AI ADOPTION IN U.S. ORGANIZATIONS, 1,000+ EMPLOYEES · AS OF OCTOBER 2026

The AI Invoice Means It's Working. Is It Buying Anything?

Higher usage and higher rates can both increase an AI bill. This issue looks at how companies measure what they get for that spending.

Micah Laughlin, Chief Strategy Officer · October 12, 2026

In short. At the same rate, usage charges rise as people use more AI. U.S. companies plan to spend 52% more on AI per employee this year than last, and less than $1 in $5 of large-company AI usage spending can be tied to a result anyone measured. The company examples show how usage, pricing, and vendor defaults can contribute to a larger bill. The companies that kept their value did the same things, all after the invoice arrived: they put one control layer between the company and every AI vendor, set a cost per result rather than restricting who could use the tool, changed the defaults, made one person responsible, and showed finance the invoice monthly. Every one came out using more AI. The position: set two numbers before the rollout, the target outcome and the cost per result the company will accept; choose the layer that enforces them; keep comparing the two for as long as the AI runs. That comparison produces the one figure boards have asked for and not received: what AI returns after its running cost. It is not a reason to slow down. It is how the return gets collected.

52%

How much more U.S. companies expect to spend on AI per employee in 2026 than in 2025: $2,068, up from $1,358.

Source: Atlanta Fed, Survey of Business Uncertainty, March 2026.

Less than $1 in $5

The share of large companies' spending on AI usage that they can connect to a specific financial result they measured.

Source: Accenture Research, September 2026, 750 executives at $1B+ companies, 17 countries (global).

FIVE TERMS, USED THROUGHOUT.

Target outcome

the financial result the AI was supposed to produce, written down before go-live, in a number the board already tracks.

Cost per result

running cost divided by what it produced, staff and AI together.

Control layer

one piece of software the company controls, between its people and every AI vendor, through which every AI request passes; it can be built or bought.

Running cost

what the AI costs to operate once people use it, mostly usage charges counted in "tokens" (roughly, a word or part of a word the AI reads or writes).

Default settings

the choices the vendor makes unless the company changes them: which version of the AI (the "model") answers, how much it reads, how much it may spend.

Other terms are defined where they first appear.

01

Three things that are all true.

> **In brief.** AI spending per employee is rising about 50% a year, and a few companies are spending most of it. How fast the invoice grows is set by the vendor's defaults, not the company's plan. Most companies cannot say what the spending was supposed to produce. And every surprise AI invoice with a company's name on it arrived after adoption worked, at the companies that have gone furthest first.

THE DETAIL

AI spending is rising. U.S. companies spent $1,358 per employee on AI in 2025 and expect $2,068 in 2026, up 52%. More than half expect to spend no more than $200 per employee; the top 10% plan $2,800 or more, 14 times as much. The gap is not two kinds of company. It is how far one company's spending rises once its people use AI heavily.

The vendor sets the pace. In April, Anthropic changed which model answers by default in Claude Code, its tool for software developers, and its own published cost estimate went from $6 to $13 per developer per working day. A spokesperson: "There was no pricing or product change." An independent measurement found the same work cost about 40% more after one model update at the same advertised price.

Most companies cannot say what the spending was for. Four in five executives at large companies say AI is producing measurable results; less than one dollar in five of their AI usage spending can be tied to a measured financial result. In the U.S., 82% of senior leaders at companies investing in AI are concerned about usage costs; 64% say their company tracks usage with clear limits. More are worried than have controls, and more have controls than can answer the board's question: what was this supposed to do for us?

What it looked like. Uber used up its 2026 budget for AI coding tools about four months into the year; Priceline's renewal for one tool came back at four to five times the price; Salesforce told investors in August that AI usage costs were part of why it did not raise its profit forecast. Those invoices arrived in the spring. By autumn they had reached filings, forecasts and the surveys boards read, which is why the question is live in budget season. The cases, and the evidence that cuts the other way, are in section 05.

02

The position.

> **In brief.** Set two numbers before the rollout: the target outcome, and the cost per result the company will accept. Then control the cost per result rather than who may use the tool, and keep checking both for as long as the AI runs. That is how the return gets collected. It is not a case for restricting AI.

THE DETAIL

The two problems in this issue are one problem. A company that never wrote down what the AI was supposed to produce has nothing to judge the invoice against. A company that did, but left the vendor's defaults in place, watches the running cost grow with success until it eats the result. The companies that kept their value did both halves: a target, and a cost per result, held together and rechecked as models, prices and users changed. They came out using more AI, not less.

No company in the evidence put the controls in before the rollout; every published account runs roll out, surprise invoice, then controls. The advice here comes from what that order cost them, and it works whether a company is planning its rollout or already has tools running.

