Can AI pay for itself?

The true cost of AI and the value required to justify it

EY Total Cost of Agents series | Paper 2

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

EY

Shape the future with confidence

The token is just the entry fee

1/3

of the cost to run an agent comes from the token.

When we began the Total Cost of Agents series, our first objective was to develop an estimation model for the costs incurred after the first tokens were spent by an AI.

Seven cost lines surface on the final bill of an agent, starting with tokens, then including subscriptions, platform infrastructure, governance, organization change, expected failure and emerging regulation. Our first paper concluded that the token cost, which dominates the preoccupation of most business leaders, is just a portion of the total AI bill. That finding points to a broader imperative: **Organizations need a new enterprise capability that governs agent capacity like capital.**

In this paper, we analyze the supply chain that must be financed and operated before a single token is consumed by an agent. This costly supply chain is what sets the price that organizations must reconcile as we consume more tokens, sparking the question: What value has to be created for AI to pay for itself?

The question has become urgent. AI economics are no longer hidden or subsidized. Introductory pricing and investment-funded usage are giving way to explicit cost recovery. As AI suppliers face compute commitments, debt financing and public-market scrutiny, cost is moving downstream toward the enterprise customer. And the token is only the beginning of the bill.

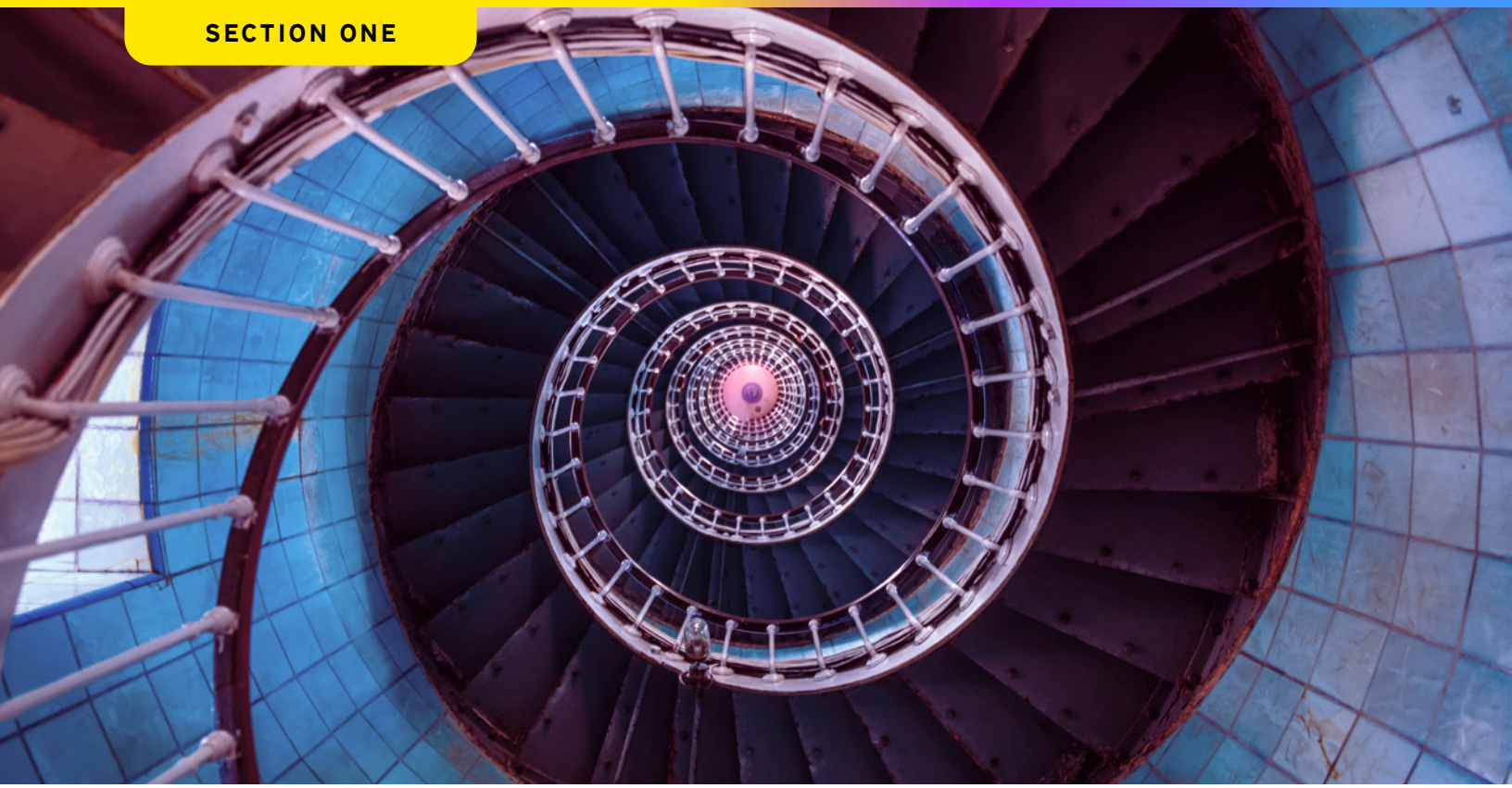
Companies that lead with AI will direct investment toward value pools large enough to justify it, rather than treating substitution as a funding strategy.

