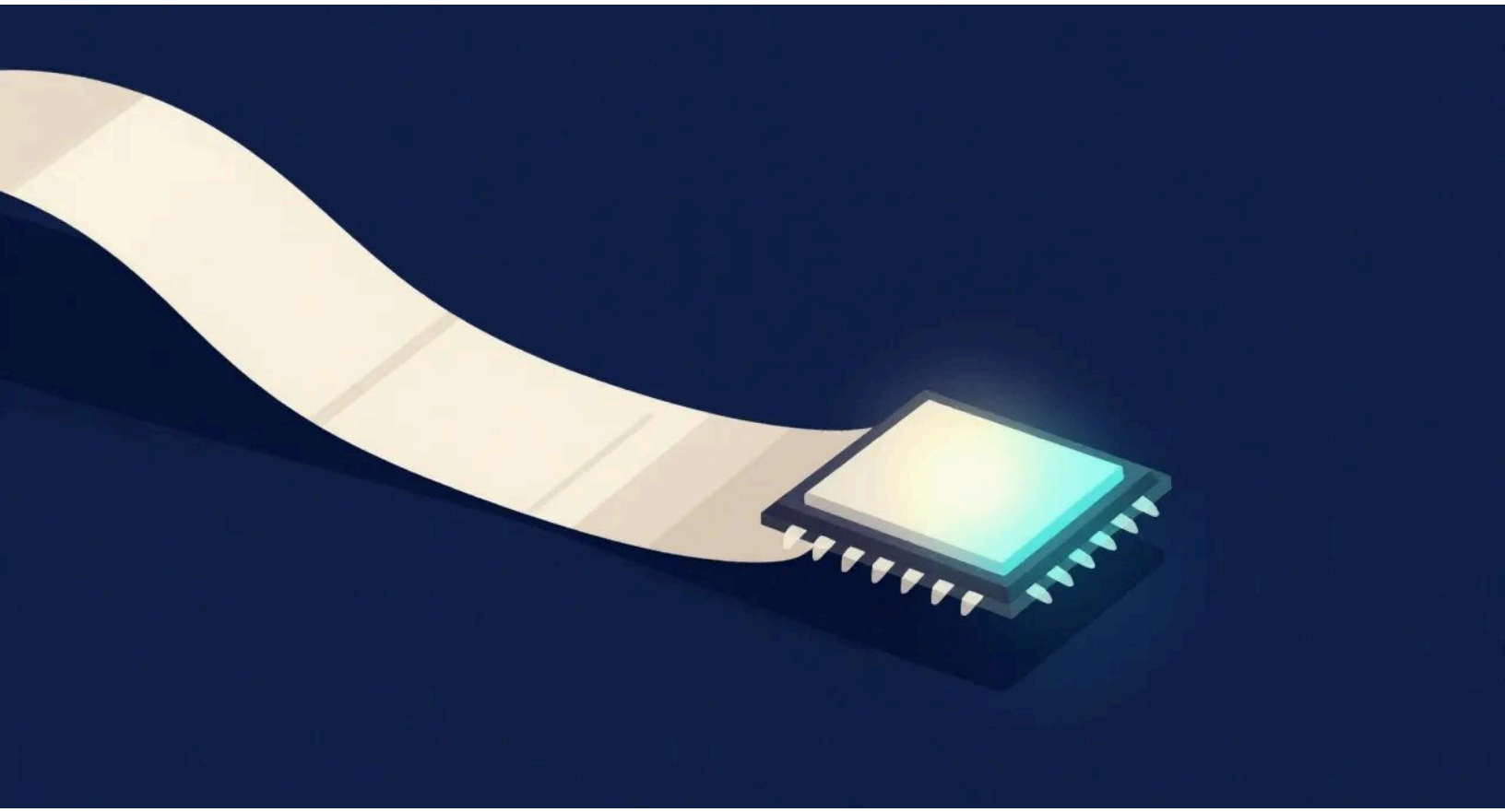


EXECUTIVE PAPER · AI

The Real Cost of AI in Production

The pilot invoice tells a CFO almost nothing about what an AI system costs once customers use it every day. Budget for the full stack of operating costs and judge each system by its cost per outcome.