What would prove this wrong. A company (roughly 1,000 to 10,000 employees) that gave AI tools to most employees with no cost-per-result target and still showed a

measured financial benefit without restricting access. Evidence that typical spending stays flat as use deepens. A vendor default change that lowered the cost of the same work as often as it raised it.

03

What the companies that kept their value did.

> **In brief.** The companies that kept their value did one thing the others had not: they put a single control layer between the company and every AI vendor and ran their cost control from there. Large companies built it. A company can buy it, and choosing it well is the most consequential decision in the plan. From that layer they controlled a cost per result (per software change, per booking, per document), changed which model answers by default and how much it reads rather than taking the tool away from people, made one person responsible, and showed the invoice to the finance team every month. Every one did it after the invoice arrived, and every one came out using AI more, not less: four times the users on the best tools, 30% more software shipped, usage near a company high at half the spend. Two software companies published the numbers; no company that only buys AI has published both halves of its own, which is why the two numbers have to be set in-house.

THE DETAIL

The companies that kept their value reduced cost by changing the settings, changing which model they used, or accepting lower profit, rather than by restricting who could use the tool. Uber and Priceline added spending caps, and Priceline relaxes its limits for work that brings in revenue. One company reversed, in one division. These are the practices, in their words, starting with the one that makes the others possible. The invoices they were responding to are in section 05.

They put one layer between the company and the AI vendors. The companies with the clearest results did it from one place: a piece of software the company controls, sitting between its people and every AI vendor, through which every request passes. Coinbase's engineers built a gateway that "routes everything" through a single point:

it chooses the model, reuses material already read, keeps a record of who spent what, and applies redaction, logging and cost limits "before anything reaches a vendor." Man Group built its around the firm's data and a different model for each kind of task. OutSystems, a software company with about half its customers in Europe, built a gateway its chief executive calls an LLM router for its own engineers after its token spending peaked in June and July; by October the company was spending less than it had forecast for the quarter, "because of optimizations largely," and he said most of what it had been doing with AI "we could do just as well, a lot cheaper with a better harness." The arXiv paper on AI agent software explains why the layer matters: the software the vendor bundles already decides which model answers, what it reads and how much it reuses, so "it picks the rate on the price sheet and sets the volume bought at it," and a company that keeps the defaults "inherits these choices and their bill."

Most companies have only part of one. KPMG's global survey in September found security and data-access controls the most common capabilities in companies' AI control layers, at 43% each, and model routing, "which decides which model handles which task," the least common, at 23%. KPMG's reading: "Most organizations appear to have built a gate before they built a switchboard. A gate lets AI be deployed. A switchboard decides where work goes, and that is where orchestration and consumption economics begin." Among companies reporting an established return, 41% have the switchboard.

Large companies built theirs with engineering teams the size of Coinbase's. A company does not have to. The vendors bundle their own version of the layer with their products, and the arXiv paper proposes a ladder for deciding whether to rent that, buy an independent one, or build. This is where the evidence ends and the position begins. Done well, one layer gives a company four things at once: its data kept off public AI services, through the redaction and data-access controls that KPMG found are the most common capabilities companies build first; the freedom to switch models as prices and capabilities change, without rebuilding what sits on top; AI agents that can be moved to a different model when the market moves; and one

place where the cost of every agent and every team is visible against a limit. Without the layer each of those is a separate project; with it they are settings. The wrong choice locks a company into one vendor's models and prices for as long as the AI runs. The right one is only as useful as the numbers it is set to enforce, which is why it comes after the two numbers in the plan, and why which layer, built or bought, set against which targets, is a different answer for every company.

They controlled the cost of each result, not the number of people allowed to use the tool. Booking Holdings, with roughly 9,000 technology staff, measures its software teams on total technology cost per finished software change, one number combining staff salaries and AI usage charges. Its chief financial officer told Fortune that AI spending can go up as long as the cost of each software change that goes live comes down, and those teams have increased the changes going live by about 30%, counting only changes that passed quality checks. That is the rule: one cost-per-result figure, staff plus AI, with total spending free to rise as long as the cost of each result falls. The evidence is one conference talk with no dollar figure, and Booking is far larger than a company; it is the approach to copy, not a comparable company. Uber's chief technology officer, four months after the company had used up its annual budget for AI coding tools: the price Uber paid per token fell while the number of employees using the most capable AI tools quadrupled, "Not because we've restricted access, but because we've treated efficiency as an engineering problem rather than a budget problem." Fortune's caveat stands: Uber did not say whether the total invoice fell.

