

Capgemini
RESEARCH INSTITUTE

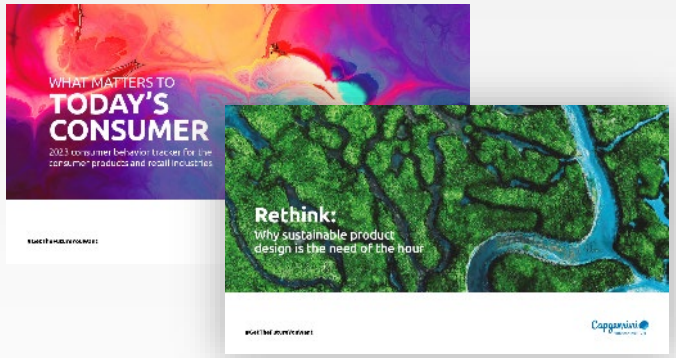
The cost of intelligence: Governing AI economics in the agentic enterprise

Research Snapshot

This publication is part of our *snapshot series*

Snapshots are short notes to discuss market evolutions with targeted primary research

Research reports



Long-form publications with detailed actionable recommendations:

- 10,000–15,000 words
- In-depth global surveys
- 15–30 focus interviews with senior executives sharing best practices
- Detailed recommendations

Research snapshots



Short, sharp pieces with insights on the latest market trends:

- Visual storytelling
- ~1,500 – 2,000 words
- Short surveys
- Key implications and recommendations

Journal: *Conversation for tomorrow*



Our flagship journal with curated content from CXOs, policymakers, Nobel laureates, top academics, and startups:

- Executive and CXO interviews
- Multi-perspective essays
- Infographics and visual summaries
- Trend overviews

Company Confidential © Capgemini 2026. All rights reserved.

Executive summary

The economics of AI can no longer be managed as a technology bill. As organizations scale agentic AI, they need to connect consumption, architecture, human oversight, risk, and successful business outcomes.

- **Scaling AI has created a cost control problem.** Nearly three in five organizations exceeded their AI budgets, with spending running 24% over budget on average among those affected.
- **Agents will add to the pressure.** Agentic AI use is multiplying token usage (token usage grew 14x in six months), and only about half (56%) of organizations are confident in forecasting how AI costs will evolve as adoption scales.
- **Organizations lack an outcome-level control plane.** Six in ten lack an enterprise-wide view of AI costs, and human oversight and rework are the least measured. Even firms with established tracking can link only 45% of AI spend to verified value, and just one in five calculates AI Yield.
- **Economics must be engineered into the AI stack and workflows.** Architecture choices, such as model routing across a multi-model portfolio and the use of open-weight models, can significantly contain AI costs. First-pass success rate is set to become the top-priority AI metric.



This snapshot is based on a survey conducted in September 2026 of 500 CXO-level executives from organizations with annual revenues above USD 1 billion, across 10 countries and 13 industries.

What should organizations do?

- Measure the full cost and value of every AI outcome
- Strengthen governance to prevent AI costs from skyrocketing
- Engineer AI workflows for cost, quality, and reliability
- Adopt a comprehensive strategy to optimize AI Yield
- Promote a culture of responsible experimentation and value realization among the workforce

Company Confidential © Capgemini 2026. All rights reserved.

01

Agentic scale is turning AI economics into a C-suite control-plane challenge

Nearly three in five organizations overshot their AI budgets, with costs running 24% over budget on average

~60%

of surveyed executives say their organization's AI-related expenditure **was significantly higher than expected or budgeted** during the past 12 months.

24%

Among organizations that experienced an AI-related cost overrun in the past 12 months, expenditure was, **on average, 24% above expectations or budget.** One in five organizations (21%) overshot their budget by more than 30%.

> **"The rate of change in AI consumption is unlike anything we have managed before."**
>
> Arvind Joshi, COO & CFO of Global Technology, JPMorgan Chase

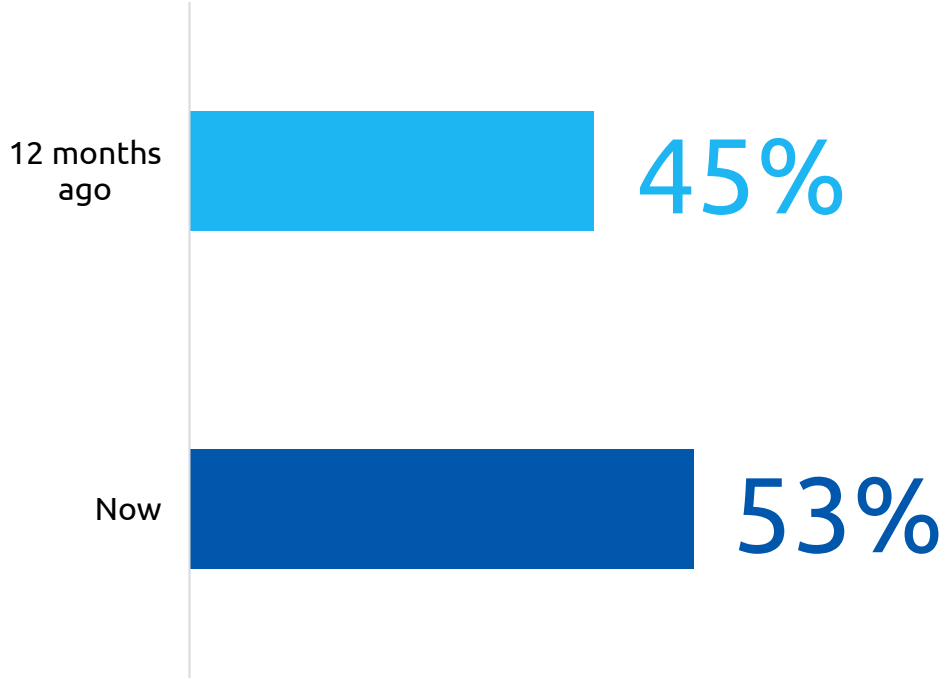
Two-thirds of organizations that overshot their AI budget **faced negative financial consequences,** such as lower free cash flow or a reduction in discretionary spending

Source: Capgemini Research Institute, AI economics survey, September 2026, N - 500 senior executives. Tech Times, "Tokenomics Foundation Launches," August 2026.

Company Confidential © Capgemini 2026. All rights reserved.

53% of executives now worry about the cost of AI

Percentage of executives who are concerned about the overall financial cost of AI



Source: Capgemini Research Institute, AI economics survey, September 2026, N = 500 senior executives who rated their concern as 5-7 on a scale of 1=Not concerned at all to 7=Extremely concerned.

Source: The Stack, "Tokenomics Foundation has big hitters, complex challenges to tackle", August 2026; The Information, "Uber CTO Shows How Claude Code Can Blow Up AI Budgets", April 2026

In the past few months, the fast-changing dynamics of AI economics have caught many organizations off guard, disrupting budgets and forecasts.

Only 56% of organizations are fully confident in their ability to forecast how AI costs will evolve as adoption scales

"I'm back to the drawing board, because the budget I thought I would need is blown away already"

Praveen Neppalli Naga, CTO, Uber

Company Confidential © Capgemini 2026. All rights reserved.

