INSIGHTS · AI ECONOMICS & COST CONTROL

Your AI Bill Just Became a Variable Cost

Seat licenses were a fixed cost: budget once, pay monthly, explain nothing. The meters replacing them charge by usage, across vendors whose units don't reconcile. Most finance teams are controlling the result the way they controlled a 2010 travel budget: they set the number, then explain the overrun when the statement arrives. The tools to do better are already in the controller's kit. They just haven't been pointed at AI.

EvoNova Advisors | Insight | October 2026 | 12 min read

BY THE NUMBERS

53%

of CFOs at North American companies with $1B+ in revenue say their primary practice is to review AI usage costs after the fact, from bills, reports or dashboards.

Deloitte CFO Signals, Q3 2026 (n=200)

7%

respond automatically or in near real time when AI cost controls fail. 74% take several business days to several weeks to pause the usage.

Deloitte CFO Signals, Q3 2026 (n=200)

83%

of respondents in a survey of 2,000 finance executives say consumption-based AI costs pushed spending above expectations; 38% of all respondents are more than 25% over.

Pigment CFO Index, Q3 2026 (vendor survey, n=2,000)

The statement arrives after the money is spent

Budget season is under way, and many 2027 plans contain a line for AI that nobody can decompose. It was built the way SaaS lines are usually built: this year's run rate, plus the renewals, plus a growth assumption someone in IT supplied. That method worked when the line was a seat count. It does not work when the line is a meter.

Deloitte's third-quarter CFO Signals survey, published October 21, asked 200 CFOs at North American companies with at least $1 billion in revenue how they govern AI costs. Fifty-three percent said their primary practice is to review AI usage costs retrospectively, through vendor bills, periodic reports, dashboards or manual analysis. When their cost controls fail, 74% typically pause the usage within several business days to several weeks. Nineteen percent act the same business day, and 7% respond automatically or in near real time. Only 20% use advanced AI-specific controls such as token- or workload-level attribution, AI gateways or caching. The most-cited obstacles were a lack of AI cost governance, processes or policies (40%) and limited timely insight into token usage and unit costs (38%).

© 2026 EvoNova Advisors

Two vendor surveys published within weeks of Deloitte's show how common overruns are, though they disagree on how common. Pigment's CFO Index surveyed 2,000 finance executives in the US and Europe: 83% of respondents said consumption-based AI costs had pushed spending above expectations, and 38% of all respondents put the overrun at more than a quarter. Datarails surveyed 270 US finance leaders and found 32% had exceeded their AI budget by at least 10% over the past year, while 13% said no one clearly owns the AI budget at all. The two surveys asked different populations different questions, Pigment about expectations and Datarails about budget, and their findings do not reconcile. A cost with no agreed unit cannot have an agreed variance.

A 2010 travel budget ran on the same control model: an annual allocation per department, a policy document, and a card statement that arrived a month after the trip. Overruns surfaced in the variance report and were explained after the fact. Travel became controllable when the policy moved from the expense report to the booking screen, where an out-of-policy fare could be stopped before anyone bought it. AI spend is still in the expense-report era, with one difference that makes it worse. A traveler books a few trips a month. An agent can make thousands of billable calls in an afternoon.

The seat died and the meter replaced it

A seat license is a fixed cost. You negotiate a number of users and a price, and the monthly charge does not move with how hard anyone works. Agentic AI breaks that arrangement from the vendor's side first: the vendor's own compute cost now scales with the work an agent does, so its price does too. Salesforce, Microsoft, Oracle, Workday, SAP, ServiceNow and HubSpot now all meter AI work by consumption, and no two of them count the same thing.

Salesforce sells Flex Credits at $500 per 100,000, and a standard agent action consumes 20 of them, which makes it ten cents. In Microsoft's Copilot Studio, Copilot Credits cost a cent each on pay-as-you-go terms, and the count depends on what the agent did: two credits for a generative answer, five for an agent action, ten for grounding an answer in the company's Microsoft 365 data. Oracle sells pooled AI units at $1,000 per 100,000 and derives them from tokens. Each action's input and output tokens are divided by 10,000, rounded up, and multiplied by a factor, from zero to twenty, that depends on the type of action and the tier of language model it used; Oracle's included models consume nothing. Workday's Flex Credits charge per lookup or per task and carry no public dollar price. HubSpot charges for outcomes: 50 credits, or 50 cents, for each customer conversation its agent resolves.