AUTHOR

Chris Watkins

SERVICE

AI

FOR

CFO · 12 min read

AT A GLANCE

The problem. AI in production carries an operating cost that scales with every request: model calls, embeddings, retrieval infrastructure, evaluation runs, logging, retries, and the people who handle what the model cannot. The same workload can cost three times as much depending on design choices most finance teams never see.

Why it persists. Pilots are funded as projects, so their small bills set expectations. Pricing is per token and changes every few months. Costs are scattered across cloud invoices, marketplace charges, and software subscriptions, and few companies tie any of them to a business result.

The recommendation. Give every production AI system an explicit operating budget with a named owner. Measure cost per completed outcome, including human fallback, and review it monthly. Use cloud APIs by default, and move a workload to self-hosted models only when sustained utilization or a data constraint makes the numbers favor it.

The Decision in Front of You

Most companies approved their first AI systems as experiments, with pilot bills small enough to fit an innovation budget. The decision now is how to fund and control those systems once they run daily operations.

Gartner's May 2026 forecast put worldwide AI spending at $2.59 trillion for 2026, up 47%, with more than $32 billion going to AI models alone [1]. A 2025 CloudZero survey of 500 US software leaders found average monthly AI spending rising from $62,964 in 2024 to a planned $85,521 in 2025, and the share of organizations planning to spend more than $100,000 a month more than doubled [2]. Only 51% strongly agreed they could track the return [2].

The CFO has to settle how much operating budget each AI system gets and which measure decides whether it earns that budget. A related choice follows once volume grows: whether to keep renting model capacity from a provider or run models on hardware the company controls.

What's Happening

AI is now a measurable share of the cloud bill

Across CloudZero's customers, AI and machine learning services averaged 1.34% of cloud spend in 2024 [3]. By late 2025 it was running near 5%, after peaking at 6.5% in September, roughly triple the 2024 level [3]. The FinOps Foundation's 2026 survey found 98% of FinOps teams now manage AI

spending, up from 31% in 2024, and AI cost management was the skill they most wanted to add [4]. In Flexera's 2026 survey, 30% of respondents named cost unpredictability as a top challenge in scaling AI workloads [5].

Pricing is per token, and the spread is wide

A token is a fragment of text, roughly a short word. Providers charge per million tokens, with separate rates for text sent in and text generated. Anthropic lists its largest current model at $4 per million input tokens and $20 per million output tokens, and its smallest at $0.10 and $0.50 [6]. OpenAI's GPT-5.4 and GPT-5.5 models range from $0.20 to $5.00 per million input tokens and from $1.25 to $30 per million output tokens, according to a pricing tracker that checks the provider's published rates [7]. Embeddings, the numeric representations used to search company documents, cost far less: Google charges $0.15 per million tokens [8].

The discounts matter as much as the list prices. Text the provider has cached from a previous request is billed at a fraction of the normal input rate: a twentieth on Anthropic's larger models and about a tenth on OpenAI's [6][7]. Both offer half-price batch processing for work that can wait [6][7]. Data rules carry a price too: Anthropic adds 10% when a customer requires processing inside the United States [6].

The bill is hard to forecast

The FinOps Foundation now runs formal training on AI costs because the price mechanics differ from ordinary cloud spending [9]. Each model counts tokens differently, so the same request costs different amounts on different models, and reserved capacity lowers the rate while capping throughput at peak [9]. In mid-2026 the Linux Foundation and the FinOps Foundation announced plans for a Tokenomics Foundation to standardize token cost measurement. Coverage of the announcement noted that per-token price declines are easing and newer models often cost more per token [10]. A cloud provider's FinOps lead warned at a 2026 conference that one unguarded experiment can consume a quarterly budget in days [11].

Two recent cases

An AI bill that slipped past anomaly detection, May 2026. A customer running Anthropic's Claude Opus through Amazon Bedrock expected a bill in the hundreds of dollars and received a Bedrock inference charge of $30,141 [12]. The customer had configured AWS Cost Anomaly Detection to flag any increase over $100 or 40%, and no alert arrived. The charges were billed through AWS Marketplace, which that tool did not cover at the time, and startup credits had silently absorbed the first $8,026 [12]. AI spend often arrives through billing channels that existing cost controls do not watch.