Agentic workloads increase consumption, making architecture and outcome economics critical

Agent design can raise token costs

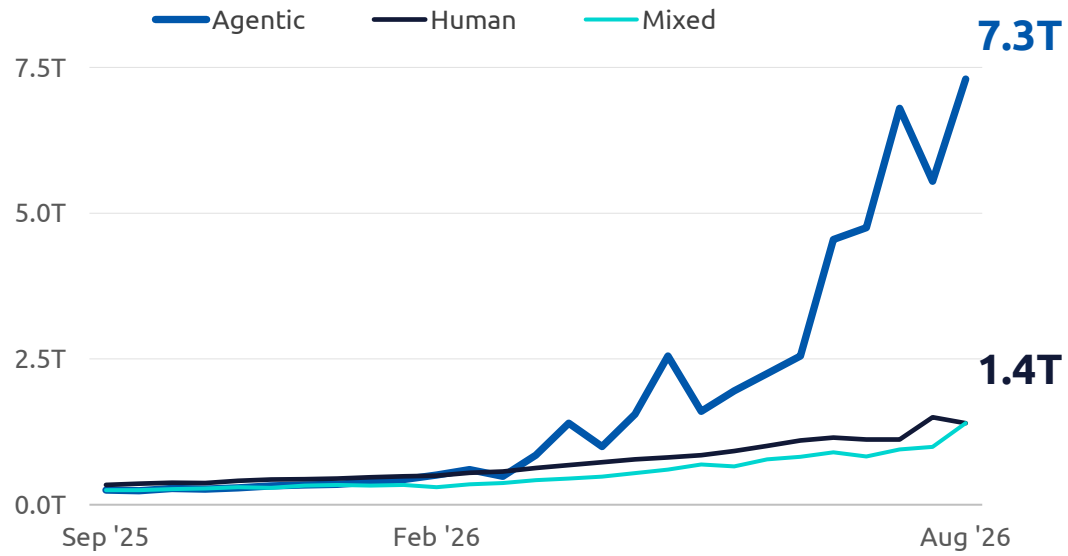
> *Token costs have fallen drastically. But with more usage, and depending on how you design the agent, costs can rise substantially."*
>
> - Bernd Meurer, Chief Technology Officer, Field CTO Europe, BT Group

Source: Capgemini Research Institute interview

> *The amount of infrastructure needed for an agent is meaningfully higher than for a chatbot."*
>
> - Jeetu Patel, President & Chief Product Officer, Cisco

Agentic token use at one AI gateway company grew 14× in six months – human use only 2.8×

OpenRouter, 7-day average token use by request type (trillions)



Agents are uniquely expensive: agentic coding tasks use ~1,000× more tokens than code chat and the same agent on the same task can vary up to 30× in cost.

Source: Stanford Digital Economy Lab, "How are AI agents spending your tokens?", May 2026; OpenRouter data via The Decoder, "AI is becoming AI's biggest customer as agentic token usage jumps 14x on OpenRouter", August 2026; Financial Times, "'We created a monster': companies rein in AI usage as costs strain budgets", June 2026

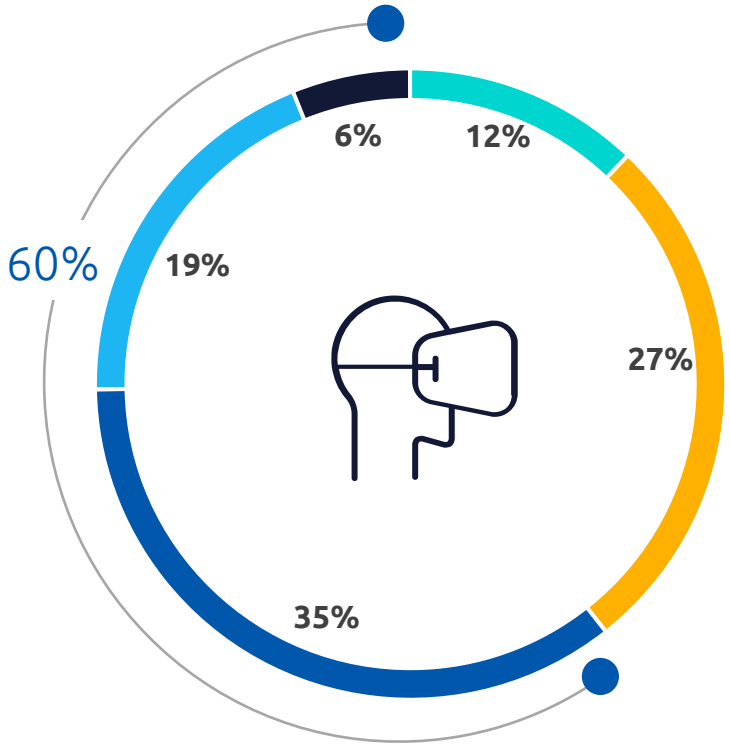
Company Confidential © Capgemini 2026. All rights reserved.

02

Most enterprises cannot see the full cost of AI

Six in ten organizations lack a complete view of AI costs

- **Only a third of all** executives and **one in four CFOs** are fully confident of estimating indirect AI costs.
- **1 in 2 organizations** approve AI spending without adequate visibility into the full cost
- **$500 million:** AI usage bill of one company in one month after failing to set usage caps



Percentage of executives who agree with the following by how effectively their organization can account for AI costs

- We can track most or all AI costs and value across the enterprise
- We can track AI costs across major heads such as technology, business, and operations for a majority of the organization
- We can track AI costs for major initiatives but not consistently across the enterprise
- We track direct AI tech spending but not indirect costs such as governance, human oversight, and rework
- We do not have systems and processes to track AI costs

Source: Capgemini Research Institute, AI economics survey, September 2026, N – 500 senior executives; Forbes, "AI costs more than the people it replaced," July 2026.

Company Confidential © Capgemini 2026. All rights reserved.

Human oversight and rework are the largest AI cost blind spots

Direct costs

Indirect costs

AI cost components	Organizations actively measuring this cost	Among those measuring it, organizations actively optimizing this cost
AI models, applications, and consumption	79%	65%
Cloud, compute and infrastructure	81%	70%
Data preparation and management	67%	58%
Integration and workflow orchestration	61%	55%
Evaluation, security, risk, and compliance	61%	56%
Human oversight, skills, and workflow change	51%	49%
Ongoing operations, retries, failures, and rework	48%	45%
Energy use and environmental-impact costs	28%	28%

Easily measurable/ being optimized

Hard to measure/ not being optimized

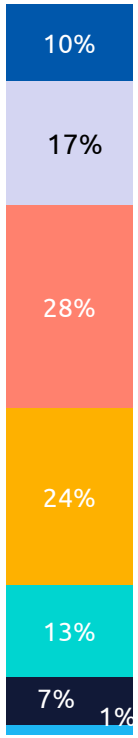
Source: Capgemini Research Institute, AI Economics Survey, September 2026, N = 392 senior executives with responsibilities for AI cost management, finance, AI strategy, AI platforms, AI value realization, procurement, or corporate finance.
Note: Active optimization was assessed only among organizations that measure the respective cost component; bases therefore vary by cost component.

For organizations, the least visible AI costs sit inside the operating model and not the monthly model bill

Company Confidential © Capgemini 2026. All rights reserved.

Only one in five organizations calculate AI Yield

Percentage of executives on which statement best describes how your organization measures AI Yield



- We do not formally measure AI costs or business value
- We track AI costs or usage, but not successful business outcomes
- We track AI costs and business outcomes separately
- We connect AI costs with successful business outcomes, but do not calculate AI Yield
- We calculate AI Yield for selected AI-enabled workflows
- We calculate AI Yield across most material AI-enabled workflows
- Not sure/ Don't know

80%
Do not calculate AI Yield

20%
Calculate AI Yield

Source: Capgemini Research Institute, AI economics survey, September 2026, N - 500 senior executives.