By our estimate, the token accounts for roughly a third of what an agent actually costs to run. The remaining two-thirds accumulate across the subscriptions, platform infrastructure, governance, organizational change, expected failure and emerging regulations that surround every deployment. Taken together, the full enterprise cost of AI runs to roughly three times the token invoice alone. A cost base of that size means "AI saves time" is no longer a sufficient business case.

This paper starts from the bill and works backward. With the spending already committed, we ask what value has to be created to cover it.

Modeling both sides of that equation leads to an uncomfortable finding. To earn a market return on the projected AI build-out, the white-collar economy must either grow its output by roughly 10.5% by 2031 or shed nearly 15.8% of its labor costs — one dollar in six. This ramps up to one dollar in five by 2035. In a static, single-firm view, growth and cost cutting look like the same destination. They are not. On one company's economics, AI is simple labor-for-capital substitution: trade a payroll dollar for a compute dollar and book the difference. But a wage is also income, and that income is another company's revenue. When one firm cuts labor costs to fund AI, it captures a private saving; when every firm cuts at once, the aggregate wage bill falls and appears as lost personal income weighing on final demand.

That is the difference. Pure labor cost reduction can weaken income and final demand, while productivity-led growth can lower unit costs and expand potential output. For a single enterprise, cost savings may generate a real return on AI and provide a compelling reason to act. But productivity and growth create far more room to maneuver. Across the economy, the build-out cannot be financed through cost cutting alone.



The cost behind the token

Gartner forecasts worldwide AI spending to rise from $1.76 trillion in 2025 to $2.6 trillion in 2026.[1] And this continues its ramp, with current trajectories approaching $9.1 trillion a year by 2035. A decade-long infrastructure build-out, one that is now stretching investors' ambitions toward space for scale, converges on a surprisingly small billable unit: the token.

These figures are only part of the story. Data center and hardware infrastructure is the single largest line, around $1.4 trillion today, yet it is barely more than half of the total. The rest is spread across model development, AI software and application platforms, data and integration, and the security and services that stitch them together. These are the costs that never appear on the token invoice, and they compound.

To bring the full picture of the cost to light, we divide them into these **eight streams.** Each is its own network and business landscape, a complex chain of transactions with a claim to profit at every step:

1. Data centers and facilities capital
2. Accelerators and silicon supply
3. Server, networking and lifecycle
4. Energy, cooling and operations
5. Model development and evaluation
6. Inference-serving runtime
7. Provider orchestration and service layer
8. Risk and compliance

[1] Gartner Forecasts Worldwide AI Spending to Grow 47% in 2026," Gartner.com, www.gartner.com/en/newsroom/press-releases/2026-05-19-gartner-forecasts-worldwide-ai-spending-to-grow-47-percent-in-2026, 19 May 2026.

The infrastructure behind every token

Before anyone consumes a token, suppliers have already aggregated data, infrastructure, hardware, energy, reliability, safety and service operation costs.

