# Unlocking agentic value: a new investment discipline for the agentic era

As agentic AI reshapes the economics of enterprise technology, organizations have an opportunity to govern run-rate exposure as a growth investment, protecting operating budgets and margin while compounding enterprise value.

**Whitepaper 1: EY Total Cost of Agents series**

The better the question.
The better the answer.
The better the world works.

Shape the future with confidence

**The loudest conversation about AI right now is about what it costs.** The more important one is about what it earns. Effective governance treats that spend as a growth investment, directing it toward measurable enterprise value.

Token costs are the most visible line item in an AI budget and are often the first sign that the economics of agentic AI are changing. They are not the full cost, but they matter because they are where inference intensity, model choice, orchestration complexity and upstream compute scarcity first surface, often only after the work is done.

The issue is becoming more important as token pricing is ultimately tied to a constrained physical supply chain: chips, power, data centers, cloud capacity and model provider infrastructure. Enterprises may experience that constraint in the form of token bills, usage caps, rate limits, model access restrictions or unexpected repricing. For many companies pushing AI adoption, the last few months have been a shock to the budget.

Current pricing may also understate the long-term economics of agentic AI if some portion of compute cost is being absorbed, subsidized or strategically priced by upstream providers. As agentic use cases become more valuable and more compute-intensive, those costs are unlikely to remain invisible indefinitely. Someone will ultimately pay for the compute that agents consume. Many providers are already shifting from subscription-based fees to consumption-based models, ending the days of vendor-subsidized pricing.

But while tokens are the story today, the real economics of running agentic workflows at scale include infrastructure, operations, people, risk and the hidden cost of engineering around AI's limitations. Optimizing tokens without understanding total cost of ownership is like managing a factory by watching the electricity bill.

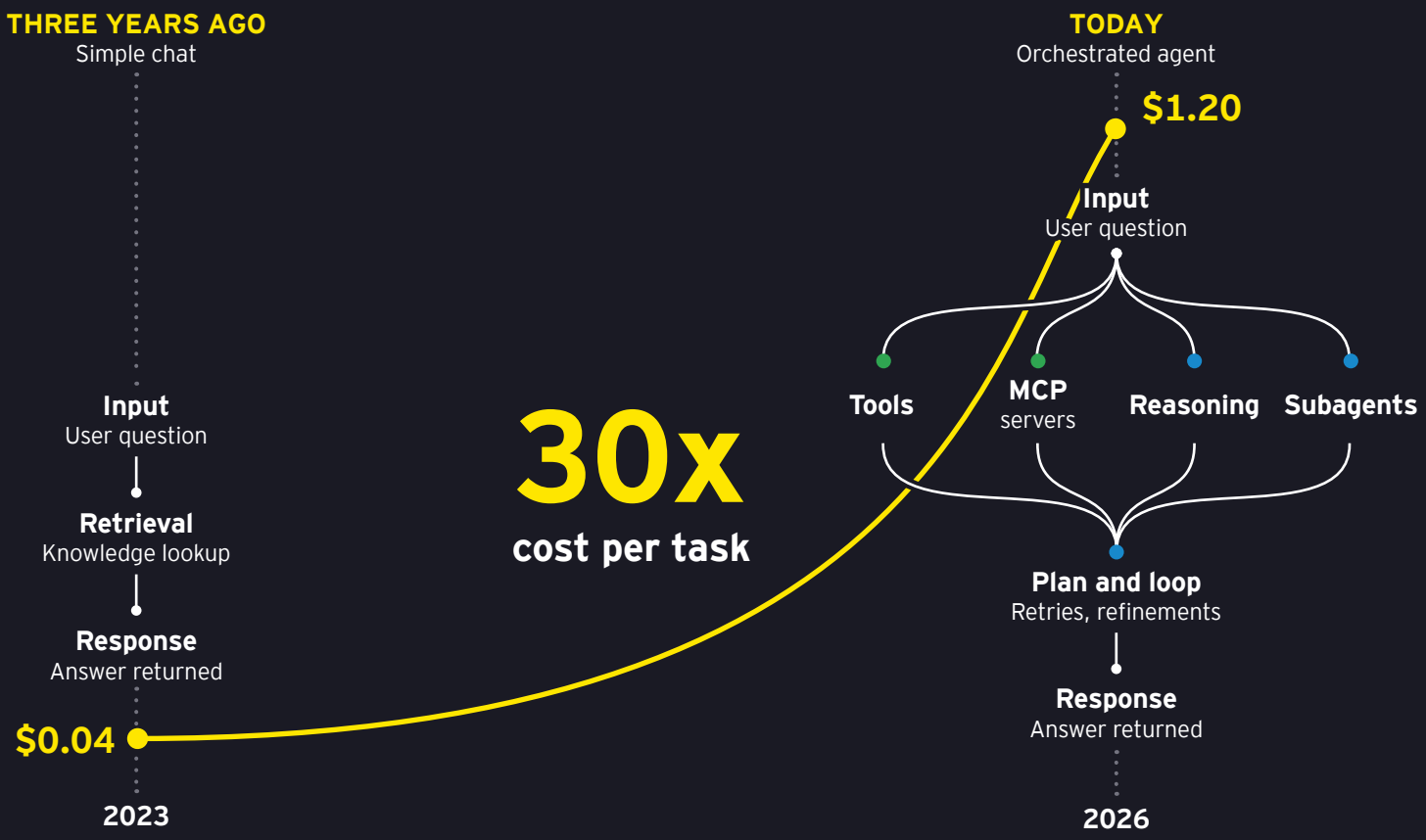
The story previously presented to many boards was straightforward: an expensive predictable line of human labor traded for a cheaper predictable line of software. But an agent is not a license. It lives in an operating stack of compute, models, data pipelines, governance, oversight and redesign work, each of which compounds rapidly. As a result, the tokens AI consumes, serving as the metered proxy for much of that operating stack, become a highly variable operating cost.