Read across Exhibit 1 and the meters sort into three kinds. Oracle counts inputs. Salesforce, Microsoft and Workday count activity, and HubSpot counts outcomes. A controller would recognize the situation at once if it involved physical suppliers: one bills by weight, one by the hour, one by the piece, and the product you sell needs all three.

The units also move. In April 2025 Microsoft cut the charge for an autonomous agent action from 25 messages to 5, and that September it renamed the unit itself, from messages to Copilot Credits. In June 2025 Oracle said it charged no additional subscription fee for the AI agents embedded in Fusion and flagged usage costs for agents built on other models; its October 2026 terms meter actions on paid model tiers in pooled units. Salesforce's August 2026 rate card adds a "Headless Platform Interaction" meter, which Salesforce describes as covering calls that registered outside agents make into Salesforce. Its rate is still to be announced, and metering has not started. When a vendor re-rates the unit and leaves the price per unit alone, the price change shows up nowhere a procurement team normally looks.

© 2026 EvoNova Advisors

EXHIBIT 1

Five AI meters, and none of them counts what finance counts

Vendor	Meter	List price	How a unit of work is counted	What it counts, and what to watch
Salesforce Agentforce	Flex Credits	$500 per 100,000 ($0.005 each)	20 credits ($0.10) per standard agent action; 30 per voice action; 2–16 per prompt	**Activity.** Confirm in the order form what counts as one action.
Microsoft Copilot Studio	Copilot Credits	$0.01 pay-as-you-go; $200 a month for 25,000 in a capacity pack	2 per generative answer, 5 per agent action, 10 for grounding in Microsoft 365 data	**Activity.** An agent action was re-rated from 25 messages to 5 in April 2025; the unit became Copilot Credits that September.
Oracle Fusion Applications	Pooled AI Units	$1,000 per 100,000 ($0.01 each)	Input plus output tokens ÷ 10,000, rounded up, × a factor from 0 to 20 set by action type and model tier; included models consume none	**Inputs.** Test environments draw down units, unused units are forfeited, and Oracle may block access below zero.
Workday	Flex Credits	No public price per credit	Self-Service Agent: 1 credit per information lookup, 5 per autonomous task	**Activity.** The console alerts at 80%, 90% and 100% of credits and switches nothing off.
HubSpot	HubSpot Credits	$0.01 each	50 credits ($0.50) per text conversation its Customer Agent resolves	**Outcomes.** The only meter in this table priced on a result.

Source: vendor price lists, rate cards and licensing documentation current at October 2026: Salesforce Flex Credits rate card (August 31, 2026); Microsoft Copilot Studio licensing guide and billing-rate page (updated October 2026); Oracle Fusion Cloud global price list and service descriptions (October 2026); Workday Flex Credits page and Platform Consumption Console documentation; HubSpot product and services catalog. List prices; negotiated rates are not public.

The larger exposure comes when one workflow crosses vendors. Gene Dai, in a May piece about HR workflows, called the result the agent action tax. An agent that starts in Teams, grounds its answer in SharePoint, retrieves policy from Workday and opens a ServiceNow case can trigger a different meter at each stop. Each vendor bills its slice in its own unit and on its own cycle. No invoice shows the workflow, and the cost of what the business did is scattered across four statements.

Finance has seen this before, with a different cast

When infrastructure moved to the public cloud, servers that had been bought once and depreciated over five years became a bill that moved every hour. Respondents to RightScale's 2017 State of the Cloud survey estimated that 30% of their cloud spend was wasted. By 2020, Flexera's respondents were running over their cloud budgets by an average of 23%. The discipline that grew out of that period, FinOps, got its foundation in February 2019 and its standard text, *Cloud FinOps*, that December.