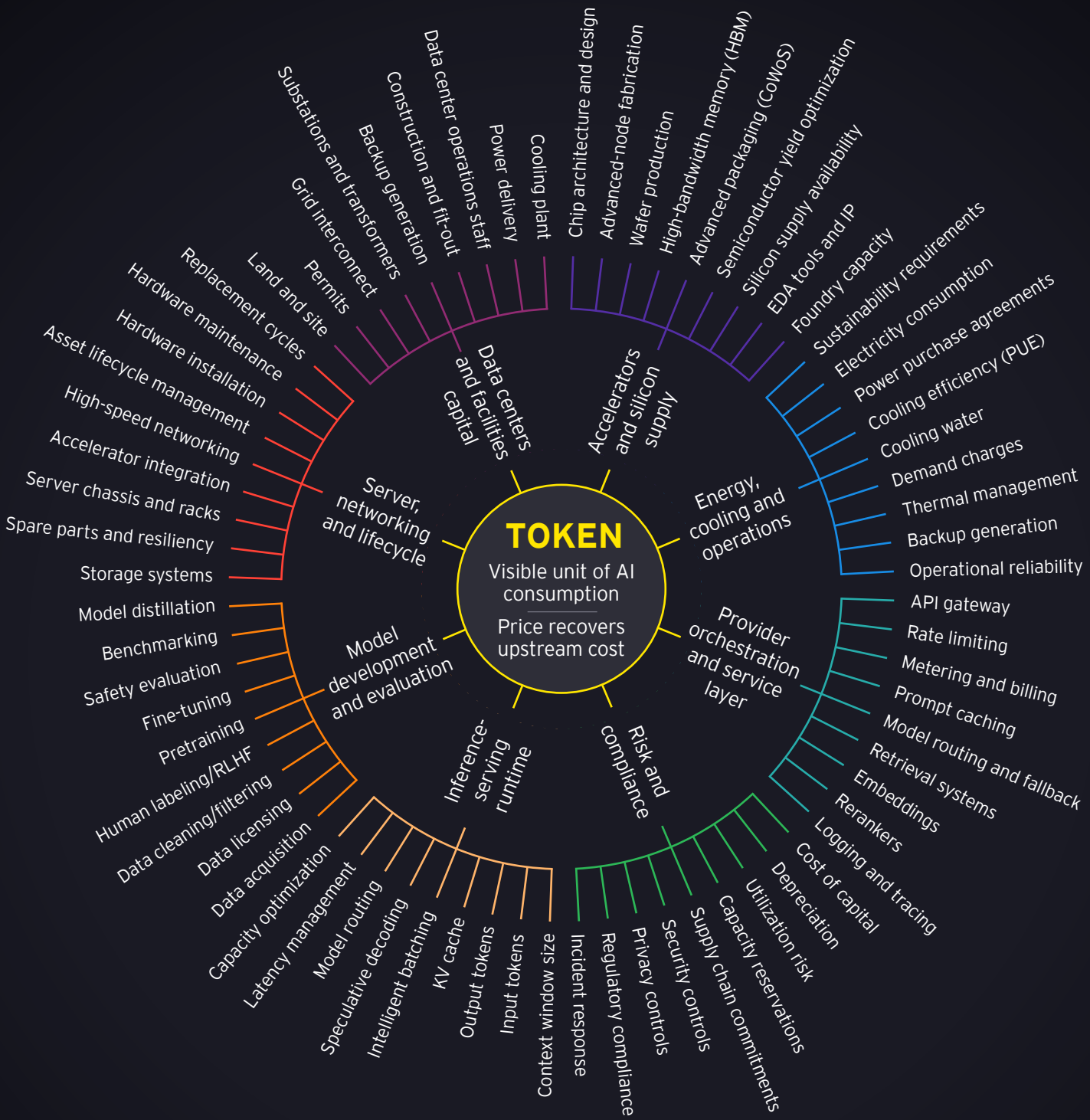
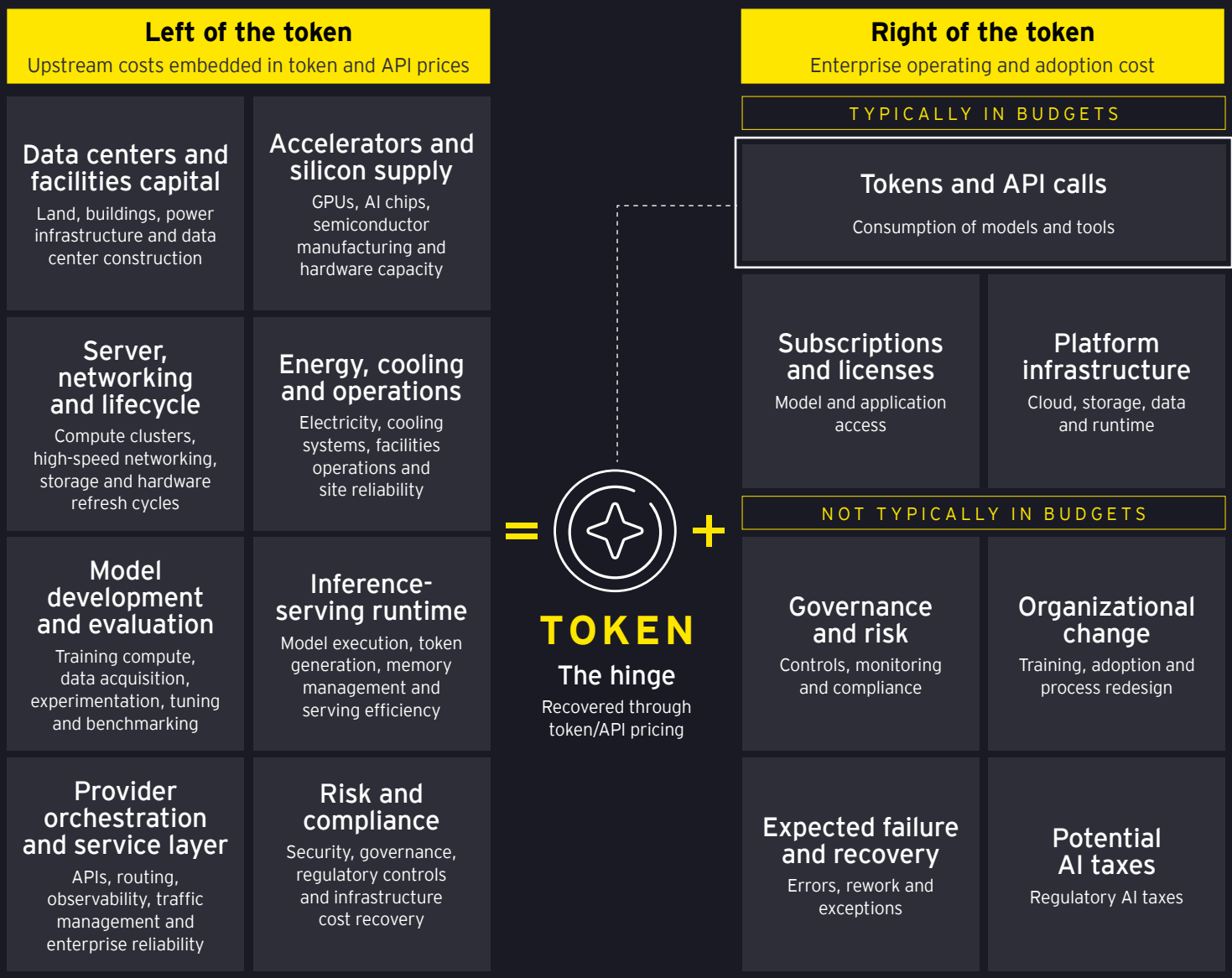


