David Soden

MARKET ASSESSMENT | AUGUST 2026

The Unit Economics of Agentic AI

Per-task token counts are outrunning per-token price declines. What that does to enterprise AI budgets, and to SaaS margins if agents replace seats as the billing unit.

David Soden

Independent analyst and practitioner, enterprise automation and applied AI

davidsoden.com/market-assessment

REPORT DS-2026-24FB · 2026 · VERSION 1.0 · PUBLIC

About this report

Every substantive claim in this report carries an evidence classification and, where applicable, a numbered source. Facts are separated from vendor claims, third-party estimates, the author's assessments, and untested hypotheses. Forward-looking sections use scenarios with observable tripwires rather than point forecasts. Stated assumptions are listed before the analysis begins.

Market Assessment reports are published and commissioned at **davidsoden.com/market-assessment**. This document is analysis, not investment, legal, or procurement advice.

Scope and evidentiary standard

This assessment covers the cost structure of agentic AI as it appears on an enterprise P&L: what a completed agent task costs, how that cost is moving, and what happens to software pricing if the billing unit shifts from a seat to an agent. It runs from mid-2025 through 16 August 2026, the date DeepSeek's revised API pricing took effect. It does not cover model capability, agent frameworks, or the security posture of autonomous systems except where those bear directly on cost.

The published numbers in this category need heavier discounting than usual, for a specific reason. Almost every circulating statistic about agentic token consumption is a multiplier, and a multiplier is only as good as its denominator. The denominators here are not standardised and are frequently not disclosed. There is no agreed definition of a task, no agreed boundary for what counts as one agent run, and no convention for whether failed attempts and retries are counted. A figure like "agents use fifteen times more tokens" can therefore be true and useless at the same time.

Compounding that, the sources with the strongest incentive to publish these multipliers are the vendors selling cost observability, token optimisation, caching layers, and FinOps tooling. That does not make their numbers wrong. It does mean an unattributed multiplier from a vendor blog is treated here as an illustration, not as a measurement, and is labelled accordingly.

Eight anchor claims framed this research and each was checked against primary sources. Three did not survive in the form they circulate, and one survived only with its attribution corrected. Rather than quietly dropping them, this report reproduces the audit in full, because which numbers fail is itself part of the finding.

Label	Definition	How to treat it
FACT	Verifiable in the cited primary source: a filing, press release, official documentation, or standards body announcement.	Reliable as stated. Check the source if the exact wording matters.
VENDOR CLAIM	Reported by a vendor about its own business or product. Self-measured and not audited by any third party.	The vendor said it. Directionally useful, not a measurement. Definitions of the metric are usually undisclosed.
THIRD-PARTY ESTIMATE	A research firm forecast or survey result. Methodology is proprietary and generally not reproducible.	Use for direction and order of magnitude only. Do not build a model on it.
ASSESSMENT	The author's reasoned judgment from the cited evidence. Falsifiable, and the reasoning is shown so it can be argued with.	This is analysis, not reporting. Disagree with it where your evidence differs.
HYPOTHESIS	A proposition offered for testing. Not currently supported by sufficient evidence to assert.	Treat as a research question, not a conclusion. Each one names what would confirm or refute it.
SCENARIO	A structured future state with a stated subjective probability. The probability is the author's judgment, not a calculated figure.	Use the leading indicators attached to each, not the probability. The indicators are observable; the probability is not.

Assumptions this analysis rests on

Five premises are assumed rather than demonstrated. Each is stated with what fails if it turns out to be wrong.

Frontier per-token prices keep falling in aggregate, but unevenly

Independent measurement shows inference prices falling at very different rates depending on the capability being purchased, and the fastest declines are the newest and least likely to persist. This analysis assumes the capability tier agents actually run on, which is the reasoning tier and not the cheapest tier, declines more slowly than headline averages suggest. If this is wrong and frontier reasoning prices fall as fast as commodity chat prices did, the cost-per-task problem resolves itself inside twenty-four months and the repricing scenarios in this report lose most of their urgency.

Reported adoption percentages measure intent and light usage, not production load

Survey figures showing most large organisations using agentic AI are read here as covering everything from a single departmental pilot to genuine production deployment. If this is wrong and reported adoption already reflects production-grade workloads, aggregate token demand is

materially higher than any figure in this report, and the capacity argument strengthens rather than weakens.

Agent task design does not become dramatically more token-efficient in the next four quarters

Cache reuse, smaller routing models, and shorter reasoning traces all reduce tokens per task. This analysis assumes those gains are real but incremental, in the range of a few multiples rather than an order of magnitude. If a step change arrives, from persistent memory tiers or from architectures that stop replaying full context on every loop iteration, the two curves in the next section converge and the central argument weakens substantially.