AI Yield captures the verified value of an AI-assisted outcome after accounting for its full safe-delivery cost

AI Yield

=

Verified business value

- What outcome was produced?
- What value was created?

Minus

Full cost to achieve it safely

- Model usage
- Infrastructure & SaaS
- Integration effort
- Human oversight
- Governance & risk
- Rework/error trail

"AI initiatives must begin with clear business goals, whether revenue growth, margin improvement, cost reduction, or better customer outcomes. We answer those questions first, rather than approving the technology and looking for the value later."

- Mikaël Zakarian, CFO, Adobe

Source: Capgemini Research Institute interview

Note: Refer to Issue 01 of The AI Compass from Capgemini for details on AI Yield

Company Confidential © Capgemini 2026. All rights reserved.

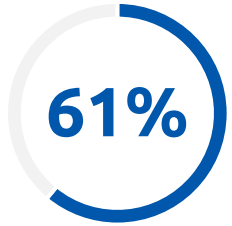
Even firms with established tracking cannot link over half of AI spend to verified outcomes

Among organizations able to estimate expenditure coverage, what share of total AI spending they can track and connect to verified business value.

Total AI expenditure



Share of AI expenditure that can be **systematically identified and tracked currently**



AI expenditure that can be **linked to verified business value**



Share of AI expenditure that **does not ultimately create business value***



Top business value attributed to AI

Productivity gains (cited by 59% of executives), cost reduction (56%), faster processes (50%), and improved customer experience (42%)



57%

say business value is difficult to demonstrate

> "The purpose of token economics is not to minimize token consumption. It is to connect token consumption to value."
>
> - J.R. Storment, Executive Director, FinOps Foundation

Source: Capgemini Research Institute, AI economics survey, September 2026, N = 88 senior executives from organizations with established AI spend-tracking capabilities and with responsibilities in AI cost management, finance, AI strategy, AI value realization, or corporate finance
*This may reflect experimentation, broad employee access, delayed value realization, or failed initiatives. The FinOps Foundation, "Token Economics: The Atomic Unit of AI Value", May 2026

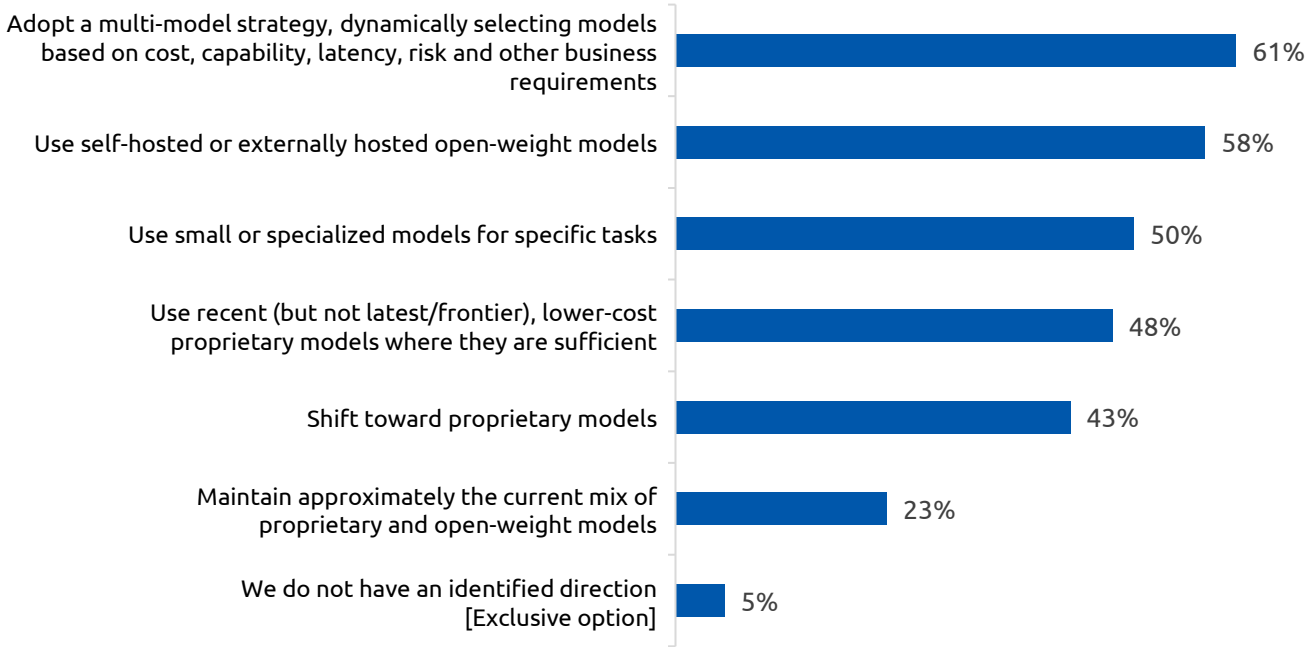
Company Confidential © Capgemini 2026. All rights reserved.

03

How model portfolios change AI economics

Multi-model portfolios are becoming the default architecture for balancing cost, performance and risk

Percentage of executives on how they expect their organization's model portfolio to change over the next 12 months



Source: Capgemini Research Institute, AI Economics survey, September 2026; N = 259 executives working across AI cost management, AI strategy and investment, AI technology and infrastructure, or technology procurement.

"The way you really run up costs is you use the most expensive models with no guardrails around them."

- Katya Andresen, Chief data, digital, and AI officer, Cigna

"The question for the user is, 'what models are best suited [for the job]?' You're not always looking for a frontier model."

- Kyle Daigle, COO, GitHub

Source:
Fortune, "CIOs and CTOs spent years lauding AI. Now, with costs rising, they're putting limits on how it's used", August 2026

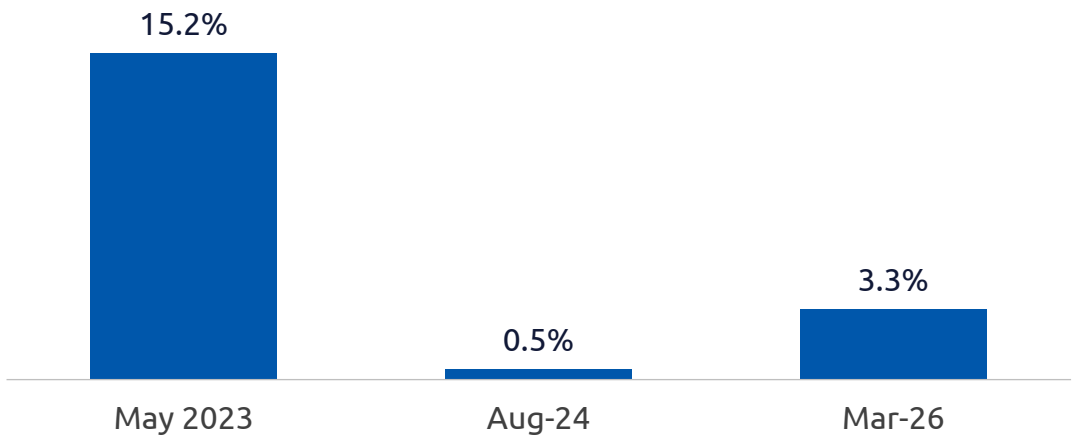
Financial Times, "'We created a monster': companies rein in AI usage as costs strain budgets," June 2026.