They changed the default settings instead of lowering the spending caps. Coinbase's AI spending fell by nearly half in a quarter while usage stayed near a company high. Its engineering lead: "91% of employees weren't hitting their usage caps. Instead of lowering caps, we set cheaper model defaults to cut spend." The share of material reused rather than re-read rose from 5% to 60%. Accenture found that 54% of AI requests go to a more expensive model than the task needs, and that 90% of more than 9,000 tasks it analyzed did not need the most capable model; its advice is to have central software decide, "not as guidance left to each developer's judgment."

They made one person responsible and showed the AI invoice to the finance team every month. Smartsheet made the team that manages its technology budget responsible for total AI spending, set warnings before people reach a limit, and published cost reports by department and by manager. Bristol Myers Squibb's technology chief prepared the CFO and board in advance to expect "pretty high token consumption," telling the Journal it would be "orders of magnitude higher than what we spend today"; the only case in the record where the cost was forecast rather than discovered.

The cheapest model is not always the cheapest. A cheap model that is bad at a task can cost more than an expensive one that finishes it. On long tasks with many steps, the most expensive model cost less in total than the next cheaper one; researchers confirmed this by recalculating the cost of about 10,000 real work sessions. At Docusign it ran the other way. And a model can be chosen for neither reason: Salesforce rebuilt its AI assistant on models made for specific jobs and published speed, not cost, about 26 thousandths of a second to decide how to handle a request against about 1,446 for an all-purpose model, saying outright that the result "isn't a smaller stack, or a cheaper one." A slow answer holds up the work that depends on it. So "can we afford the more expensive model here?" has an answer only against the financial result that AI project was set up to produce, and the answer changes as prices, models and the work change.

What cuts against the pattern. The order of events. Salesforce deliberately let its developers use Anthropic's AI without spending limits for six months, with a fallback plan ("worst case scenario, we would pull back"), and only then began choosing a model per task. It presents "roll out first, control later" as deliberate. Salesforce's August 2026 outlook projected a 34.3% adjusted operating margin for fiscal 2027. Its deputy CFO cited token costs as one reason the company kept that forecast unchanged (section 05). Companies with fixed AI budgets may have less room to try the same approach.

Two companies published the numbers. Docusign, 7,044 employees, processes more than a million agreements a day; software decides which size of model handles each document and cuts what the model reads from the equivalent of a long book to a few pages. In August it published the results for the smaller model it trained for the job: roughly a twentieth of the cost of the large model it replaced, about eight times the documents in the same time, and 7 percentage points more accurate. Cost went down and accuracy went up at the same time; ratios with no dollar figures, from the company's own blog. Figma, 1,886 employees, told the Securities and Exchange Commission that the direct cost of delivering its product rose $32.6 million in one quarter, "primarily due to a $27.1 million increase in technical infrastructure and hosting costs relating to AI and increased usage of our platform by paid users," and the same report is the first to include a sentence about managing that cost: changes "intended to manage usage and optimize the efficiency of our AI-related spend." The $27.1 million covers both AI costs and the cost of serving more paying customers; the report does not say how much is each. Both companies sell AI, so a rising AI invoice is matched by sales income. A company has the opposite situation, an invoice for tools its own employees use with no sales income tied to it, and no company of that kind has published that invoice as a number; where it appears in a report to regulators, it is buried inside an existing expense line.

The number nobody has. How often does running cost overtake the result at a typical U.S. company under 10,000 employees? No source measures it, and this issue does not estimate it. The companies that publish a cost attribute no result to it; the companies that publish a result publish no cost. And mid-sized companies are the weakest at proving a result at all: 14% of companies with $100 million to $1 billion in revenue can show their chief financial officer a measurable result from AI spending, against 34% of smaller companies, 17% of those at $1 billion to $10 billion, and 23% above $10 billion. Global survey; countries not disclosed. Until a company puts both halves in one document, nobody knows, and a company that sets both numbers for itself will be the first to.

The same choices, made before the invoice. The companion guide to this issue scores three invented AI ideas at a 1,200-person professional services company this way: customer support assistance, with a result already defined and a cost that grows with every request; reporting automation, cheap to run but with no baseline measured; and help writing sales proposals, worth eight people's selling time once the pricing rules are fixed. The order of funding came out of the scoring, before a vendor was in the room. A company that does not score its ideas this way discovers the cost, and the lack of a result, when the invoice arrives.

04

The test and the plan.

> **In brief.** Five questions to ask about one AI investment before it gets budget, and seven steps that keep the two numbers together for as long as the AI runs. The two numbers are not a one-time approval check; the model, the price, the users and the work all keep changing, so the comparison runs from the first day and is redone on every change. Kept that way it produces one figure the board has not had: what AI returns after its running cost. The CEO names the target outcome and the owner, the CFO sets the cost per result and receives it monthly, the IT leader chooses the control layer and the settings that hold it, in that order. The comparison starts the day the AI goes live, but a project is judged against its cost per result only once it is out of its development phase, with a dated end and a spending cap, because the cost is real from day one and the result takes time to stabilize. Seven steps, two numbers, and a one-page monthly report.

