, owner, and cost center, organizations can move from aggregate vendor invoices to meaningful unit economics, such as cost per task, case, code review, or customer interaction.
- *Policy and governance.* A control plane enables organizations to enforce rules in real time, including approved model access, role-based permissions, token budgets, spend thresholds, context-window limits, premium-model approvals, and exception workflows. This helps prevent avoidable spend and unmanaged risk before they occur.
- *Routing and optimization.* A control plane can route workloads to the right model or provider based on cost, quality, latency, risk, and availability. It can also identify optimization opportunities such as caching, prompt standardization, agent-loop reduction, output-length controls, and model right-sizing.

agent loops, and recommending where open-weight models could replace proprietary models without compromising business outcomes. Over time, AI systems will be able to continuously optimize other AI systems.

Optimize spend

Most organizations initially approach AI optimization as a model selection problem. The opportunity is significantly broader. The largest savings often come from optimizing the systems that consume tokens.

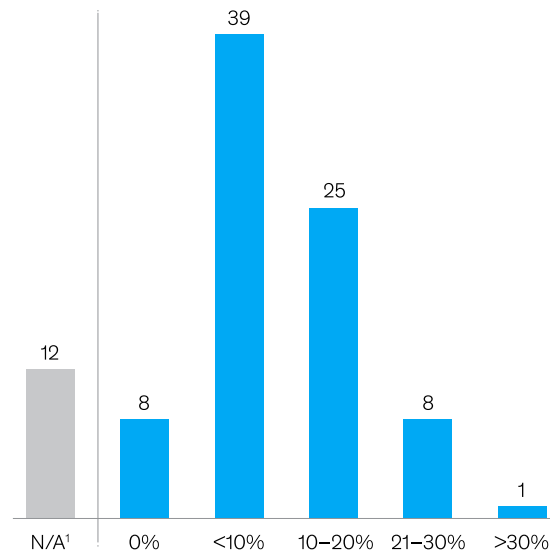
AI cost management is complex because it can fluctuate dramatically based on model choice, token consumption, routing decisions, agent behavior, orchestration patterns, and user adoption. Our analysis has shown that there are about 40 levers that generate the best outcomes in optimizing spend. Prompt caching (reusing static prompt context), for example, reduces repeated input-token costs by up to about 90 percent, especially for retrieval-augmented generation (RAG) and agents with large, stable prefixes.

AI-assisted optimization can also improve prompt efficiency by recommending shorter prompts, reducing unnecessary context, and standardizing prompt templates. About a third of organizations surveyed have already achieved savings of 20 to 30 percent through active optimization actions (Exhibit 3).

Exhibit 3

By actively optimizing their AI spend, companies can capture significant savings.

Savings realization in past 3 months, % of respondents



¹No optimization steps taken.
Note: Excludes responses of "not measured" (7%).
Source: McKinsey Enterprise AI FinOps survey, May 2026

McKinsey & Company

The following activities have the most impact on optimizing AI costs:

- *Selecting the right model type.* Few tasks warrant the use of expensive frontier models. Companies should route each task to the lowest-cost model that can deliver the required

quality.

- *Reducing bloated token inputs.* Shorten prompts, limit context windows, pass only relevant sections of documents or tool outputs, and summarize long conversation histories before sending them back to the model. Reusing static prompt context can reduce repeated input-token costs by up to about 90 percent, especially for RAG and agents with large, stable prefixes.
- *Controlling output-token consumption.* Define expected response formats, use structured outputs where appropriate, and cap response lengths by use case so models do not generate unnecessary text.
- *Managing agent and workflow volume.* Set limits on retries, tool calls, agent loops, and escalation paths; redesign repetitive or poorly structured workflows upstream rather than simply “throwing agents at the problem.”
- *Batching and caching.* Batch nonurgent, high-volume requests where latency is less important, and cache repeated prompts, stable context, and reusable intermediate results to avoid paying repeatedly for the same inference.
- *Using open-weight models selectively.* Consider open-weight models as alternatives to frontier models for simpler tasks that do not need the smartest, most expensive models (see sidebar “A closer look at AI models”).