Consider a customer service AI assistant built two years ago to answer product questions. A chat that once cost $0.04 may now involve tool retrieval, planning and subagents that turn it into a $1.20 orchestration. Additional costs — such as knowledge-base updates, agent evaluation and human-collaboration design — may not appear on the model vendor's invoice but still surface to the enterprise.


When a CFO asks what one agent costs, they are often shown a vendor invoice that captures only part of the total financial exposure.

## Three years of cost evolution

One customer-service interaction, then and now.




**Same workflow. New architecture. New economics.**

EY analysis based on observed enterprise customer-service AI deployments

The challenge is that agentic AI shifts enterprise AI from a fixed-cost labor comparison to a dynamic compute consumption model. Boards and CFOs are therefore looking for investment cases that capture the full operating picture, including usage, orchestration, governance, change, risk controls and remediation.

This also changes how AI is budgeted and managed. CFOs need visibility into consumption across use cases. CTOs need to understand inference volume, model mix and retrieval load. CEOs need to assess not only labor savings, but the ongoing technology run rate required to deliver them.

Gartner predicts more than 40% of agentic AI projects will be canceled by the end of 2027 due to escalating costs, unclear business value or inadequate risk controls.[1] These outcomes are not inevitable. They are more likely where organizations scale agents before establishing agentic FinOps to manage cost, risk, usage and value together.