THE DETAIL

The test. A company that can answer all five has a result it can defend and an invoice it can explain.

	QUESTION	WHAT A PASSING ANSWER CONTAINS
1	What should it return, and how will we measure it?	A number the board already tracks, measured before go-live, a target, a date.

	QUESTION	WHAT A PASSING ANSWER CONTAINS
2	What will it cost to run as more people use it?	Fixed monthly fees and usage charges listed separately; how much usage the fee includes and when, at the planned level of use, it runs out and overage begins; the default model and the material it reads named; a cost per result.
3	What work has to change?	The way the work is done changed from start to finish, not the AI tool added to the old way. Which tasks people stop doing.
4	What has to be in place before we start?	The data cleaned up and business rules written down; the model, settings and spending caps chosen; the layer that enforces them in place, built or bought; the cost report running; the current performance measured.
5	Who will own it once it's running?	One name accountable for the result and the cost per result together, reporting monthly to finance.

A company that can answer the first question and not the second has a result the invoice may wipe out. A company that can answer the second and not the first has a cheap tool and no idea why it bought it.

Development first, then production. Two things make the division harder than it sounds. The invoice stacks a variable charge on a fixed base (fees per user, the platform, the control layer itself), so a project is charged only the variable cost it causes; the base is the cost of having AI at all, reported once. And the result is unstable while a project is being tuned, while the cost is real from the first request, so a project measured in its second month looks like it is losing money even when it is on track. So every project gets a development phase with a dated end and its own spending cap, during which the cost per result is measured and reported, not judged. From the date it moves to production, the three verdicts in step 4 apply. How long that takes differs by project, so the plan sets the date per project. Nothing in the evidence says how long a typical project takes to reach a stable cost per result; this is the position, not a finding.

Why the numbers have to be kept, not set. A new default model, ten times the users, an agent that now reads twice as much: each moves the cost after the plan was approved, so the two numbers are compared from the first day and rechecked whenever anything changes. Kept that way, the comparison produces one figure the

board has not had: what AI is producing for the company minus what it costs to run, by project, added together, month by month. Beside it, what the company is doing with that value: some new revenue, some costs eliminated, and most of it, in the evidence so far, hours saved, which the company has to reassign to other work. Reassigned to what, and was that measured? That is the part of the value most often lost.

Who does what. 75% of board audit committees rank oversight of AI among their top three priorities, and it is the area they feel least confident overseeing, at 56%. Finance is responsible for AI costs and returns at only 5% of companies. Companies where one named person is responsible for the costs and returns of AI are 3.7 times more likely to be able to show their chief financial officer a measurable result; among the 12% with no such person, none could. So the order is by seat. The CEO names the target outcome and the owner in the same sentence, before the budget is approved. The CFO defines the cost per result at the planned level of use and receives it monthly next to the result. The IT leader chooses the control layer and sets the defaults that keep the cost per result within the limit. A company that starts with the settings ends up with a smaller invoice and no result to point to; a company with tools already running starts with whichever is missing.

The plan. Every company in this issue that kept its value did some version of this after the invoice arrived. The plan is to do it before, or to wrap it around what is already running; step 7 is where that starts.

THE PLAN

Step 1. **Name the target outcome for every AI project, in a number the board already tracks.** Measure the starting point before go-live. Set the target and the date. Decide, now, what work the saved hours will go to.

Step 2. **Set the maximum cost per result the project is allowed to reach.** Forecast it at the level of use the plan expects, not the trial's, including the month the allowance runs out and overage begins. Name the measure: cost per resolved customer request, per software change, per accepted proposal, staff plus AI. Set the development phase: its end date and its own cap.

Step 3. **Choose the layer that will enforce it, built or bought, and set the defaults there.** Which model answers by default, how much it reads, when a request is passed to a more expensive model, the spending cap that actually stops spending, and cost reported by team and by week, including every background agent as if it were a team. Choose the layer against the two numbers, not a vendor's feature list. Control the cost per result, not who gets to use the tool.

Step 4. **Put both numbers on one page, monthly, for each AI project.** Making money, breaking even or losing money, from the date the project leaves development. Finance reads it, and one named person answers for both numbers together.

Step 5. **Rebalance on every change.** A new model, a new price, more users, a changed AI agent, a task that has to be done faster. Each reopens which model that project can afford, and the answer depends on the target outcome, not the price list.

Step 6. **Add it up to one number for the board, and break it down.** New revenue, costs eliminated, hours saved, and what work those hours went to, measured.