FinOps is the obvious template, and its practitioners have already moved on AI. In the FinOps Foundation's 2026 survey, 98% of respondents said they manage AI spend, up from 31% two years earlier. But FinOps grew up inside engineering at companies with very large cloud bills. Seventy-eight percent of FinOps teams report to a CTO or CIO; 8% report to a CFO. Eighty percent of that survey's respondents work at companies with at least 1,000 employees. Flexera's 2026 cloud report notes that enterprises typically have dedicated cost management teams while smaller firms lean on their DevOps engineers, and it found enterprises far more likely than smaller firms to measure unit economics, 53% against 32%. A company with a few hundred employees sits on the DevOps side of that line, and it is unlikely to hire a FinOps team just to watch its AI bill.

The cloud bill also had a structural advantage the AI bill lacks. It came from one or two providers, with a tagging scheme that let an engineer attribute every resource to a team. AI spend arrives through SaaS contracts the CIO never sized, metered by vendors whose usage reports were designed to support their own billing.

Falling prices will not fix this either. Stanford's AI Index found the price of querying a model that performs at GPT-3.5's level fell more than 280-fold between November 2022 and October 2024. Spending rose anyway. Menlo Ventures estimates enterprise generative-AI spending more than tripled in 2025, to $37 billion. In March 2026 Gartner forecast that inference on a trillion-parameter model will cost providers more than 90% less in 2030 than in 2025, and in the same release predicted that overall inference costs will increase, because token consumption is rising faster than token costs are falling. Agentic models, Gartner said, need 5 to 30 times more tokens per task than a standard chatbot. The Linux Foundation's June announcement of a new standards body said that per-token prices had leveled off and that new models' token prices were rising. For finance, the variable that mattered was always volume.

Cost the workflow, not the token

The technology industry has picked its unit. When the Linux Foundation announced its Tokenomics Foundation in June, its chief executive, Jim Zemlin, said tokens had become "the new unit of technology spend." That is the right unit for an engineer tuning a model. A controller needs a different one, for the same reason no manufacturer costs a product in kilowatt-hours drawn by one machine: in cost accounting terms a token is a cost driver, and the cost object is what the business produces.

For agentic work, the cost object is a completed workflow: an invoice exception cleared, a customer dispute resolved, a reconciliation prepared and ready for review. Every meter the workflow touches rolls up to it, the way freight, duty and handling roll into the landed cost of a unit of inventory. Finance has built landed costs from several suppliers' invoices for as long as it has imported goods, and a workflow that fires four vendors' meters is the same calculation with stranger units.

Completed matters because agents fail, loop and retry, and every attempt burns meters whether or not it produces anything. In a plant that is scrap, and a standard cost carries an allowance for it. Cost per completed workflow captures the scrap. Cost per token or per credit hides it, because a failed attempt and a successful one look identical on a vendor's usage report. When first-pass completion drops, cost per completed workflow rises even though nothing on any price list has changed.

Exhibit 2 shows a standard cost card for one workflow, a customer billing dispute that an agent resolves across four systems. The quantities are illustrative; the rates are the vendors' published list prices, apart from the model rate, which is an assumption.

EXHIBIT 2 · ILLUSTRATIVE

One resolved billing dispute touches four meters and costs $2.75

Step	Meter	Standard quantity per attempt	Standard rate	Standard cost
Agent reads the case, pulls the account, updates the record and drafts the reply	Salesforce Agentforce	6 actions	$0.10	$0.60
Agent checks the invoice and payment history in the ERP	Oracle Fusion AI	60 AI units	$0.01	$0.60
Credit approval routed to the account manager in Teams	Microsoft Copilot Studio	25 credits	$0.008	$0.20
Reasoning and summary through the company's own model API	Model API	100,000 tokens	$8 per million	$0.80
Standard cost per attempt				**$2.20**
Allowance for failed attempts at a standard first-pass yield of 80% ($2.20 ÷ 0.80 − $2.20)				$0.55
Standard cost per completed workflow				**$2.75**

Source: EvoNova Advisors illustration. Quantities are hypothetical. Rates are vendor list prices at October 2026, with Microsoft at the capacity-pack rate ($200 for 25,000 credits); the model API rate is an assumed blended rate for input and output tokens.