Exhibit 1: The upstream cost stack — the eight cost streams that produce a token

Upstream meets downstream

Upstream supplier costs are embedded in the token price; downstream, the enterprise adds its own operating layer from first paper. The token is the hinge where supplier and user meet.



The price of a token is high as soon as it leaves the silicon, but it's only the start of a long journey across the market. The cloud provider that hosts the workload, the model provider that needs to fund the next generation and often a software integrator all have to clear margins in the journey to the token buyer. Supplier cost is the floor beneath a price that margins push higher. Measure the business case against that supplier cost, and the real bill gets understated.

Supplier costs are recovered through token and API pricing. How much an enterprise uses AI multiplies that price across seats, agents, retrieval calls and agentic processes. That first term is line one of our first paper's seven cost lines, now opened up. The entire upstream stack lives inside the token price.

The enterprise then adds its own operating layer downstream of the token — the seven cost lines from our first paper. Combining upstream supplier cost with downstream enterprise cost, joined at the token, we arrive at a model for the total enterprise cost of AI.

Is AI-driven efficiency enough?

The tempting question some organizations are asking is **if AI replaces human work, how much work must it replace to pay for itself?** This question unfairly reduces AI's value to labor substitution. Still, following the logic through shows just how demanding the cost cutting case would have to be.

In 2026, global white-collar industries' labor cost is estimated at $31 trillion, which grows to $38 trillion in 2031 and $45 trillion by 2035.

Productivity, which is a legitimate driver of enterprise returns, can support higher potential output, income and demand. This paper tests a narrower issue: Whether productivity gains captured solely as cost savings are sufficient to fund the AI build-out across the economy.