Exposed keys turned into AI charges, April and May 2026. Attackers scraped public Google Cloud API keys and used them to call Gemini models. One company chief executive whose bills normally stayed under $50 a month was charged $10,138, with $8,000 of it arriving within about five minutes [12]. A developer with a $250 cap was automatically moved to a higher spending tier and billed about $12,000

[12]. Google reversed the charges for two of the developers after press coverage [12]. Keys that can call a paid model deserve the controls a company already applies to corporate cards: a named holder, a limit, and a record of use.

Unbudgeted AI does not stay unused

IBM's 2026 Cost of a Data Breach study found employee use of this "shadow AI" in 43% of the incidents it studied, up from 20% in 2025 [13]. Those breaches cost $5.39 million on average, about $760,000 more than breaches without shadow AI involvement. Among organizations with an AI-related incident, 68% lacked AI governance policies and 92% lacked proper AI access controls [13]. A sanctioned tool with its own budget line gives employees an alternative and gives the company a record of where its data goes.

Why the Usual Approach Fails

The pilot sets the budget. Pilots run on clean data and short prompts. Production adds longer conversations, retrieved documents in every request, retries, and safety checks, each of which multiplies token volume.

Token price gets all the attention. Teams compare list prices and overlook how often a person has to finish the job. Exhibit 1 shows the human fallback costing more than every model call combined.

Evaluation is left out of the budget. Careful teams test every prompt change and model upgrade against a library of sample cases, often with a second model grading the answers. That cost recurs with every new model release.

Costs are scattered across cloud invoices, marketplaces, direct contracts, and software subscriptions. CloudZero found 57% of companies were still tracking AI costs in spreadsheets and 15% had no formal tracking at all [2].

The Framework: Cost per Outcome

The measure that matters is the fully loaded cost of one completed business outcome, such as a resolved ticket or a processed invoice, including the human work for cases the system cannot finish. Version 1.5 of the FOCUS billing data standard, due December 2026, adds native AI token tracking [11].

In the hypothetical example below, a mid-size distributor uses an AI assistant to resolve 100,000 customer service emails a month. The assistant searches company documents, drafts a reply through several model calls, and escalates to a person when it cannot resolve the case.

Exhibit 1. Illustrative cost per ticket: design choices triple the AI bill, and human fallback outweighs both

Cost line (per ticket)	Tuned: mid-tier model with caching	Baseline: mid-tier model, no caching	Unmanaged: largest model everywhere
Model calls, including 10% retries	$0.0838	$0.1716	$0.3432
Query embedding	$0.00005	$0.00005	$0.00005
Retrieval infrastructure	$0.0150	$0.0150	$0.0150
Production grading of a 5% sample	$0.0025	$0.0025	$0.0025
Logging and monitoring	$0.0050	$0.0050	$0.0050
Release evaluation runs	$0.0180	$0.0180	$0.0180
AI operating cost per ticket	**$0.1244**	**$0.2121**	**$0.3837**
AI cost per AI-resolved ticket (80% resolved)	$0.16	$0.27	$0.48
Human handling of 20% escalations	$1.20	$1.20	$1.20
Fully loaded cost per ticket	**$1.32**	**$1.41**	**$1.58**
Monthly AI operating cost	$12,437	$21,214	$38,375
Annual AI operating cost	about $149,000	about $255,000	about $460,000

Source: author's illustrative estimate. Assumptions: 100,000 tickets a month; six model calls per ticket at 10,000 input and 600 output tokens each; 10% retried in full. Rates per million tokens from Anthropic's pricing [6]: mid-tier $2 in, $10 out; largest $4 in, $20 out; cached input $0.10, with 70% of input cached in the tuned case. Embedding: 300 tokens at $0.15 per million [8]. Retrieval $1,500 a month; logging $0.005 per ticket. Grading: 5% of tickets, largest model, 10,000 in and 500 out. Release testing: four runs a month of 2,000 cases at baseline cost plus grading, about $1,773. Escalation: $6.00, ten minutes of a $36 per hour loaded agent. Excludes build cost and taxes.