Some of the overrun is good news nobody can prove

A standard cost lets finance run the analysis it needs, which is a flexible budget. Many AI budgets are static: a dollar figure for the year, divided by twelve. Against a static budget, every month in which the business used AI more than planned shows up as an overrun, and the report cannot say whether the overrun is waste or success.

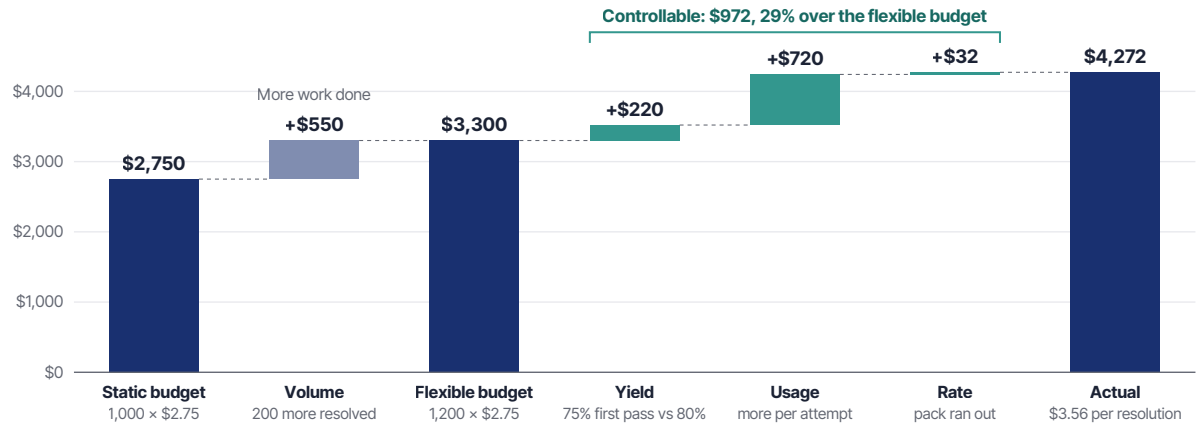
In Exhibit 3 the dispute workflow came in 55% over its static budget. Recast against a flexible budget, $550 of the $1,522 overrun is volume. The agent resolved 1,200 disputes instead of 1,000, which is the business doing more work, and if each resolution costs less than the person who used to handle it, that is good news. The remaining $972 is controllable, and it has three causes a manager can act on. Yield fell from 80% to 75%, so the agent needed more attempts per resolution. Usage per attempt rose: more Salesforce actions, longer prompts to the model, and some Oracle actions routed to a more expensive model tier. And the Microsoft capacity pack ran out mid-month, so the overflow billed at the higher pay-as-you-go rate the company had set up as a backstop.

Oracle's meter shows how hard this is to see from an invoice. Because Oracle multiplies the token count by a factor tied partly to the model tier, choosing a more capable model raises the number of units consumed while the price per unit stays the same. A decision a controller would call a rate decision reaches the bill disguised as quantity, and Exhibit 3 books it as usage because that is how the invoice presents it.

© 2026 EvoNova Advisors

EXHIBIT 3 · ILLUSTRATIVE

A 55% overrun on a static budget is 29% once volume is taken out



Source: EvoNova Advisors illustration, continuing Exhibit 2 for one month. Volume is valued at standard cost per completed workflow; yield at standard cost per attempt; usage at standard rates (Salesforce +$160, Oracle +$240 from routing some actions to a higher model tier, model API +$320); rate on actual quantities (40% of this workflow's credits, its share of a tenant-wide capacity pack, overflowed to pay-as-you-go billing at $0.01 against the $0.008 pack rate).

Pigment's 83% looks different through a flexible budget. Some unknown share of those overruns is the business using a tool that works. Without a standard cost per completed workflow, nobody can say how large that share is, and the finance team that cannot say is left defending every overrun as if it were waste.

The purchase order and the purchase are now the same call

Every other variable cost finance controls has a gap between the decision to spend and the spending. A purchase order commits the money before the invoice arrives, and public-sector accounting goes further by recording the commitment as an encumbrance so the budget reflects it. Travel booking tools stop out-of-policy fares before they are bought. An agent has no such gap. The decision to make a billable call and the call itself happen in the same instant, made by software, thousands of times a day. No approval step can run at that speed, so the control has to be a limit set in advance.