Step 7. **Start with what is already running.** The AI projects with the largest invoices get their two numbers set first. The ones with no target get a starting measurement before they get more budget.

The answer the evidence in the next two sections gives to the question that opened this issue: at the companies that have gone furthest, the invoice did eat into the value where nobody had written down the two numbers, and the ones that kept their value controlled the cost per result, not who could use the tool, and kept expanding. Nobody has measured the return for a typical company, which is why the two numbers have to be set in-house. Seven steps, two numbers, one page a month. It is less work than a surprise invoice, and it is the difference between AI that costs money and AI that pays.

05

When the invoice passed the value.

> **In brief.** These are the invoices the practices in section 03 were a response to. Every one with a company's name on it came after adoption worked: a budget used up by April, a renewal that came back at four to five times the price, AI usage costs given by one company as a reason it did not raise its profit forecast for the year. Across the market, one large company in three uses up its annual AI usage budget before the year ends. Two things are true that sound like opposites. Companies invested less in AI overall than they predicted a year ago; the big plans came in small. But where they did put AI to work, the usage charges ran past the budget. The surprise invoices hit the companies that have gone furthest first, in the spring; by autumn they had reached filings, profit forecasts and board surveys. And cost is not the reason the typical company has held back; most have not reached the point where the invoice matters yet.

THE DETAIL

Uber used up its 2026 budget for AI software-writing tools about four months into the year, and a $1,500 monthly cap per employee per tool followed. The harder problem was not the money. It was that Uber could not say what the spending had produced. Its president and chief operating officer, Andrew Macdonald, in a podcast interview reported by Fortune: "Maybe implicitly there's more that is getting shipped, but it's

very hard to draw a line between one of those stats and 'Okay, now we're actually producing like 25% more useful consumer features.'"

Priceline. When its contract for an AI coding tool came up for renewal, the new price was four to five times higher, an employee told TechCrunch; the vendor's usage reports did not match the company's own data, and usage limits on certain teams followed. "With AI, you're putting the credit card in the hands of the end user," its senior director of IT finance told the Wall Street Journal. "If you have no control over that, or if the end user is not educated enough, they're going to run up that tab." **Microsoft** cancelled AI coding tools it was buying from another vendor for one division, the only case here where a company took a tool away entirely; cost is the stated reason, and Microsoft sells a competing product.

Salesforce is the one company where AI usage cost affected a figure investors track: its adjusted operating margin guidance. Its deputy chief financial officer told an investor conference on August 27 that about six months earlier the company "unleashed Claude in our R&D cycle," giving its product developers Anthropic's AI without limits, and: "It is part of the reason we did not raise margin guidance on the year is because we are covering some of the token spend that we have got going." The results announcement filed the day before maintained full-year fiscal 2027 non-GAAP operating margin guidance of 34.3%.

Across the market. "Every CFO got the token bill in February and March," the chief executive of OutSystems said in an October podcast interview. "It was an oh shit moment for the whole world." One in three large companies use up their annual AI usage budget before the year ends; only 53% can tell which team or person generated the usage. One chief technology officer, in an account relayed to TechCrunch by the head of Faros AI, described an engineer who spent $40,000 on AI usage in a month and said he genuinely did not know whether to stop him or tell everyone else to be like him. That is the problem in one sentence. Without a stated financial target, a large invoice and a large success cannot be told apart.

How many pulled back. In KPMG's mid-2026 global survey, 49% of companies had scaled back, paused or questioned projects using AI agents (AI programs that carry out multi-step tasks on their own) once the expected cost began to outweigh the value. A survey of what people say, from a sponsor that sells governance work; it confirms the direction, not the figure.

The evidence on the other side. Total AI investment came in below plan: a year ago 35% of U.S. senior leaders at companies investing in AI expected to be spending $10 million or more by now; only 23% are. Companies did less than they planned, and the part they did ran over. At companies already paying for usage-billed AI tools, more say the cost led them to use AI more (37%) than to cut back (15%). And cost is not why the typical company has held back: among U.S. businesses not planning to use AI, "too expensive" ranks eighth of eleven reasons, at 6.9%. The invoice has hurt the companies where the most employees use AI the most. The named cases arrived between February and June; from August the same costs reached a quarterly filing (Figma), a profit forecast (Salesforce), and the surveys boards read in budget season. It reaches the rest in the order they adopt.

06

How the invoice behaves.

> **In brief.** AI products charge a small fixed fee per user and a large variable charge for usage. Four things drive the usage charge: the default model, how much background material the AI reads each time, how much extra work each new vendor release makes it do, and AI agents left to work without a person watching. The usage included in the monthly fee runs out earlier in the month as adoption grows, and overage starts; a background agent can do this with nobody watching. The price per token is falling and the invoice still rises, because the price cuts apply to older models, newer models do more work per request, and companies use the savings to do more with AI. The only way to judge the numbers is one AI project at a time: what that project costs to run against the result it was measured to produce.