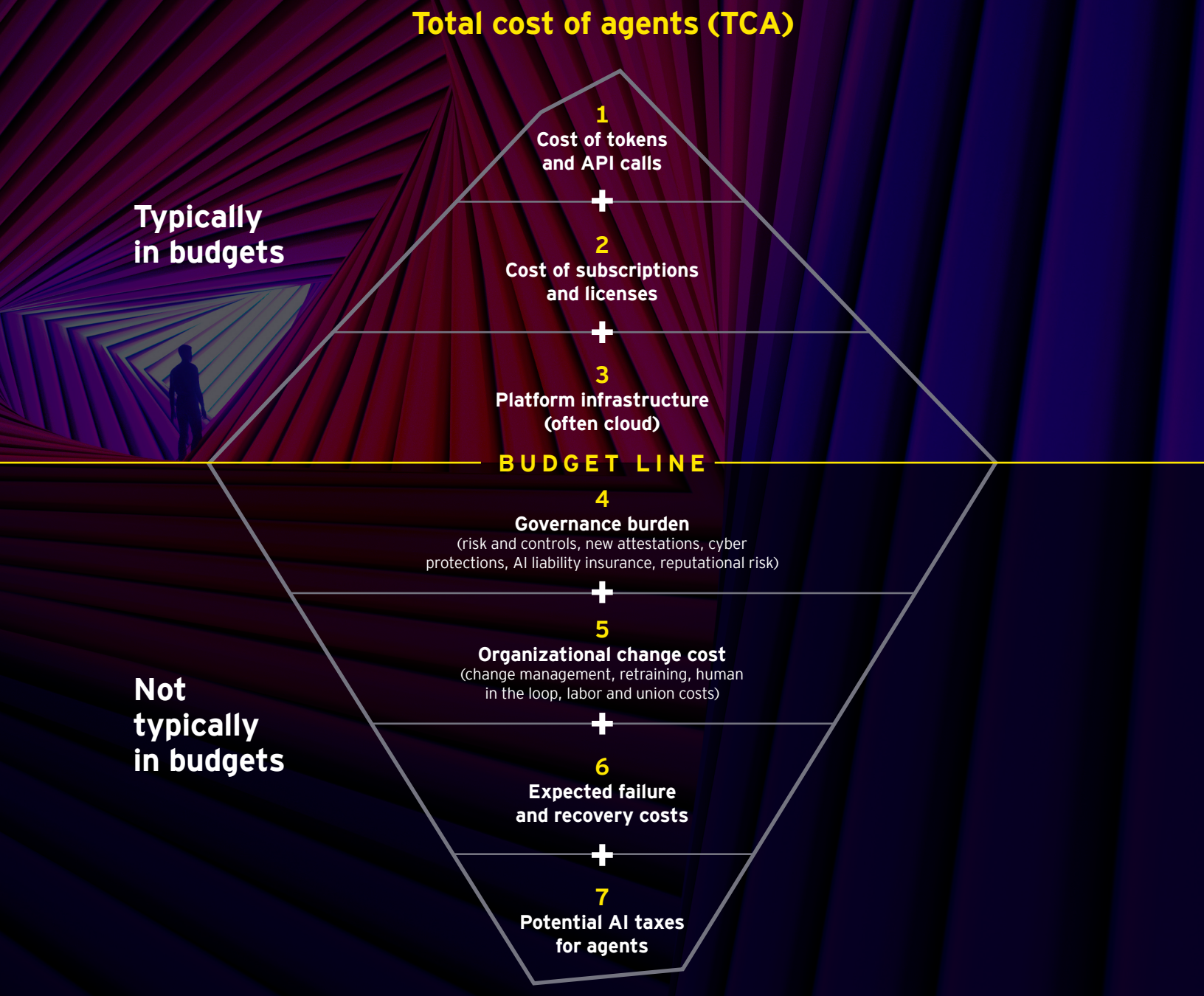
**40%**

of agentic projects are predicted to be canceled by the end of 2027 (Gartner).

[1] "Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027," *Gartner.com,* https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

# How to break down the cost of an agent

The agentic cost problem is twofold: a lag problem, where the true cost reveals itself after the work is complete; and a fragmentation problem, where the true cost resides over budgets that rarely get fully reconciled. **The total cost of an agent is not one item on an invoice; it is seven.**



## How to break down the cost of an agent

**Most companies only include costs 1-3 in their agentic investments and business cases, with costs 4-7 often emerging later in the lifecycle as agents scale.**

### 1. Tokens and API calls

The cost that shows up directly on the invoice; i.e., input and output token volume, model selection, reasoning intensity and retries. Agentic workflows can consume hundreds of thousands of tokens in a single session, much more than the hundreds needed to support a more traditional generative AI chat experience.