Design choices swing the AI bill by a factor of about three. The unmanaged version costs about $311,000 a year more than the tuned version for identical work. Caching repeated context and choosing a mid-tier model account for the entire difference.

The human fallback is the largest line. At an 80% resolution rate, escalations cost $1.20 per ticket, about ten times the tuned AI cost. Raising the resolution rate from 80% to 85% would save $0.30 per ticket, or $30,000 a month, which is more than the tuned system's entire AI bill. A CFO should ask for resolution rate before asking for a cheaper model.

Evaluation is a real line item. In the tuned case, release testing and sampled grading together make up about 16% of the AI operating cost. Cutting them removes the evidence that the system still works after each change.

Rent or run

Self-hosting means running an openly licensed model on servers the company controls or leases, trading the per-token charge for a fixed monthly cost in hardware and staff. It pays only when the hardware stays busy, since idle capacity costs the same as used capacity.

For the tuned case, assume a dedicated inference server at a fully loaded $20,000 a month plus half an engineer at $10,000 (author's assumptions). At $0.0838 of model cost per ticket, the API bill passes $30,000 a month at about 358,000 tickets, so the distributor at 100,000 should stay on the API. That breakeven also assumes the self-hosted model resolves as many tickets as the commercial one. If its resolution rate is two points lower, the extra escalations add $12,000 a month at 100,000 tickets and push breakeven further out.

Capacity has its own lead time. GEICO's infrastructure chief said in 2024 that cloud providers wanted three-year reserved pricing up front and nine months to deliver dedicated AI capacity, which made a cloud operating expense behave like a capital purchase [14].

Exhibit 2. Cloud API or self-hosted: decide on utilization and data constraints

Factor	Favors cloud API	Favors self-hosted
Volume and utilization	Low, uneven, or growing unpredictably	High and steady enough to keep hardware busy most hours
Model quality needed	Task needs the strongest commercial models	An openly licensed model meets the quality bar on your own test set
Data constraints	Provider terms and regional processing options satisfy your obligations, at a known surcharge [6]	Contracts or regulation prohibit sending the data to a third party
Staff	No team experienced in running GPU servers	An operations team already runs production hardware
Capacity timing	Shared capacity on demand, within rate limits	Can plan dedicated capacity a year ahead; cloud GPU reservations carry similar lead times [14]
Price change risk	Accept that providers reprice every few months [10]	Prefer a fixed cost and accept the hardware refresh cycle

Source: author's analysis; supporting facts as cited.

Costs, Risks, and Tradeoffs

Exhibit 3. What disciplined AI cost management costs, and when it is the wrong call

Area	Cost or risk	Mitigation
Measurement work	Tagging model calls by system and outcome takes engineering time	Route all calls through one gateway that records cost and purpose
Evaluation spend	Test suites and grading add a recurring cost	Size the suite to the risk of the task; run full suites only on model or prompt changes
Over-optimizing tokens	Teams cut context or switch to cheaper models and lose resolution rate	Judge every change on cost per outcome, never on token cost alone
Spending controls	Hard limits can stop a production system mid-day	Set alerts at 50% and 80% of budget and a hard stop only on non-production keys
Self-hosting	Idle hardware, staffing, and model upgrades become your problem	Self-host only after sustained volume clears the breakeven with margin
When this is wrong	A company with one small assistant and a bill of a few hundred dollars a month gains little from per-outcome accounting	Set a monthly cap, restrict keys, review quarterly

Source: author's analysis.

How to Act

Exhibit 4. A one-year roadmap

Window	Actions	Owner	Milestone
First 30 days	Inventory production AI systems, provider accounts, marketplace subscriptions, and API keys that can call paid models. Name one budget owner per system.	CFO with CTO	Complete inventory; every system has an owner and a monthly budget
First 30 days	Restrict or rotate keys found in code or shared files. Set spending alerts on every billing channel, including cloud marketplaces.	Security lead	No unrestricted keys; alerts tested on each channel
First 90 days	Route model calls through a single gateway that logs tokens, cost, system, and outcome. Define the business outcome for each system.	Engineering lead	Cost per outcome reported for the top three systems by spend
First 90 days	Turn on caching and batch processing where they apply. Assign routine steps to smaller models and test the effect on resolution rate.	Engineering lead with product owner	AI cost per outcome down with no loss in resolution rate
First 365 days	Make AI operating cost its own budget line with a forecast per system. Fund evaluation suites. Run the rent-or-run analysis for systems near breakeven volume.	CFO with CTO	AI budget variance within 15% of forecast; written sourcing decision for each high-volume system
Ongoing	Monthly review of cost per outcome and resolution rate for each production system.	CFO and system owners	Review held monthly with owners present

Source: author's recommendation.