Company Confidential © Capgemini 2026. All rights reserved.

Open-weight models trail frontier models by a few months, while costing a fraction as much

The performance gap has narrowed from 15% to about 3%...

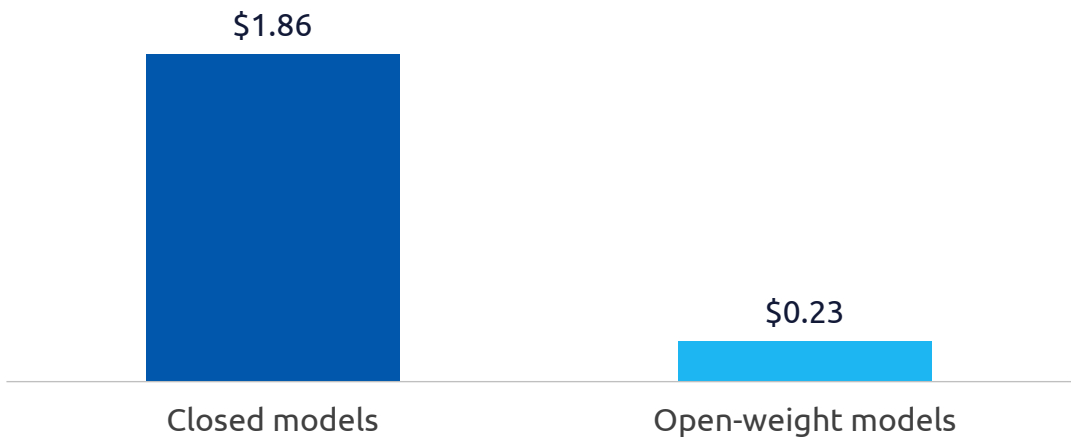
Score gap between the best closed and the best open-weight model on Arena



In July 2026, Chinese model Kimi K3 became the first open-weight model to take the absolute #1 spot overall on Arena – an evaluation platform assessing AI coding performance.

...while open-weight tokens cost 87% less.

Average price per million tokens on OpenRouter, May–September 2025 (USD)



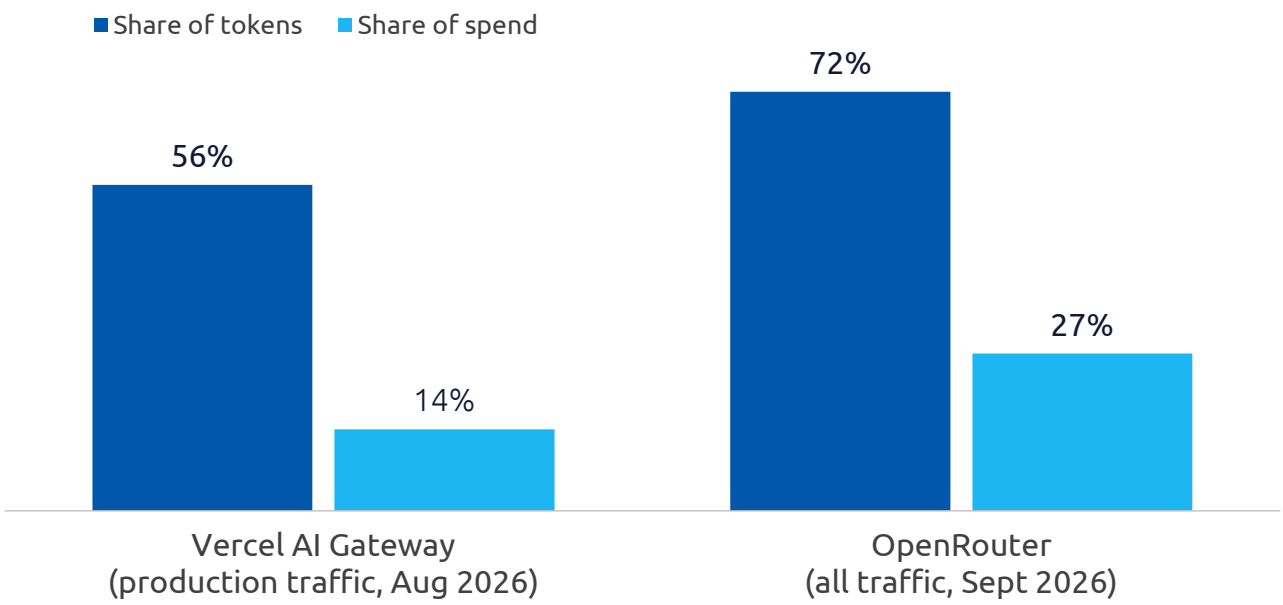
Open models deliver nearly 90% of the capability at roughly 16% of the price

Source: Stanford HAI, AI Index Report 2026; F. Nagle and D. Yue, "The Latent Role of Open Models in the AI Economy", November 2025; Epoch AI, "Open models lag state-of-the-art closed models by 4 months", May 2026; Mozilla, The State of Open Source AI, September 2026

Company Confidential © Capgemini 2026. All rights reserved.

Open-weight models now process most tokens on AI gateways – yet take only 14–27% of spend

Open-weight models' share of tokens vs. share of spend
Traffic through two AI gateways, latest month or week available



Source: Vercel, AI Gateway Production Index, September 2026; OpenRouter, "The Token State of Play", September 2026; Financial Times, "Corporate America embraces cheaper 'open' AI models»", September 2026

AT&T's shift to open-weight models

~40% to 70%
of AI workloads on open models within a year

"We're doing 45bn tokens a day. At that scale, costs become super important. If we can move that traffic from a super expensive closed-source model to an open model that we can often serve at [a fraction] of the cost, and still maintain the accuracy, then we'll do that."

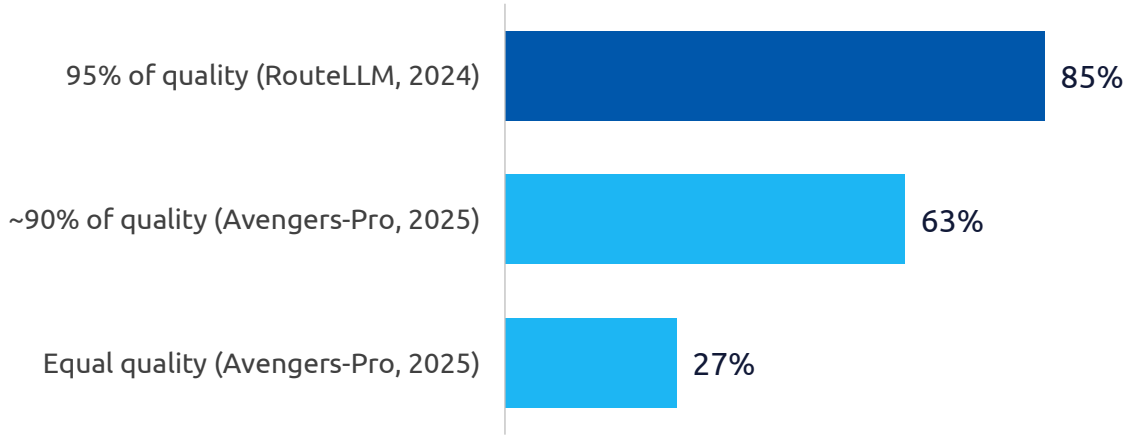
- Andy Markus, Chief Data and AI Officer, AT&T

Company Confidential © Capgemini 2026. All rights reserved.