### 2. Subscriptions and licenses

The fixed commitments made before any agent runs; i.e., model provider contracts, SaaS and Orchestration platform licenses, committed-use agreements with cloud and AI vendors.

### 3. Platform infrastructure

The compute and services that keep agents running but never appear on the model vendor's invoice; i.e., orchestration runtimes, sub-agent steps, application environments. This typically lands on the cloud consumption bill and gets filed as infrastructure. As agents scale, GPU availability, power and architectural choices become inputs to manage rather than assumptions to inherit. This is a theme that the next papers in this series will return to.

### 4. Governance burden

The incremental investment required to keep agents safe, auditable and compliant; i.e., guardrails, cyber protections, new attestation procedures and human-in-the-loop reviews. These vary depending on the nature of the industry regulatory landscape and often manifest themselves as different teams doing "work" against the initiative. This cost compounds as agents scale out but becomes more predictable when designed in from the start.

### 5. Organizational change

The cost of the transformation to reorganize the operating model around the new agentic workflow; i.e. change management, workforce retraining, role-redesign, human-in-the-loop architecture, labor relations, and in some industries, union negotiations. These investments are front-loaded against each workflow an agent enters, but then often recur with every major model upgrade or capability change.

### 6. Expected failure and recovery

The probabilistic cost of agent failures, including hallucination remediation, back testing, reputational damage and poorly governed agents without kill switches. These costs are zero until they are significant, and most organizations are only beginning to plan for expected loss or insurance. Black swan failures, such as a coding assistant deleting a core database, are too catastrophic to price and must be designed out entirely, via hard-coded boundaries, write-protected environments and human-in-the-loop gates for irreversible actions.

### 7. Potential AI taxes for agents (speculative)

The regulatory cost that does not exist yet but has already been signaled. Various governments and regulatory bodies are exploring AI-specific reporting obligations and potential tax on AI agents as a result of job loss. This carries a major question mark today. Price volatility, capital intensity and concentration risk could ultimately reshape how AI is governed, priced and consumed, a shift a future paper in this series will examine in depth.

## How to break down the cost of an agent

The table below is our working framework for agentic FinOps, illustrating how agent costs are organized and evaluated.

**The key challenge:** The total cost of an agent, at most organizations, is structurally invisible until designed for visibility. This is not because the costs are hard to capture, but because they surface irregularly across uncoordinated budgets.

| | Cost | Where it hides | How it moves | When you know | Budget owner |
|---|---|---|---|---|---|
| 1 | Tokens and API calls | Invoice — model vendor | Highly variable based on usage | At the end of the month after the work is done | CIO/CTO/BU leader |
| 2 | Subscriptions and licenses | Invoice — multiple vendors; visible but fragmented | Fairly predictable, based often on license count or tiered consumption patterns | Fairly constant and predictable | CPO/CIO |
| 3 | Platform infrastructure | Cloud bill – often filed under infrastructure | Step-fixed by tier, variable on top; rarely goes down | Before spend, can be estimated within a range | CTO/CIO |
| 4 | Governance burden | Risk and compliance budget – often headcount | Compounding, every agent widens the surface; limited economies of scale. | Baseline is scopable; compounding over quarters | CRO/CAE |
| 5 | Organizational change | Workforce budget, HR, L&D and consulting | Front-loaded per workflow | Initial cost is plannable; recurrence is triggered by someone else's roadmap | COO/CHRO |
| 6 | Expected failure and recovery | Nowhere – until the incident | Zero until it isn't. Probabilistic moves with scale of agents | After the damage is done | CFO |
| 7 | Potential AI taxes for agents | Doesn't exist yet; regulatory signals only | Unknown | When/if regulation lands | CRO/GCO/CFO |