Where a vendor charges a monthly fee per user, it also charges for usage on top, and the fee is the small part: Anthropic's business plan is $20 per user per month plus usage at standard rates.

Four things drive the usage charge, and the company controls none of them unless it changes the defaults.

The default model. The software that runs an AI agent, breaking a job into steps and sending each to the AI, is usually supplied by the vendor. It decides which model answers, what it reads, whether it reuses material already read, and which extra AI helpers it starts. The vendor's software chooses both the rate and the volume, and a company that leaves the defaults in place accepts both.

How much the AI reads. In AI coding tools, reading is where the cost is: re-reading the company's existing code accounts for about 83% of the spending, while the new code the AI writes is 0.4% of what it processes.

Volume, driven by the vendor's releases. Companies in Accenture's survey expect the amount of AI they use to rise 78% over two years while the price per token falls 19%. Our arithmetic on those two figures: the invoice ends up roughly 44% higher if nothing else changes.

Agents left running. In the longest 1% of AI coding sessions Jellyfish measured, sessions with little human supervision produced about eleven times as much usable code as sessions of the same length where a person checked each step. Unsupervised AI work produces the most output, and runs up usage charges fastest.

Caps, allowances and overage. A spending cap does not reliably stop spending. OpenAI's pricing page says the AI stops answering once a monthly budget is reached, then adds: "There may be a delay in enforcing the limit, and you are responsible for any overage incurred." Google's Gemini accounts are moved automatically to a

higher spending level, so a $250 limit becomes $20,000 to $100,000 once an account is 30 days old and has paid $1,000; Google confirmed this. Most plans bundle a set amount of usage into the monthly fee and charge for everything beyond it; Anthropic's plans draw all of a user's activity "from the same pool," reset on a rolling five-hour window with weekly limits on top, and at the limit the choices are to wait, move up a plan, or "turn on usage credits to keep working at standard API rates," that is, pay for overage. The vendor sizes the allowance; adoption uses it up.

Nobody is watching the agents. An AI agent running in the background, with no one at a screen, has no one to notice its overage. Anthropic bills its managed agents at $0.08 per hour of active runtime on top of token charges, so the clock runs whether or not the work is useful. One company's prepaid credits ran out silently and $30,141 of charges accrued before its cost-alert tool said anything. A background agent that quietly adds an overage charge every month is the most ordinary form this takes.

Why falling prices do not fix it. The price of a fixed level of AI performance falls roughly 13-fold a year, and the price U.S. businesses actually paid per million tokens fell 41% between March and September 2026. The invoice still rises, for three reasons. The cuts apply to last year's level of performance, so the newest model, which is what vendors sell today, has not been discounted yet. The work the AI does per task grows faster than the price falls: the cost of running the best available model rises 3 to 18 times a year. And companies spend the savings: nine in ten executives say cheaper usage prices would lead them to use AI for more things rather than spend less.

So a pilot's invoice predicts nothing about the invoice after a rollout: one model, a few people and short sessions, against the vendor's default model, all employees, and long sessions nobody is watching. The only way to judge it is project by project: what this AI project costs to run, against the result it was measured to produce. That is what cost control means in this issue. It does not mean deciding who may use the tool.

Method.

Correction, November 24, 2026.

The 34.3% Salesforce figure is its projected fiscal 2027 non-GAAP operating margin. An earlier version described it as actual profit per dollar of sales.

What this issue is.

A synthesis of published evidence, with a stated position and the evidence that cuts against it. What follows is what the evidence proves and does not, how the sources were weighed, and which widely quoted figures were refused.

Thesis.

AI spending rises with how much people use it, not with what the budget planned; how much it rises is determined by the vendor's default settings unless the company changes them first; and a company can judge the invoice only against a target outcome it set before the rollout. **Position.** Set both numbers in the same plan, before the rollout; control the cost per result, not who gets to use the tool; keep comparing the two for as long as the AI is in use. **What would prove it wrong** is stated in section 02.

Who wrote it.

Stratos Edge is the AI adoption intelligence company. It is an AI strategy consulting firm that helps companies decide what to do next with AI, what it should return, and who should deliver it.

What it does not prove.

That any company put the controls in before the rollout. That running cost has overtaken results at a typical company; no source measures it. That a company's internal AI invoice looks like Figma's or Docusign's; those are the costs of products

they sell. That seeing the invoice produces a result: two surveys disagree on it, and this issue relies on neither.

How the sources were weighed.