Questions to Ask Your Team

1. Which AI systems are in production today, and who owns the budget for each one?
2. What does one completed outcome cost for each system, including the people who handle what the AI cannot, and how has that moved since launch?
3. Through how many invoices, marketplaces, and contracts do we pay for AI, and does our anomaly detection cover all of them?
4. Which API keys can spend money on paid models, and where are they stored?
5. How much do we spend testing a new model version before we switch to it?
6. At what monthly volume would running our own models cost less, and how far are we from it?

Closing

AI in production is an operating cost that moves with volume, design choices, and provider pricing, much like freight or payment processing. Give each system its own budget line and owner, measure it by the cost of a completed outcome including human fallback, and keep it on cloud APIs until sustained volume or a data obligation makes running your own models the cheaper and safer choice.

Sources

[1] CIO Dive. Coverage of Gartner forecast of worldwide AI spending in 2026. 2026. https://www.ciodive.com/news/global-AI-spending-2026/820656/.

[2] CloudZero. "The State of AI Costs." Survey of 500 US software professionals, March 2025. 2025. https://www.cloudzero.com/state-of-ai-costs/.

[3] CloudZero. Cloud Economics Pulse, November 2025. 2025. https://www.cloudzero.com/blog/cloud-economics-pulse-november-2025.

[4] FinOps Foundation. "State of FinOps 2026." 2026. https://data.finops.org/.

[5] CFOtech UK. Coverage of Flexera 2026 State of the Cloud Report. 2026. https://cfotech.co.uk/story/flexera-warns-ai-cloud-costs-strain-technology-budgets.

[6] Anthropic. "Pricing." Claude Platform documentation. https://platform.claude.com/docs/en/about-claude/pricing.

[7] Morph. "OpenAI API Pricing (2026): GPT-5.5, GPT-5.4 and the Full Per-Token Table." Third-party tracker of OpenAI published rates. 2026. https://morphllm.com/openai-api-pricing.

[8] Google. "Gemini Embedding now generally available in the Gemini API." Google Developers Blog. 2025. https://developers.googleblog.com/en/gemini-embedding-available-gemini-api/.

[9] Techzine. "Why is AI so hard to cost? FinOps Foundation codifies 'price mechanics' of AI." 2025. https://www.techzine.eu/blogs/devops/131881/why-is-ai-so-hard-to-cost-finops-foundation-codifies-price-mechanics-of-ai.

[10] IT Brief UK. Coverage of the Linux Foundation and FinOps Foundation Tokenomics Foundation announcement. 2026. https://itbrief.co.uk/story/linux-foundation-launches-tokenomics-foundation-for-ai-costs.

[11] nOps. Recap of FinOps X 2026 day 2 keynotes. 2026. https://www.nops.io/blog/finops-x-2026-keynote-day-2/.

[12] WebProNews. "Cloud Giants Face Backlash Over Unchecked AI Charges." Coverage of reporting by The Register. 2026. https://www.webpronews.com/cloud-giants-face-backlash-over-unchecked-ai-charges/.

[13] Baker Donelson. "Ten Takeaways from IBM's 2026 Cost of a Data Breach Report." 2026. https://www.bakerdonelson.com/ten-takeaways-from-ibms-2026-cost-of-a-data-breach-report.

[14] The Stack. Interview with GEICO's vice president of platform and infrastructure engineering on cloud repatriation. 2024, updated 2025. https://thestack.technology/warren-buffetts-geico-repatriates-work-from-the-cloud-continues-ambitious-infrastructure-overhaul.

The views in this paper are the author's own and do not represent any employer. Scenarios are hypothetical. This paper is not legal, financial, or compliance advice.