Routing can cut AI costs by 27–85% at near-frontier quality

What routing has delivered in research

Cost reduction vs. always using the strongest single model, at the quality level shown



Routing turns a model portfolio into AI Yield –
but only if every route is measured on cost, quality and risk.

Source: Y. Zhang et al., "Beyond GPT-5: Making LLMs Cheaper and Better via Performance-Efficiency Optimized Routing", October 2025; NVIDIA, "Route AI Agent Workloads Across Models with NVIDIA NeMo Switchyard", August 2026; VentureBeat, "Nvidia's Switchyard router reshuffles AI models mid-task, cutting task costs to a third in its own tests", August 2026.

NVIDIA NeMo Switchyard model routing in action

LANGCHAIN

74% lower cost

Deep Agents benchmark;
6-point accuracy tradeoff

COGNITION

28% lower cost

Devin Desktop staged routing
with near-frontier coding results

RAMP

58% lower cost

Ramp SWE-Bench matched
frontier results; runtime fell 33%

Company Confidential © Capgemini 2026. All rights reserved.

Decision models add a low-cost tier for routing and checks

What a decision model does

Scores a fixed set of answers in one pass – it writes no text

In:	the context, a question and the allowed answers
Decision model:	one pass that scores every allowed answer
Out:	a typed answer with a calibrated confidence – e.g., “escalate”, 0.91
Used for:	routing requests to the right model or tool, classifying and prioritizing, checking outputs and guardrails, choosing an agent’s next step, and flagging uncertain cases for human follow-up.

	Jev (TypeSafe)	OpenAI Decisions API	Microsoft-Decision-1
Launched	15 Sept 2026	Preview 29 Sept; public beta 6 Oct 2026	9 Oct 2026
Built on	Proprietary “System One” model	GPT-6 Luna	Open-weight Qwen3.5-9B, post-trained by Microsoft
Price per million input tokens (output free)	$0.042	$0.10	$0.042
Vendor speed claim	Up to 194× faster than LLMs on TypeSafe’s own workflows	~10× faster than GPT-6 Luna via the Responses API	~35× faster than GPT-6 Sol (median latency)
Early signals	Fastest-adopted model on Vercel’s AI Gateway: ~13% of paid teams in 24 hours; raised $870M at a $7.5B valuation	Open to all developers; text and image inputs	Xbox Research labeled 10,000+ feedback items ~200× more cheaply than with GPT-6 Sol

The economic advantage comes from using decision models for structured choices and reserving generative models for work that requires deeper reasoning.

Source: Respective company websites

Company Confidential © Capgemini 2026. All rights reserved.

Task-based model routing converts a diverse AI portfolio into outcome-level cost control

Lower costs per AI request do not necessarily reduce total spending. As usage grows and agents perform more steps, organizations need to choose models according to each task's cost and requirements.

A portfolio approach to controlling AI costs without compromising performance

Decision Models
(e.g., Jev and Clef)

Returns clear decisions, and probabilities instead of generating text.

Takes content and typed questions as input, returns selected option and a probability for every option

Fast, low-cost and structured classification.

Model routers

AI models that dynamically select the optimal model for tackling a given task

Assign work based on empirical evidence (cost, speed, quality) rather than fixed practice.

Frontier Models

Closed proprietary LLMs

Maximum reasoning capability for complex, high-stakes tasks

Typically, higher cost of inference, best reserved for genuinely hard, complex tasks

Open - weight Models

High-throughput general models

Customizable and community-driven

Typically provide lower cost inference

Best for handling the routine majority of enterprise queries

Portfolio of SLMs

Targeted Small Language Models tuned for specific domains rather than shrunken generalists.

Superior performance in specific domains owing to targeted training

Efficient, on-device AI for fast inference and lower costs of operation

Company Confidential © Capgemini 2026. All rights reserved.

04

AI economics is shifting from cost to outcome

Most organizations constrain spending rather than redesign AI workflows after an AI cost overrun

Primary responses of organizations to AI cost overruns

53%

Controlled expenditure



- Tightened budgets or controls for the affected initiative
- Introduced organization-wide AI spending controls
- Reduced the initiative's scale or scope

22%

Continued based on expected value



- Continued because the expected value justified the cost

20%

Rethinking tools/processes



- Reassessed which activities should be AI-led versus human-led
- Changed the AI solution, models used, configuration, or model routing

5%

Paused or stopped affected initiative



- Temporarily paused the initiative
- Permanently stopped the initiative

Source: Capgemini Research Institute, AI economics survey, September 2026, N= 297 senior executives who reported that their organization's AI-related expenditure had been significantly higher than expected or budgeted over the past 12 months

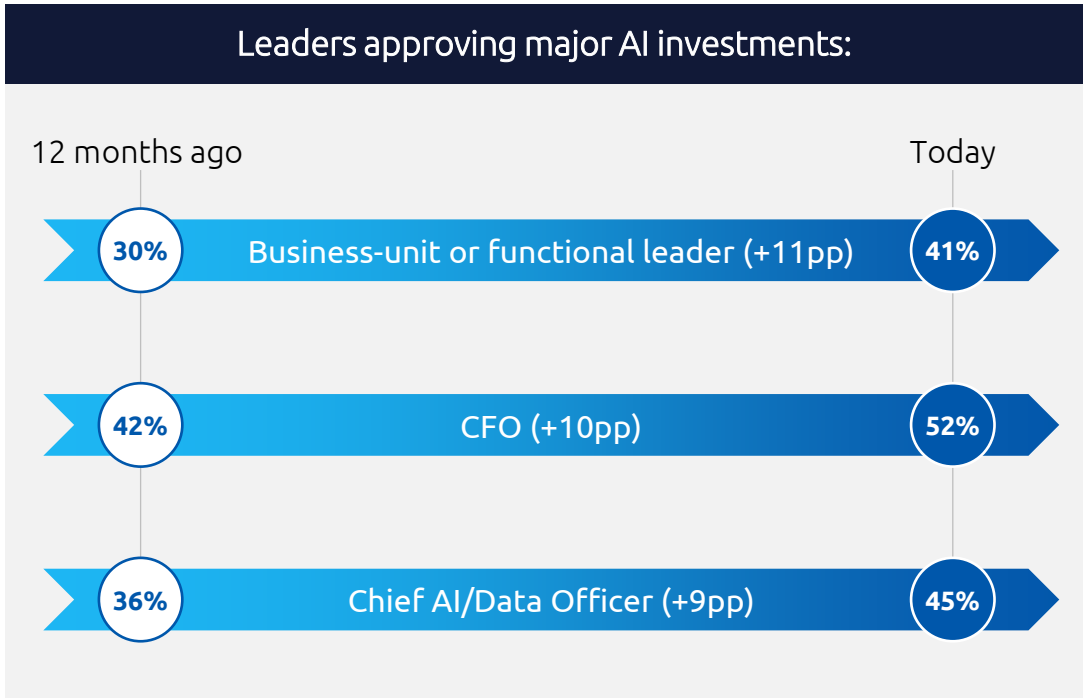
AT&T reduced AI costs by up to 90% using an AI Gateway that intelligently routes workloads to the most cost-effective model while maintaining quality. The platform dynamically balances cost, speed, and performance, delivering savings worth millions.

Source: AT&T, "The Tokenomics Equation: Balancing Cost and Performance", July 2026

Company Confidential © Capgemini 2026. All rights reserved.