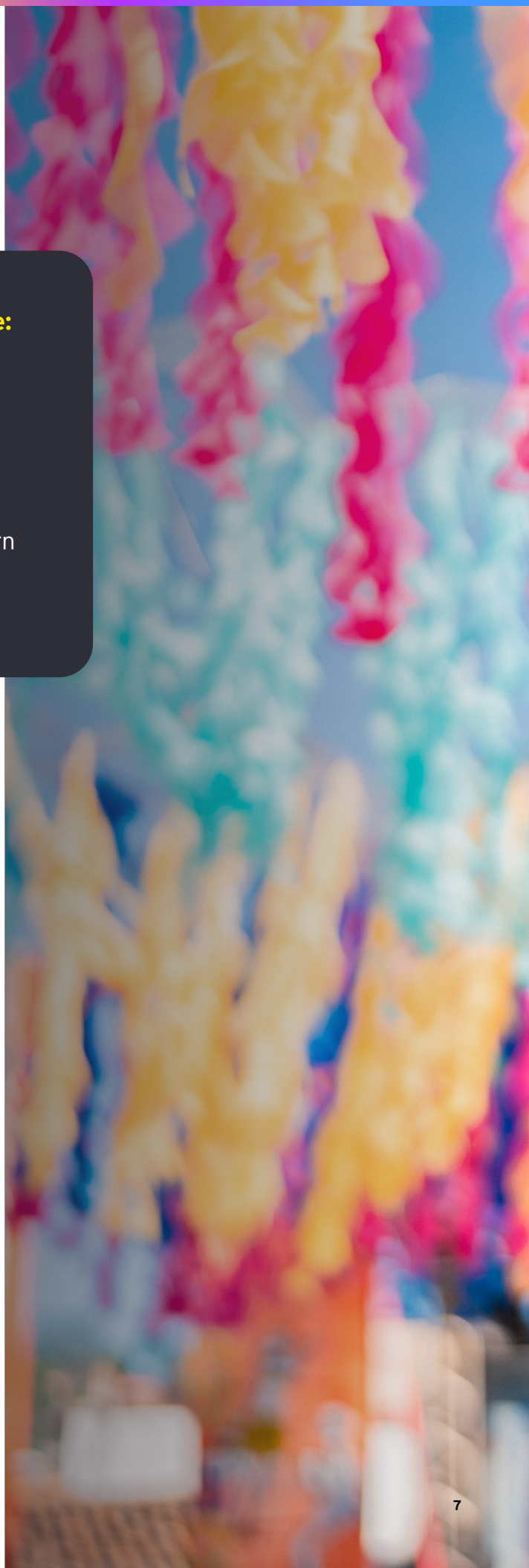
# How to break down the cost of an agent

**The practice of agentic FinOps does three things at once:**

- It allows enterprises to estimate the costs, including lag.
- It assigns an owner to each row **before** the spend, not after the invoice.
- It gives leadership the basis to decide which initiatives earn the right to scale based on the value they unlock against their fully loaded cost.

Referencing Table on page 6

Two years ago, a developer's AI assistant was a fixed-price seat. Today it is an agent that plans, retrieves, uses tools and hands off to other agents. The seat license and model contract are still there (2), but they now sit on top of metered token consumption; in the projection of one regulated entity in the finance sector, nearly a third of developers exhaust their monthly token allotment early (1). Underneath that runs a routing, retrieval and orchestration platform that lands on the cloud bill rather than the model invoice (3). Around it sits the governance burden of making agent output safe to ship, (4) and the organizational cost of redesigning roles, retraining engineers and standing up the human-in-the-loop reviews the workflow now depends on (5). And because an agent that can change production code is a different risk profile than one that can only chat, the program has to budget for the rare failure that gets through the controls (6), with an emerging regulatory multiplier on top of all of it (7). Read the bill alone and the program looks expensive; read it against the value it unlocks and a different investment thesis emerges.



# What leaders can do now

The pace of the market does not allow for a year of observation. **Three key priorities** will help put most enterprises on sharper footing within a single planning cycle.