That is the gap Deloitte measured. A window of several business days to several weeks fits a manual chain: the overrun has to show up in a report, get escalated, and reach someone with access to the vendor's console. The vendors have built the control for themselves. Oracle's terms say it may limit or block access to AI functionality once a customer's balance of units falls below zero. Once usage reaches 125% of a tenant's prepaid capacity, Microsoft disables the tenant's custom agents, unless a pay-as-you-go meter picks up the overage. Workday's consumption console alerts administrators at 80%, 90% and 100% of their credits and does not switch anything off. The buyer gets the alert. Where an automatic stop exists, the vendor set it, at a threshold the vendor chose.

A workable throttle is set the way a credit limit or a purchase-order tolerance is set: per workflow, in the unit finance already uses. Two thresholds do most of the work. A unit-cost ceiling, say twice the standard cost per completed workflow, catches an agent that has started looping. A period ceiling, the month's flexible budget plus a

tolerance, catches volume that has outrun the plan. Hitting either one need not stop the work. The workflow can drop to a cheaper model, route to a human queue or wait for approval. Someone accountable for the budget makes that decision before the month starts, and the system enforces it at the moment of spend. The AI gateways that Deloitte counts among advanced controls can enforce limits like these on calls a company routes through its own systems. They cannot reach inside a vendor's own agents, which is why the contract matters as much as the software.

Prepaid credits are inventory with a shelf life

Most of these meters are sold in prepaid blocks: Salesforce credit packs, Microsoft's monthly capacity packs, Oracle's pooled units, SAP's AI units. Oracle forfeits unused units at the end of the services period. SAP's AI units expire after twelve months if unused, and so do Workday's credits, according to CIO.com's reading of Workday's terms. To a controller, a prepaid block that expires is inventory with a shelf life. Whatever goes unconsumed is breakage, and breakage is the vendor's revenue.

Renewal is the cheapest moment to change that. Deloitte found 27% of CFOs already see contracts as a sticking point, citing pricing, commitments or data egress terms. The most valuable ask is an action-level usage export that carries an identifier the customer controls, such as a workflow or transaction ID, so meters from different vendors can be joined into one cost per workflow. Every other control in this article depends on it. Beyond that, ask for the unit definition to be fixed for the term (credits per action as well as price per credit), rollover or true-down rights on unused prepaid units, overage billed at the committed rate up to a cap the customer sets, and the right to reconcile the vendor's meter counts against the customer's own logs.

The objections, taken seriously

"The amounts are too small to matter." The per-workflow numbers are cents, by design. The materiality is in the forecast error. Forty percent of Pigment's respondents said unpredictable usage makes AI costs hard to forecast, and a line that misses its plan by a quarter is not small for long when every function is adding agents to it.

"This belongs to IT." In Datarails' survey, 84% of finance leaders said finance owns the AI budget, and Deloitte found nearly 80% of CFOs say finance has a formal role in AI decisions. IT can supply the meter data. Deciding what counts as a resolved dispute, and what one should cost, is an accounting judgment.

"Our vendors don't export usage at the workflow level." Most don't yet, so it belongs at the top of the renewal list. Until then, an engineered standard built from a sample of traced runs beats having none. Standard costs have always started as engineering estimates and been corrected by their variances.

"An enterprise agreement makes this moot." Salesforce now sells a fixed-price, fixed-term Agentic Enterprise License Agreement with no overages. A flat fee turns the variable cost back into a fixed one, at a price the vendor sets. Whether that price is a good deal depends on a number only the customer can produce: what a completed workflow costs today, and how many of them the business expects to run.

© 2026 EvoNova Advisors

Where to start this budget season

None of this requires a FinOps team. Start with an inventory of every AI meter the company pays for, including the ones bundled into SaaS renewals. Then take the three workflows with the most volume through the five steps in Exhibit 4. Define each workflow as a cost object and trace a sample of runs end to end. Build a standard cost card for it. Budget it flexibly, at standard cost times expected volume, so next year's variance report can separate growth from waste. Set the two thresholds. Carry the contract asks into every AI renewal in 2027.