CFOs now help approve major AI investments in over half of all organizations

AI investment governance is becoming a shared C-suite responsibility



Source: Capgemini Research Institute, AI economics survey, September 2026, N – 500 senior executives; Thomson Reuters, "AI won't fix broken finance departments — CFOs say the real work is on them", September 2026

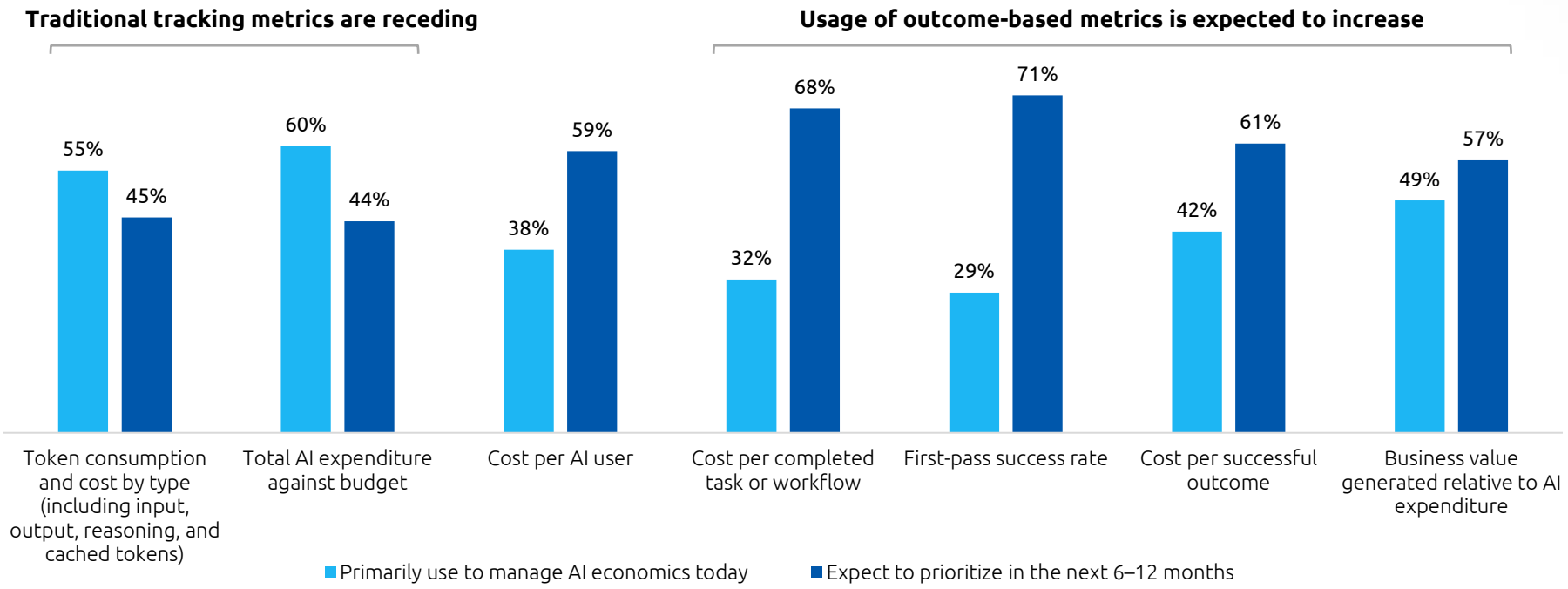
"Finance is finally getting pulled into the room before the decision is getting made, not after. Finance is also being asked to lead things like business strategy now, or tech strategy and data strategy."

- Barbara Larson, CFO, Workiva

Company Confidential © Capgemini 2026. All rights reserved.

Organizations expect to move from controlling AI spend to governing outcomes

Percentage of executives by KPIs used to manage AI economics today, and KPIs expected to be used in the next 6–12 months



> "Buying more seats or using more tokens doesn't tell you much. What matters is whether the work gets done well and what it really costs."
>
> - Sarah Friar, CFO, OpenAI

Source: Capgemini Research Institute, AI Economics Survey, September 2026, N = 368 senior executives from organizations with responsibilities in AI cost management, finance, AI strategy, AI platforms, AI value realization, or corporate finance; OpenAI, "What building an AI-native finance function taught me", August 2026.

Company Confidential © Capgemini 2026. All rights reserved.

Recommendations - Five priorities for governing AI value and scaling economically sound workflows

Measure the cost of AI outcomes

- Assign AI spend to workflows and owners with AI FinOps
- Compare cost per verified outcome with non-AI baselines
- Establish a common AI cost taxonomy

Govern investment and spend

- Set executive owners and shared finance, business and technology budgets
- Use enterprise guardrails to prioritize major investments
- Reinvest productivity gains in growth

Build efficient AI workflows

- Benchmark frontier, open and specialized models on real tasks
- Route work to the lowest-cost model capable of completing the task while reserving advanced reasoning for high-stakes tasks
- Use techniques such as caching and lean context
- Define permitted agent actions and autonomy levels, with risk-based guardrails and human escalation and strengthen AI-ready data and code

Manage a portfolio for AI Yield

- Forecast AI Yield as costs, capabilities and scale change
- Scale, redesign, rescope, or retire investments based on realized value, reliability, and risk

Build a culture of responsible experimentation

- Tailor budgets by user type while protecting innovation
- Reward verified outcomes over AI activity
- Upskill teams to reduce oversight and rework
- Include environmental costs in decisions

Company Confidential © Capgemini 2026. All rights reserved.

Glossary



Frontier AI Models:

The most advanced AI models available at a given moment, trained on massive datasets to deliver state-of-the-art performance across many tasks, representing the leading edge of AI capability.



Open-source AI models:

AI models made available under terms that grant the freedom to use, study, modify, and share the system for any purpose. These freedoms require access to the preferred form for making modifications to the system.



Open weight AI models:

AI models whose final weights and balances – parameters that determine how the model interprets input data and generates outputs – are shared under an OSI Approved License, empowering others to fine-tune, adapt, or deploy the model for their own projects.



Decision models:

Decision models evaluate supplied context against defined questions and return structured answers, often with probabilities. The application defines the permitted outcomes and uses the result for tasks such as routing, classification, scoring, or review.

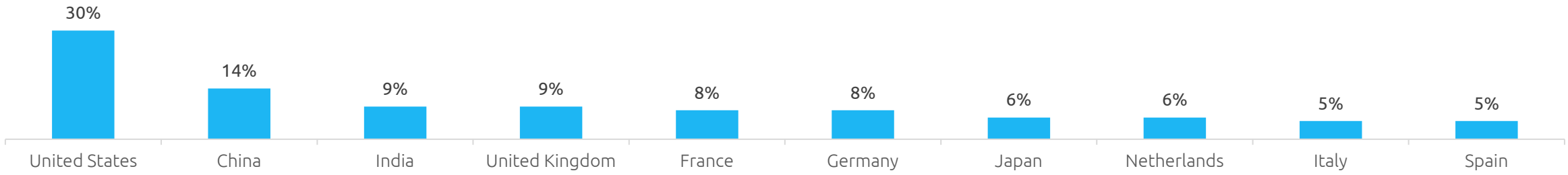
Source: NVIDIA, "How do frontier models work?" Accessed October 2026; Open Source Initiative, "Why we need Open Source Artificial Intelligence," Accessed October 2026; Open Source Initiative, "What are Open Weights?" Accessed October 2026; Vercel, "What are decision models?" October 2026.

Company Confidential © Capgemini 2026. All rights reserved.