The supplier build-out must be recovered through user charges. Those charges are augmented by enterprises, with direct spending on software, services, data and implementation. Add them together with industry-expected recovery lives and an annualized recovery charge under a 12% capital recovery model, and the economy's total AI bill runs from roughly $1.1 trillion in 2026 to $6 trillion in 2031 and to $9.8 trillion by 2035. In the absence of an industry-standard rate of return, capital could be deployed in other areas rather than speculative AI projects.

In a cost reduction scenario, that bill must be covered by reducing the cost of high- and upper-middle-income economies' white-collar labor employer costs. In 2026, global white-collar industries' labor cost is estimated at $31 trillion, which grows to $38 trillion in 2031 and $45 trillion by 2035. Why white-collar industries? These are the sectors most anticipated to benefit from AI without significant additional spending on physical AI adaptations. Finance, professional services, information and other white-collar industries are those most expected to benefit from AI.

The question: How much labor reduction would be required to pay for these costs in the most affected industries?

To earn that result, AI would need to remove 16% of total white-collar industries' labor cost in upper- and upper-middle-income countries by 2031, and nearly 22% by 2035. Or, on the growth side, white-collar output would need to expand by roughly 10.5% by 2031 and 14.5% by 2035. Even holding the return requirement to the capital investment and implementation costs, recovering operating spend dollar for dollar with no return on it, the requirement lands in the high single digits by 2035.

Model at a glance

Values for baseline scenario in which costs are recovered through white-collar industries in upper-middle and high-income countries

Year	2026	2031	2035
Total AI spend ($b)	$2,596	$6,388	$9,068
Required user cost with 12% return	$1,078	$6,003	$9,829
White-collar employer cost	$30,758	$37,904	$44,859
White-collar industries' GDP	$46,603	$57,431	$67,968
Required displacement	3.50%	15.80%	21.90%
Required GDP growth of white-collar industries	2.30%	10.50%	14.50%

The arithmetic is illustrative and is not intended as a prediction. By 2031, paying a market return on the AI build-out means removing one in six dollars of white-collar labor cost or growing white-collar GDP by 10.5% – by 2035, one dollar in five, or 14.5% growth. And the trade-off only gets steeper from there.

These figures also assume the bill is spread evenly across the full base of high- and upper-middle-income economies. In practice, it may not be. If AI delivers meaningful payback only in the wealthiest economies, the same cost is recovered against a smaller base, and every required share rises accordingly. Narrow the base to high-income countries alone, and the displacement and growth requirements move materially higher. Narrow it again to economies above a given per capita income threshold, and they climb higher still. We offer these as hypotheticals rather than forecasts, but they point in a consistent direction: The more concentrated AI's payback proves to be, the more demanding the arithmetic becomes for the economies and firms expected to carry it.

Income-threshold scenarios

Required displacement/required white-collar GDP growth

Each cell shows required labor displacement/white-collar GDP growth for that year. Baseline is the upper-middle- and high-income country set used throughout the paper; narrower rows progressively restrict the base to higher-income economies. Figures are hypotheticals, not forecasts.