Modernize sourcing and establish financial accountability

Traditional software procurement models were built around predictable license structures, annual commitments, and seat-based pricing. AI introduces a different reality that is defined by consumption-based pricing, rapidly evolving model ecosystems, fluctuating demand patterns, and continuous innovation. As a result, companies need new mechanisms to both govern AI consumption and manage providers and access.

Leading organizations from a McKinsey survey are seeing unit cost reductions of 10 to 20 percent through two sourcing and negotiation levers.⁶ The first is to develop an AI accountability model that provides visibility into who is consuming AI resources, what value is being created, and how costs are incurred across the organization. Similar to cloud FinOps, leading organizations are increasingly allocating AI costs—including token usage, API calls, infrastructure consumption, and model expenses—to the business units, products, or use cases generating demand.

Effective accountability models combine transparent chargeback or showback mechanisms with automated guardrails that monitor usage, enforce policies, and identify inefficiencies. This helps organizations move beyond viewing AI spend as a centralized technology cost and instead manage it as a business investment tied to measurable outcomes.

⁶ McKinsey Enterprise AI FinOps Survey, May 2026; 120 enterprise participants, 75 qualified respondents across five major industries.

A closer look at AI models

Enterprises increasingly have multiple ways to access AI models.

Proprietary models (including frontier models): These are the most advanced models. They typically deliver the strongest performance for complex reasoning, coding, multimodal tasks, and agentic workflows. For that reason, these tend to be the most expensive models. They are typically accessed through APIs or managed cloud services and priced based on consumption. Agents using these models can often manage an entire workflow.

Open-weight models: Open-weight models are AI systems whose core capabilities (or “weights”) are made publicly available. They provide enterprises with greater flexibility and can often offer lower-cost alternatives for many use cases. Organizations can access open-weight models through several deployment approaches:

- *Provider-hosted.* These models are hosted by a cloud or inference provider, allowing enterprises access to them without having to manage the underlying infrastructure. The trade-off is that companies still rely on an external provider for hosting, scaling, uptime, and some aspects of data governance.
- *Enterprise-hosted.* Some organizations run open-weight models on infrastructure they procure, configure, and maintain. This can provide greater control over data, customization, latency, and unit economics at scale, but it requires stronger engineering, MLOps, security, and infrastructure capabilities.
- *Edge or local models.* Smaller models can also run directly on employee devices, such as laptops or phones. These models are typically less capable than frontier models but can be useful for lower-risk, local work such as summarization, low-volume classification, and translation.

This visibility is increasingly important because it helps identify hidden costs that drive spend at scale: repeated failed/retried requests, content bloat on large prompts, conversation histories, agentic chains, and model proliferation.

The second lever involves developing an AI that goes beyond the traditional buy-versus-build questions. The growing availability of off-the-shelf AI solutions is increasing, and not reducing, the complexity of enterprise AI economics. As organizations gain access to more models, copilots, AI applications, and deployment options, the challenge shifts from acquiring a capability to governing when, where, and how that capability is consumed.

That’s why the buy-versus-build sourcing question is no longer valid. Increasingly, it is a question of finding the best mix of buy, build, host, route, and switch. Organizations must continuously evaluate which models or architectures are needed for each workload based on model performance, cost, risk, and business value to optimize their AI supply mix over time (table).

Table

AI outsourcing dynamics are significantly different from traditional sourcing.

Traditional software sourcing	AI sourcing
Seat-based licensing	Consumption-based pricing
Single-vendor strategy	Multimodel ecosystem
Fixed demand forecasts	Dynamic demand management
Long-term commitments	Flexible commercial structures
Periodic benchmarking	Continuous benchmarking

Embed governance directly into the architecture

CIOs face an unusual challenge: For the past three years, they have encouraged employees to use more AI. Now they must encourage employees to use AI more intelligently. Cost management needs to be part of a broader change management effort, rather than expecting people to learn to be cost-efficient in their AI use.