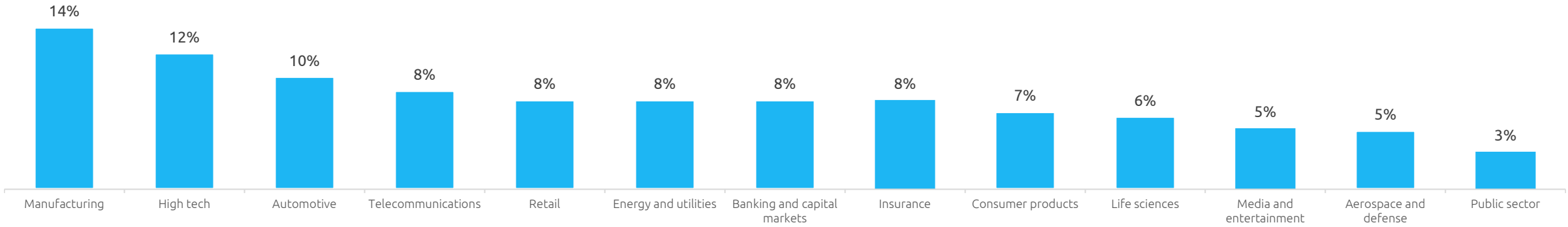
Research methodology (1/2)

In September 2026, we surveyed 500 CXO-level senior executives from organizations with annual revenues above USD 1 billion, across 10 countries and 13 industries. Respondents primarily represented AI strategy and leadership, platforms and engineering, value realization, technology finance, and AI cost management. Respondent profiles are detailed in the following figures.

Senior executives by organization's global headquarters locations



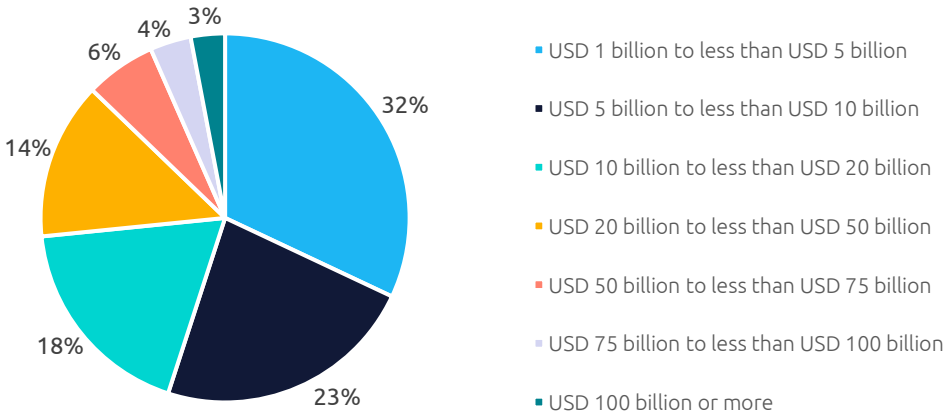
Senior executives by their organization's primary industry



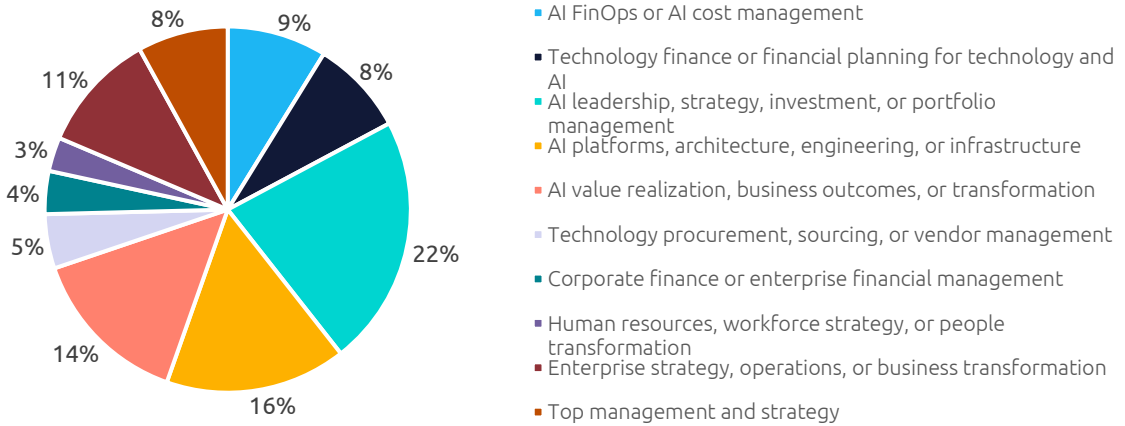
Company Confidential © Capgemini 2026. All rights reserved.

Research methodology (2/2)

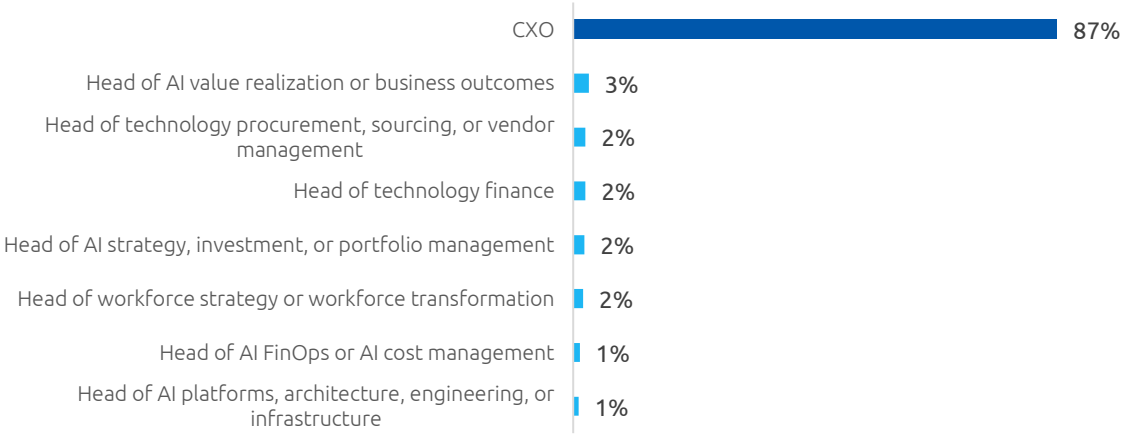
Senior executives by organization's global revenue in the last financial year



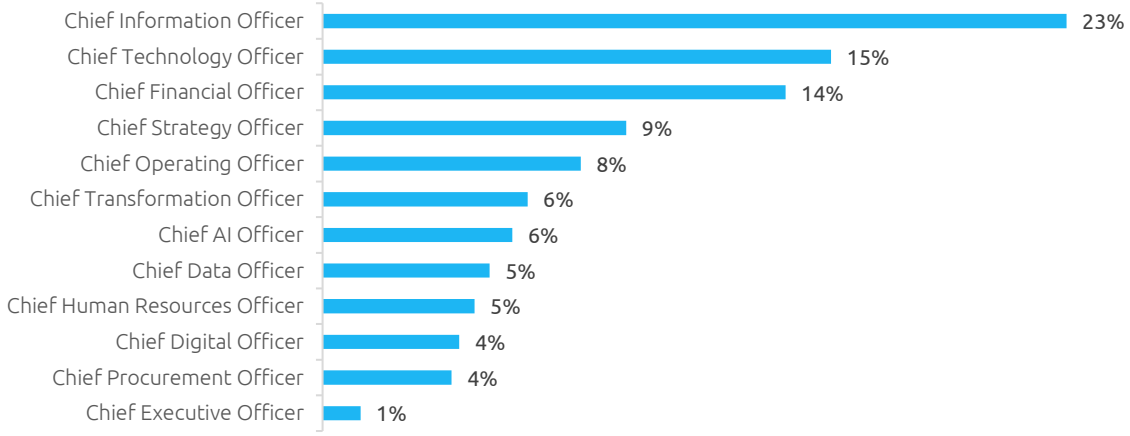
Senior executives by primary area of responsibility



Senior executives by role or title that comes closest to your current role



Senior executives by CXO role or title that comes closest to your current role



Company Confidential © Capgemini 2026. All rights reserved.