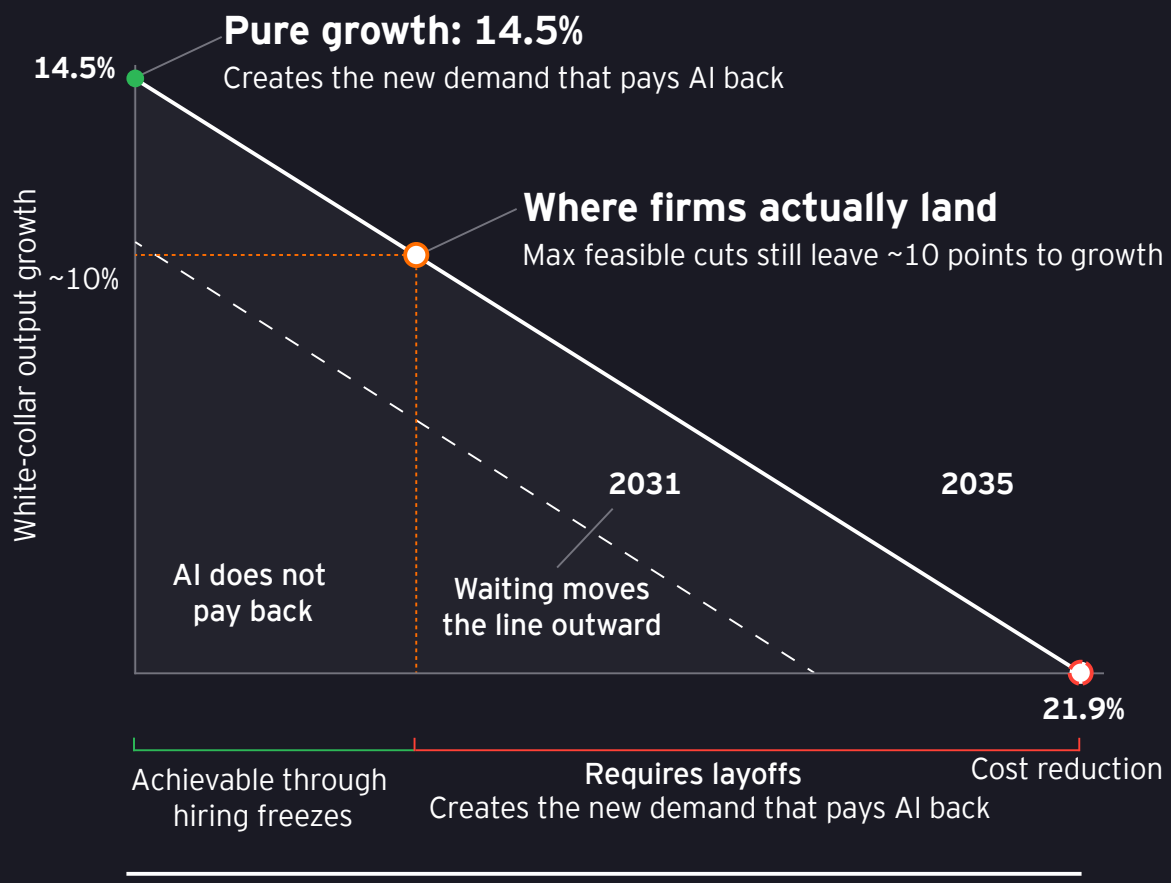
Scenario	2026		2031		2035	
	Labor	GDP	Labor	GDP	Labor	GDP
Baseline: upper-middle & high income	3.5%	2.3%	15.8%	10.5%	21.9%	14.5%
High-income countries only	5.0%	3.3%	23.2%	15.3%	32.1%	21.2%
Greater than $25,000 GNI[2] per capita	5.4%	3.5%	24.7%	16.3%	34.2%	22.6%
Greater than $50,000 GNI per capita	7.2%	4.8%	33.2%	21.9%	46.0%	30.3%
Greater than $75,000 GNI per capita	11.3%	7.4%	52.1%	34.4%	72.1%	47.6%

[2] Gross national income (GNI)

The payback frontier

The frontier defines the range of cost-reduction and growth combinations to earn a 12% on the full cost stack

The frontier shows the combinations of cost reduction and output growth required to earn a 12% return on the full AI cost stack. As projected AI spending rises, the frontier moves outward. At the 6.5% labor-reduction threshold, firms must recover approximately 10 additional percentage points through growth.



Only one end of this line is real — and every year of waiting moves it outward.

The figures presented are EY-developed illustrations based on available research and analysis. They are intended to demonstrate potential outcomes and may vary depending on the specific context and assumptions applied.

At the enterprise level, cost reduction can generate returns. At the economy-wide level, however, labor elimination alone cannot sustainably finance the AI build-out, because broad-based payroll reduction eventually erodes the demand base that funds corporate revenue.

As Keynes observed, a company's payroll is more than just a cost. It is income to households, and that income becomes spending, which in turn becomes revenue for other companies (or even the very company that made the reduction).[3]

Corporate wages paid are therefore demand created for the broader economy. What strengthens one company's balance sheet can weaken the economic system on which the revenues of every company depend.

At an extreme, when the corporate sector in aggregate tries to fund AI by reducing headcount at scale, it risks removing demand from the very economy that generates its revenue. Wages that once fueled consumer spending instead become retained capital, slowing the circulation of value. As AI infrastructure costs come due on a fixed schedule of debt and depreciation, cost-out at scale can become self-undermining. Wharton economists Brett Hemenway Falk and Gerry Tsoukalas make a related point in *The AI Layoff Trap*, arguing that competitive automation incentives can collectively erode the consumer demand firms ultimately depend on.

AI funded by layoffs risks shrinking the customers needed to pay for it.