1

## Appoint a Head of Agent Economics

Centralize accountability under a single executive. Agentic AI costs cut across budgets, so enterprises need clear ownership for model usage, cost leakage and value realization. Whether led by a Head of Agent Economics or Agent FinOps Lead, the focus should be visibility across the seven line items of AI and cloud spend, with total spend management treated as both a KPI and a capital allocation decision.

2

## Install agentic circuit breakers before scale

Benchmark current operations on a per-task or per-outcome basis to establish a defensible view of what good and bad look like. Today it is almost impossible to answer what a unit of agentic work consumes or what stops it when the agent runs away. Then, install hard kill switches: spend ceilings, call-volume caps and automatic shutoffs at the agent, workflow and BU level. Without circuit breakers, the bill is only visible after the damage is done.

3

## Embed full TCO in the business case upfront

Incorporate both capital expenditures (CapEx) and operating expenditures (OpEx) into the business case from the outset. Most agentic initiatives underestimate cost by isolating model pricing from the broader operating envelope. Capture infrastructure, orchestration, monitoring and human-in-the-loop costs to create a transparent view of total cost of ownership (TCO) and force a direct comparison against the enterprise value expected from agents.

# The capital cycle where agents get weighed instead of counted

**The era that exploration is giving way to is one that rewards discipline.**

Investment numbers are only meaningful when tied to value. Every agent should have a value metric on day one: output, revenue, productivity, speed, quality, risk reduction or decision improvement it is expected to produce. No agent should be approved, scaled or renewed without a clear measure of what it produces per dollar.

That requirement moves the discussion from spending to return. It gives the Head of Agent Economics a practical basis to shift investment away from agents that consume resources and toward agents that create compounding value over time.

The next AI cycle will not be won by companies with the most agents, the best models or the largest spend. It will be won by those who treat agent capacity as capital and allocate it deliberately, like any other investment driving growth, margin and enterprise value.

# What's next in the EY Total Cost of Agents series

This whitepaper is the first edition of the EY Total Cost of Agents series. Other papers in this series will examine:

- The path to break-even and the pivot from productivity gains to growth
- The strategic management of compute as a critical supply chain asset
- The policy and regulatory landscape that will shape how AI is governed, priced and consumed in the years ahead

# Authors


**Shawn Smith**
EY Americas Strategy, Innovation and Client Service Managing Partner

**Steve Wanner**
EY Americas Commercial Industries Vice Chair

**Whitt Butler**
EY Americas Vice Chair – Consulting

**Dan Diasio**
EY Global Consulting AI Leader


**EY | Building a better working world**

EY is building a better working world by creating new value for clients, people, society and the planet, while building trust in capital markets.

Enabled by data, AI and advanced technology, EY teams help clients shape the future with confidence and develop answers for the most pressing issues of today and tomorrow.

EY teams work across a full spectrum of services in assurance, consulting, tax, strategy and transactions. Fueled by sector insights, a globally connected, multi-disciplinary network and diverse ecosystem partners, EY teams can provide services in more than 150 countries and territories.

**All in to shape the future with confidence.**


EY refers to the global organization, and may refer to one or more, of the member firms of Ernst & Young Global Limited, each of which is a separate legal entity. Ernst & Young Global Limited, a UK company limited by guarantee, does not provide services to clients. Information about how EY collects and uses personal data and a description of the rights individuals have under data protection legislation are available via ey.com/privacy. EY member firms do not practice law where prohibited by local laws. For more information about our organization, please visit ey.com.

Ernst & Young LLP is a client-serving member firm of Ernst & Young Global Limited operating in the US.

©2026 Ernst & Young LLP.
All Rights Reserved.



US SCORE no. 31213-261US
2510-12389-CS
ED None



This material has been prepared for general informational purposes only and is not intended to be relied upon as accounting, tax, legal or other professional advice. Please refer to your advisors for specific advice.


**ey.com**