EXHIBIT 4

Five controls from the cost accountant's toolkit that need no FinOps team



Source: EvoNova Advisors.

Gene Dai closed his piece on the agent action tax with a test he aimed at vendors: pull one workflow, and show every system it touched and every meter it triggered. Put the same test to your own team before the 2027 AI line is locked. A team that can produce that invoice can build a standard cost, a flexible budget and a throttle from it. One that cannot will be explaining next year's AI variance after the money is spent.

Travel stopped surprising finance once the policy moved to the booking screen, and AI spend will stop surprising it once the standard moves to the workflow.

evonova ADVISORS

An EvoNova Advisors Insight. Survey findings are cited from the publisher's own report or release, with sample sizes, and vendor-sponsored surveys are labeled. Vendor prices are list prices current at October 2026 and change often. The cost card and variance figures in Exhibits 2 and 3 are illustrative, built on those list prices and an assumed model rate.

© 2026 EvoNova Advisors

SOURCES

Deloitte — North American CFOs cite AI cost-governance gaps in third-quarter 2026 survey (CFO Signals, 200 CFOs at $1B+ companies) — October 21, 2026

Pigment — CFO Index, Q3 2026 (2,000 finance executives, US, UK, France, Germany; vendor survey) and press release, Report: AI overspending ignored by C-suite — October 7, 2026

Datarails — 2026 CFO Sentiments Survey (270 US CFOs and finance leaders; vendor survey) — September/October 2026

Salesforce — Agentforce pricing and Flex Credits rate card — August 31, 2026; Salesforce Help on Flex Credits for agentic API calls, as reported by Vantage Point — September 2026; Flexible buying models (Agentic Enterprise License Agreement)

Microsoft — Copilot Studio licensing guide; Copilot Credits billing rates and capacity enforcement (Microsoft Learn, updated October 2026); Microsoft Copilot blog on Copilot Studio billing — January 2025; Schneider IT Management, Copilot Studio price decreases — April 17, 2025

Oracle — Fusion Cloud global price list (October 8, 2026) and Fusion Cloud service descriptions, AI Units (effective October 2026); Fusion Insider blog, 2 ways you can adopt AI agents in Fusion Apps — June 2025

Workday — Flex Credits documentation and Platform Consumption Console; CIO.com, Unpacking Workday's agentic AI pricing model — July 2026

HubSpot — Product and Services Catalog (HubSpot Credits); **SAP** — AI packages and pricing (AI Units); **ServiceNow** — Community guide to AI-native licensing in 2026 (assists)

Gene Dai — AI Agents Can Skip the UI, Not the SaaS Bill (the agent action tax) — May 15, 2026, updated October 1, 2026 (personal blog)

RightScale — 2017 State of the Cloud Report — February 2017; **Flexera** — 2020 State of the Cloud Report — April 2020 (vintage; vendor surveys)

FinOps Foundation — State of FinOps 2026 (1,192 respondents) — February 2026; About the FinOps Foundation; Storment & Fuller, *Cloud FinOps* (O'Reilly) — December 2019

Flexera — 2026 State of the Cloud Report (753 respondents) — March 2026 (vendor survey)

Stanford HAI — AI Index Report 2025 — April 2025

Gartner — Gartner Predicts That by 2030 Performing Inference on an LLM With 1 Trillion Parameters Will Cost GenAI Providers Over 90% Less Than in 2025 — March 25, 2026 (forecast)

Menlo Ventures — 2025: The State of Generative AI in the Enterprise — December 2025 (investor estimate)

Linux Foundation — Linux Foundation announces the intent to launch the Tokenomics Foundation — June 3, 2026; Tokenomics Foundation launch — August 4, 2026

An EvoNova Advisors Insight, synthesizing published survey research from Deloitte, Pigment, Datarails, the FinOps Foundation and Flexera with vendor pricing documentation and analyst forecasts. Vendor-sponsored surveys are labeled; forecasts are labeled as forecasts; cloud-era figures are cited as historical context.