This is not to say displaced labor simply disappears. History suggests much of it is eventually redeployed into new roles, industries and forms of work, and the same is likely here but with a meaningful and significant transition. The concern is one of timing and structure rather than permanent loss. Debt and depreciation come due on a fixed schedule, while redeployment unfolds over years and rarely maps cleanly onto the skills, locations or wage levels being displaced. In that gap sit real transition costs and a period in which reduced demand and fixed obligations can reinforce one another before the economy reabsorbs the affected workers.

For that reason, no single company can fully opt out of the aggregate arithmetic. If the corporate sector funds AI by cutting at scale, much of the resulting contraction does not stay on one firm's income statement; it eventually reaches the broader top line, even as labor markets adjust over the longer term.

[3] Keynes, J. M. (1936). The General Theory of Employment, Interest and Money. London: Macmillan.

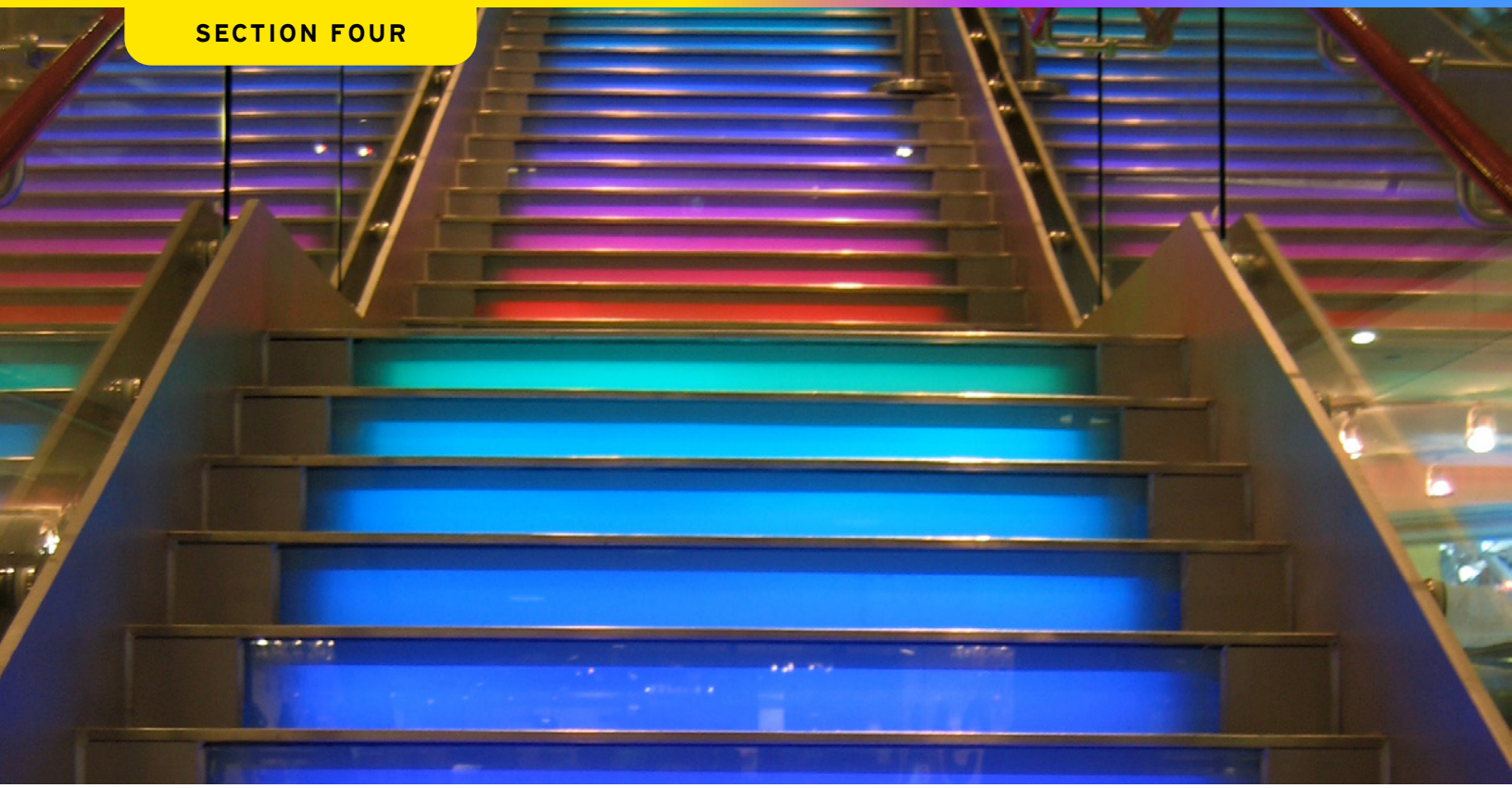
Growth as the only durable economy-wide path to payback

At the level of a single enterprise, cost reduction can generate entirely valid returns on AI, and for many firms they will be the reason to act. But unlocking growth at the same time gives a company many more degrees of freedom, and in aggregate the arithmetic is unforgiving: The economy as a whole cannot finance the AI build-out on labor elimination alone.