Tier 1: government statistics and Federal Reserve surveys, company filings read from EDGAR, and vendor documentation with the vendor on record. Tier 2: named institutional and consulting research with a disclosed method, conflict named in the sources. Tier 3: vendor data about its own customers and vendor-commissioned surveys, labeled, never carrying a claim alone. U.S. figures carry the argument; global figures are labeled and placed beside a U.S. figure where one exists. No number is built from two surveys. The Atlanta Fed's $200 describes the median U.S. company; the $1,358 and $2,068 are averages in which larger employers count for more. Named companies are described by what they published or said on record, never judged; large-company cases are contrast, not proof. Docusign's June post contains a cost sentence that contradicts itself as published and is not quoted. Booking's measure rests on one conference talk with no dollar value. OutSystems' account is a podcast interview with an automatically generated transcript, read in full; the company sells an AI platform. Coinbase's figures are its executives' own public posts as carried by four relays; Uber's monthly cap is as reported by Bloomberg.

Figures not used.

A widely circulated "73% of organizations exceeded AI budget" attributed to the FinOps Foundation does not appear in the Foundation's report and, traced back, comes from several sources that disagree. A story of one engineer spending about $180,000 on AI usage in a month has no primary source. "Token prices fell 98%" is quoted from a source quoting another source, with no original data found.

Gaps searched.

A company's internal AI invoice as a number: not found in any 2026 filing, shareholder letter or engineering blog; across every 10-K filed with the SEC this year,

and every 10-K and 10-Q filed since August 9, searches for "token costs" and "AI operating costs" find nothing. Board surveys ask about AI oversight, not spending; no Federal Reserve or Census question covers AI cost control. The nearest healthcare figure is one employee's $75 example at Jefferson Health.

How current the data is.

The main figures come from the last six months. Figma's next quarterly report (November), the next round of EY's survey (overdue), KPMG's fourth-quarter survey, Salesforce's third-quarter results (late November or early December, for the profit-forecast sentence) and the fourth-quarter CFO Survey (December 16) are when they will be rechecked. The data appendix lists, for each figure, when it was last checked.

Corrections.

No corrections have been logged to this issue since publication. A correction is recorded here with the date, the original text, the corrected text, and the reason.

Sources.

1. Federal Reserve Bank of Atlanta, Policy Hub Macroblog, "Firms' AI spending per employee," May 6, 2026, on the March 2026 Survey of Business Uncertainty.
2. Ramp Economics Lab, "September 2026 Ramp AI Index," September 9, 2026; Ramp, "Business Spending Report, Summer 2026." Card and bill-pay transactions, 70,000+ U.S. businesses. Ramp sells spend management.
3. KPMG, Global AI Pulse Q2 2026 (n=2,145, 20 countries). KPMG sells AI governance work.
4. Accenture Research, "The CIO's guide to AI tokenomics," September 10, 2026 (n=750 executives at $1B+ enterprises, 17 countries; 15 Fortune 500 interviews). Accenture sells the work.

5. EY, US AI Pulse Survey, fifth wave, newsroom release July 28, 2026 (n=534 U.S. SVP+, fielded April 24 to May 17, 2026, ±4pp).
6. Business Insider, "Anthropic doubles its estimate of what Claude Code tokens will cost engineers," April 2026, with Anthropic's spokesperson on record.
7. Jellyfish Research, "The State of AI: A Research Review on AI in Software Engineering, H1 2026," June 24, 2026. Vendor data on its own customers; global.
8. Kwartler, Aqrawi and Abbasi, "Harness Tokenomics: A Router for the Enterprise Agentic Control Plane," arXiv:2609.28919v2, September 26, 2026. Unrefereed preprint.
9. Epoch AI, "The plunging price of thought," September 22, 2026.
10. Gundlach, Lynch, Mertens and Thompson (MIT FutureTech), "The Price of Progress," arXiv:2511.23455v2, March 2026.
11. Vendor pricing pages fetched October 8, 2026: Anthropic, OpenAI, Google Gemini API, Cursor, GitHub Copilot; Google Cloud, "Spend Caps," July 28, 2026. Primary facts about pricing only.
12. The Register, on Google Gemini API tier auto-upgrades and on cloud AI billing mechanisms, May 2026, with Google and AWS on record.
13. Forbes, Janakiram MSV, "Uber Burns Its 2026 AI Budget In Four Months On Claude Code," May 17, 2026.
14. Fortune, Sheryl Estrada, on Booking Holdings' CFO at the Fortune AIQ Summit, October 2 and October 5, 2026; Fortune, on Uber's CTO and COO, August 7, 2026.
15. TechCrunch, Rebecca Bellan, "The token bill comes due," June 5, 2026.
16. The Wall Street Journal, CIO Journal, Belle Lin, "How Companies Are Managing AI Token Spend," June 30, 2026.
17. VentureBeat, citing The Verge, on Microsoft's cancellation of third-party AI coding licenses in one division, 2026.
18. Deutsche Bank 2026 Technology Conference transcript, Salesforce deputy CFO, August 27, 2026 (third-party transcription); Salesforce Form 8-K Exhibit 99.1, August 26, 2026; Salesforce newsroom, July 8, 2026.
19. Coinbase, CEO and AI-platform engineering accounts, June 27, 2026, as carried by four independent relays.
20. Bloomberg Odd Lots, Man Group on 86x token growth, July 9, 2026.