Contributors



Pascal Brier
Group Chief Innovation Officer and
member of the Group Executive Committee
Capgemini



Mark Roberts
Head of Capgemini Global AI Futures Lab,
CTO Applied Sciences
Capgemini Engineering



Jez Back
Cloud & AI Economist
and Global Offer Leader
Capgemini Invent



Etienne Grass
Global Chief AI Officer
Capgemini Invent



Craig Suckling
Executive Vice President – Agentic AI - Global
Capgemini

Meet the Capgemini Research Institute



Jerome Buvat
Head of the Capgemini Research Institute
Email: jerome.buvat@capgemini.com



Subrahmanyam KVJ
Senior Director, Capgemini Research Institute
Email: subrahmanyam.kvj@capgemini.com



Amol Khadikar
Director, Capgemini Research Institute
Email: amol.khadikar@capgemini.com



Donald Francis
Senior Manager, Capgemini Research Institute
Email: donald.francis@capgemini.com

The authors would also like to thank Alex Marandon, Dr. Philippe Cordier, Victoire Grux, Mollie Mellows, Sam Connatty, Dominique Banon, and Punam Chavan for their contribution to the report.

Company Confidential © Capgemini 2026. All rights reserved.

Capgemini helps organizations turn every AI dollar into verified AI Yield

Our AI economics offer puts this research into practice. We help leaders measure, engineer and govern AI for the highest yield, not simply the lowest bill.

01
Diagnose

Find where AI creates value, and where it leaks

The AI Economics Stress Test scores each use case on model, routing, human oversight, risk and outcome economics. It produces a CFO-ready resilience summary and a technical layer for the CIO, CTO and CAIO.

Answers: Measure the full cost and value of every AI outcome

02
Engineer

Lower the cost of every successful outcome

Seven engineering levers and a dual-mode architecture: evidence-based routing across frontier, open-weight and domain-specific small models, plus caching, context control and evaluation gates.

Answers: Engineer AI workflows for cost, quality and reliability

03
Govern

Make every AI dollar attributable

Our Token Economics Framework tracks Cost Per Verified Outcome, names an owner for each workflow, attributes every model call, and adds budget-as-code and kill-switch controls in near real time.

Answers: Strengthen governance to keep AI costs in check

THE OUTCOME

AI Yield

The highest verified value from every workflow, every token and every employee

Re-measured continuously as models, prices and workloads change

Explore More

AI Yield = Verified business value – Fully loaded costs of successful delivery

Our framework makes it measurable, connecting Cost Per Verified Outcome with evaluation gates, telemetry and named accountability.

Source: Capgemini Invent, The AI Compass, Issue 01, "The Era of AI Yield: Mastering tokenomics to maximize enterprise value," October 2026; www.capgemini.com/ch-en/insights/research-library/the-era-of-ai-yield-mastering-tokenomics/

Company Confidential © Capgemini 2026. All rights reserved.

Built for the whole C-suite, starting with the AI use cases that matter most

Who	The question we help answer	What Capgemini delivers
CEO and board	Which AI investments are creating value, and which are being scaled on faith?	Our AI Economics Stress Test scores your priority use cases and gives a plain-language verdict on each: economically sound or not, and what would break it. Scale or stop AI investments on evidence.
CFO	Can we attribute, forecast and defend every dollar of AI spend?	AI Yield tracking that measures Cost Per Verified Outcome, names an owner for every material workflow and attributes every model call to a team, in near real time. AI spend you can fund, not just cap.
CIO, CTO and CAIO	Which mix of models and architecture delivers the highest yield?	We stand up the routing layer across frontier, open-weight and domain-specific small models, fine-tune those models on your data inside your perimeter, and build the evaluation gates and telemetry that tie it back to yield.

How to get started

01 **Stress test**
Score your priority AI use cases to see which are economically sound, and what would break them.

02 **Pilot**
Move two or three high-volume, residency-sensitive workloads onto domain-specific small models, with evaluation gates from day one.

03 **Scale**
Track AI Yield in near real time, so your executive committee and engineering teams read the same numbers in the same week.

Go deeper

Read The AI Compass, Issue 01: The Era of AI Yield

Talk to Jez Back, Cloud & AI Economist and Global Offer Leader, Capgemini Invent

Source: Capgemini Invent, The AI Compass, Issue 01, October 2026; www.capgemini.com/ch-en/insights/research-library/the-era-of-ai-yield-mastering-tokenomics/

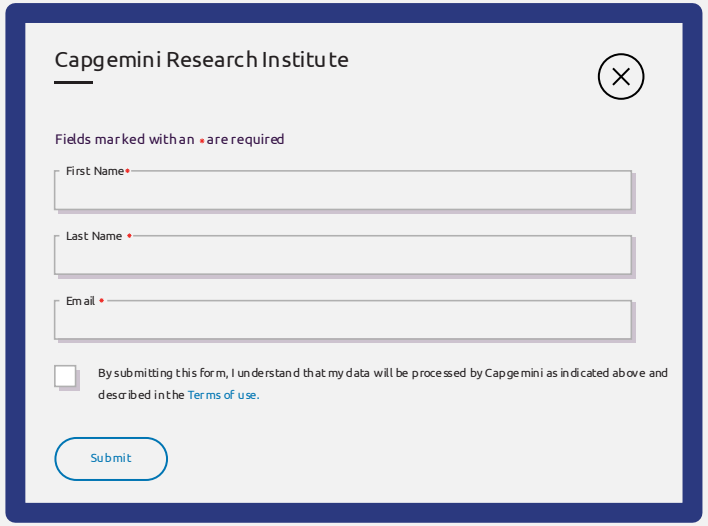
Company Confidential © Capgemini 2026. All rights reserved.

Subscribe to latest research from the Capgemini Research Institute



Receive copies of our reports by scanning the QR code or visiting

https://www.capgemini.com/capgemini-research-institute-subscription/



Company Confidential © Capgemini 2026. All rights reserved.

About Capgemini

Capgemini is the business transformation partner for enterprises in the age of AI. We help organizations imagine and build an intelligent, sustainable future, combining AI, technology and human ingenuity to transform how they operate, innovate and grow. With unique end-to-end capabilities spanning strategy, technology, engineering and intelligent operations, we bring together deep industry expertise and market-leading capabilities in AI, cloud and data to turn ambition into measurable business outcomes at scale. Supported by a robust ecosystem of partners and nearly 60 years of expertise, Capgemini is a responsible and diverse global organization of over 410,000 team members in more than 50 countries. The Group reported 2025 revenues of €22.5 billion.

Make it *real*.
www.capgemini.com

Capgemini

This presentation contains information that may be privileged or confidential and is the property of the Capgemini Group.

Copyright © 2026 Capgemini. All rights reserved.