One way to address this issue is to automate and embed governance and spend management practices directly into the AI architecture. AI gateways, control planes, policy engines, and automated guardrails help enforce approved model usage, monitor consumption, automatically apply organizational policies, and enable responsible AI controls.

One tool being piloted internally within McKinsey, for example, coaches people on how to write better prompts and thoughtfully select models as they use LLMs. Educating and guiding people when they're actually using the product, instead of only requiring separate trainings, has proven to work.

What CIOs should prioritize now

The temptation in the face of these AI cost pressures is to simply cut spend. That would be a mistake. The better approach is to focus on shaping demand to create the most value for the business. In our experience, the most successful CIOs focus on the following actions:

- *Build forecasting capabilities before costs erupt.* As AI adoption expands, organizations need a clear understanding of where consumption occurs, what drives costs, and which use cases are generating value. CIOs should therefore invest in forecasting capabilities that model demand under different adoption, pricing, and workload scenarios. AI business cases should include projected consumption curves, sensitivity analyses, and expected cost-per-outcome metrics. The goal is to move AI economics from a budgeting exercise to a planning discipline.
- *Move AI to an AI business-value accountability model.* CIOs should work with finance and business leaders to establish cost-attribution models that connect AI usage directly to the products, workflows, and business units that generate demand. The objective is to create transparency around which use cases are generating value and which are consuming

Find more content like this on the
McKinsey Insights App



Scan • Download • Personalize



resources without meaningful returns. Over time, organizations should evolve toward measuring AI investments using business metrics such as cost per claim processed or revenue generated per AI-enabled workflow rather than simply tracking token consumption.

- *Anchor AI efforts around the enterprise's most valuable workflows.* CIOs should work with business leaders to identify the workflows where AI can create disproportionate value, whether through productivity gains, revenue growth, improved customer experience, or risk reduction. These workflows should be the focus of initial AI consumption capabilities because that's where the impact will be greatest, generating savings that can be reinvested in other technology priorities.
- *Maintain flexibility in sourcing and technology choices.* Given the pace of innovation, today's optimal model or vendor may no longer be optimal six months from now. CIOs should avoid locking the organization into rigid commercial structures or narrow technology stacks. Instead, CIOs should work with enterprise architects to design architectures and sourcing strategies that preserve flexibility. Key requirements should include supporting multiple model providers, maintaining the ability to shift workloads between proprietary and open-weight models, and regularly reassessing workload placement decisions.
- *Build a permanent AI FinOps capability before spending reaches scale.* CIOs should establish a cross-functional AI FinOps team responsible for forecasting demand, monitoring consumption patterns, identifying optimization opportunities, evaluating sourcing options, and measuring cost per outcome. This team should operate as a permanent enterprise capability rather than a temporary project office. It is this team's responsibility to stay on top of the fast-changing AI cost dynamics and integrate them into the FinOps capability quickly.
- *Embed governance directly into the AI operating environment.* As AI adoption expands, automation becomes the only practical way to govern consumption at scale. CIOs should consider embedding governance directly into the architecture through AI gateways, control planes, policy engines, and automated guardrails. These systems can route requests to lower-cost models when appropriate, enforce budget thresholds, limit unnecessary context expansion, monitor agent behavior, and trigger escalation when costs or risks exceed predefined limits. The objective is to make the economically efficient choice the default choice.

AI is rapidly becoming one of the largest and fastest-growing categories of technology spend. But this is not simply another cost-management challenge. The organizations that create the most value will develop the capabilities to forecast demand, govern consumption, and continuously optimize the economics of intelligence. That will be the defining challenge for CIOs in the age of AI.

Pankaj Sachdeva is a senior partner in McKinsey's Philadelphia office; **Wasim Lala** is a partner in the Washington, DC, office; **Avinash Javaji** is an associate partner in the New York office; **Kaavini Takkar** is an associate partner in the Seattle office; and **Purva Arora** is a knowledge expert in the Toronto office.

The authors wish to thank Dinky Khatri, Kiran Nagra, Nicolas de la Flor, and Raghuvar Choppakatla for their contributions to this article.

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