21. The Cognitive Revolution podcast, Woodson Martin (CEO, OutSystems) with Nathan Labenz, episode published October 7, 2026, automatically generated transcript; as first reported by heardin.ai.

22. Docusign engineering blog, "Architecting for millions," June 23, 2026; "How a Small Model Learned to Do a Large Model's Job at Docusign's Scale," August 12, 2026; Docusign Form 10-K FY2026 (headcount).

23. Figma, Inc., Form 10-Q for the quarter ended June 30, 2026, filed August 5, 2026; Form 10-Q for the quarter ended March 31, 2026, filed May 14, 2026; Form 10-K FY2025 (headcount).

24. SEC filings, 2026: Doximity Form 10-K FY2026; ServiceTitan Form 10-Q (filed September 8, 2026).

25. Upwork, Annual Shareholder Report, April 23, 2026.

26. Deloitte and Center for Audit Quality, Audit Committee Practices Report, 5th edition, September 2026 (n=248, 88% U.S.).

27. Tokenomics Foundation, State of Tokenomics, September 23, 2026 (n=472 companies; geography undisclosed).

28. U.S. Census Bureau, Business Trends and Outlook Survey, 2026 AI supplement (~117,000 responses).

29. Becker's Hospital Review, Q3 2026 CIO roundup, September 29, 2026.

30. KPMG, Global AI Pulse Q3 2026, "AI at scale: Accountability, resilience and economics," September 2026 (n=2,131, 20 countries, fielded July 23 to August 26, 2026). KPMG sells AI governance work.

Cite this issue.

Cite this issue. Laughlin, M. "The AI Invoice Means It's Working. Is It Buying Anything?" Stratos Edge Research, Issue 2, October 2026. stratosedge.ai/insights/the-ai-invoice-means-its-working. Reproduce with attribution; the data appendix is open.

Data appendix: issue-2-data.csv (Download the PDF), one row per number in this issue with metric, value, population, sample, source, URL, date and evidence tier. Series methodology and citation forms at stratosedge.ai/research.

Questions this issue answers.

What does an AI plan return once running costs are counted?

Nobody can say in general, because no public source measures running cost against results at a typical company. A company can say for itself once it has set two numbers for each AI project: the target outcome in a number the board already tracks, and the cost per result it will accept. Whether the project is making money, breaking even or losing money then becomes a figure read off a monthly report, not a matter of opinion.

Why do AI usage costs exceed budget?

Because the invoice rises with use and is set by the vendor's default settings: the default model, how much background material the AI reads, the amount of work that jumps with each new vendor release, and AI agents left to work on their own. The usage included in a monthly fee runs out earlier as more people use the tool, and overage billing starts. One vendor more than doubled its own estimate of what its product costs to use when it changed the default model, with no change in the price per token. Companies were three times over their annual AI usage budget by April 2026.

How should a company control AI usage costs?

Control the cost per result, not who gets to use the tool. Set a cost per result (per resolved customer request, per software change, per booking), forecast at the planned level of use; set the default settings that hold it, in one control layer between the company and the AI vendors, built or bought; report it monthly beside the result; rebalance on every change to the model, the price, the users or the work. The companies that kept their value did this. None solved it by cutting people off from the tool, and all of them ended up using AI more.

What should the board be told about AI?

One figure tracked over time: what AI is producing for the company, minus what it costs to run, by project, added up, and broken down into new revenue, costs eliminated and hours saved, with what work those hours went to measured. And one name that answers for the result and the cost together.

Will falling AI prices fix this?

No. Prices for a fixed level of AI performance fall about 13-fold a year, but the price cuts apply to older models, newer models do more work per task, and nine in ten executives say they would spend a price cut on more use rather than save it.

Confidence **Before Technology**

STRATOS EDGE RESEARCH · THE AI ADOPTION INTELLIGENCE COMPANY · STRATOSEDGE.AI

© 2026 Stratos Edge LLC. All rights reserved. This article may be quoted with attribution and a link to the original. It may not be reproduced in full or used to train models without written permission.