This is why treating 14.5% growth and 22% cost reduction as equivalent paths is misleading. They arrive at fundamentally different economic outcomes. The 22% reduction in labor costs represents income that would otherwise become revenue elsewhere in the economy. Implemented at scale, it contracts the economic base. By contrast, 14.5% growth reflects the creation of new output, new demand and new markets, expanding the economic base rather than redistributing or shrinking it. One number shrinks the pie, the other enlarges it.

This applies to growth as well. If the growth comes on the back of a competitor's share, this is a mirror to the cost cutting. Growth only expands when the output is genuinely new.

Consider a regional bank. The obvious AI agenda is efficiency: automating servicing, compressing back-office processing, cutting the cost of each transaction. But suppose its leadership asks a different question: How do we grow the value of every client relationship? That reframes the deployment from a list of use cases into a value stream. Segmentation models map how clients actually behave, analytics turns that into better-targeted advice and agents carry those signals to relationship managers at the moment of decision. It is the same machine-learning-plus-agents stack a cost program would have built, pointed at growth rather than cost to serve.



What leaders should do now

AI must be governed with the discipline traditionally reserved for capital allocation

The token invoice records only the transaction; the cost cascade reveals the economic obligations that precede it and the returns that must ultimately justify it. As leaders shift from a cost-out mindset to a growth one, these three areas are where to start:

- **Initiate value streams over use cases**
  Value streams address the cross-functional constraints that limit enterprise performance, including delays, bottlenecks and fragmented data. This speeds time to market, expands product portfolio, improves customer satisfaction and helps evolve business models. Invest in capabilities that expand what the organization can do, rather than simply making today's processes more efficient.
- **Re-evaluate measurement**
  Distinguish between productivity created, captured and monetized. Time saved is only potential value. It becomes economic value when the operating model turns it into higher output or new revenue, not before. Productivity is the bridge between efficiency and growth, but only if the gain is actually banked rather than just observed.
- **Run a live portfolio**
  Scale, redesign, pause or stop each deployment based on whether its value pool is large enough to carry the weight of the agentic system. Challenge where and how this new capability can be turned on its head toward growth. If it fails the test, decisions to stop are an exercise of investment discipline.

Operationally, this discipline takes the form of agentic FinOps, introduced in our first paper. Give a single executive owner authority over agent economics, including investment, cost, captured value and scaling decisions. We call this new capability an AI Value Realization Office.

The infrastructure build-out is already underway, and its costs are moving steadily downstream. Every enterprise adopting AI will ultimately participate in financing that investment. The organizations that capture the greatest advantage will be those that treat AI as a capital allocation challenge, govern it with financial discipline and expand the value created from every dollar invested.

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

The next paper in this series turns to cost predictability: treating the compute supply chain as critical infrastructure, hardening it, sizing it and deciding what belongs on premises, in smaller models or in reserved capacity. Cost must be made governable.

Authors

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

Hyong Kim
EY Global TMT Leader and EY Americas TMT Vice Chair

Steve Wanner
EY Americas Commercial Industries Vice Chair

Greg Daco
EY-Parthenon Chief Economist, Strategy and Transactions

Dan Diasio
EY Global Consulting AI Leader

Sachin Lulla
EY Americas Industrials & Energy Vice Chair

Andrew Phillips
EY Americas Quantitative Economics & Statistics (QUEST) Leader

Contributors

Lisa Peterson
EY Brand Marketing & Communications Leader, Americas Consulting

Tommaso Babucci
EY Americas AI Strategy and Development Lead

Disclaimer

This analysis assumes that a portion of underlying infrastructure costs may be reflected in AI consumption pricing. The extent of any cost or margin pass-through across infrastructure providers, AI companies and end users is not publicly known and may vary by provider and commercial arrangement. This assumption is used solely for illustrative modeling purposes and should not be interpreted as attributing pricing changes to any individual company or segment of the AI value chain.

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.

US SCORE no. 32150-261US
2605-10326-CS
ED None

ey.com