Compute capacity, not commercial strategy, sets the near-term price floor

The report reads time-of-day inference pricing as evidence that serving capacity binds at peak. If this is wrong and it is simply margin recovery by one provider under investor pressure, then it is a company-specific event and carries no signal about the wider price floor. This is the single assumption most likely to be wrong, and the leading indicators section gives the test.

Buyers will accept variable bills provided the variance is bounded

The repricing scenarios assume procurement can be brought to consumption pricing if it comes with enforceable caps. If buyers instead insist on fixed prices regardless of consumption, vendors absorb the entire cost distribution and the gross margin consequences land on the vendor side rather than being shared.

The call

ASSESSMENT Enterprise AI budgets built on falling token prices will not break on price. They break on variance. Per-token costs have collapsed while per-task consumption has risen faster, and the resulting cost per completed task is not merely higher, it is unbounded on the right tail. That is what stalls deployments that have already cleared data quality and governance: you cannot underwrite a business case against a cost distribution you cannot cap. Cheaper tokens do not fix an unbounded tail. Hard caps do. I would abandon this position if per-task token consumption flattens for two consecutive quarters while autonomous multi-step deployment keeps climbing.

My judgment on the evidence assembled here, nothing more. There is data I can't see and data I may have missed. New evidence can move me, including to the opposite position, and I reserve the right to move.

Executive summary

The question put to this research was whether the stall between pilot and production is better explained by cost per completed task than by data quality and governance. The answer is that the cost argument is correct about the mechanism and wrong about the word binding. Cost per task

is not a wall. It is a distribution, and it is the shape of that distribution rather than its mean that stops deployments.

FINDING 1 The two curves have crossed, and the crossing is measurable.

THIRD-PARTY ESTIMATE Independent measurement puts the price of reaching a given capability milestone falling by anywhere from nine to nine hundred times per year, depending on the benchmark. Over roughly the same window, EY puts the cost of a single customer service interaction at four cents in 2023 and one dollar twenty in 2026, about thirty times higher. Both can be true only if tokens consumed per completed task rose faster than price per token fell. [9] [10]

FINDING 2 The price floor stopped falling in August 2026, and started tiering.

FACT DeepSeek raised V4-Pro output from eighty-seven cents per million tokens to one dollar ninety-eight off-peak and three dollars ninety-six at peak, effective 16 August 2026. The increase matters less than the tiering. Time-of-day pricing is what a provider introduces when serving capacity binds at peak rather than when unit costs rise, and it is the first such move by a major provider positioned on price. [2] [3]

FINDING 3 Cost does not appear in the barrier lists, and the barrier lists are the wrong instrument.

VENDOR CLAIM ServiceNow's index of 4,500 senior leaders names IT infrastructure at 71 percent, integration at 47 percent and data accuracy at 45 percent. Cost is absent. This is the strongest evidence against this report's argument and it is presented as such. The counter is that surveys record what respondents attribute, and nobody tells a surveyor that the business case did not close. [6]

FINDING 4 The fifty-point gap between adoption and autonomy is where the economics actually bite.

ASSESSMENT Fifty-nine percent of organisations report using agentic AI and nine percent report reaching autonomous multi-step workflows. That population has already cleared data quality, because it is running agents in production. What separates the fifty-nine from the nine is step count, and step count is the direct multiplier on both token consumption and failure retries. [6]

FINDING 5 Gartner's own cancellation forecast names escalating cost first.

THIRD-PARTY ESTIMATE Gartner predicts more than forty percent of agentic AI projects will be cancelled by the end of 2027, citing escalating costs, unclear business value and inadequate risk controls, in that order. The same firm projects agentic AI in a third of enterprise software applications by 2028, up from under one percent in 2024. Those two predictions are only compatible if a large share of what ships is narrow enough to be cheap. [4] [5]

FINDING 6 The intervention that worked was a cap, not a discount.

FACT Uber consumed its entire 2026 budget for one AI coding tool by April. Its response was a hard ceiling of fifteen hundred dollars per employee per month per tool. Users of advanced AI tools then more than quadrupled while total AI cost fell. If the constraint had been the price

level, cheaper tokens would have fixed it. A ceiling fixed it, which says the constraint was unbounded variance. [11]

FINDING 7 Single vendors are now running three incompatible billing units at once.

THIRD-PARTY ESTIMATE Salesforce is reported to offer Agentforce per conversation, as consumption credits, and per user per month simultaneously. That is not a transitional state to be tidied later. It is what happens when a vendor cannot predict the cost of its own product well enough to pick one unit, and it is the clearest signal in the market that seat pricing is under structural strain. [21] [12]

FINDING 8 The infrastructure bill is reweighting away from the GPU.

VENDOR CLAIM Red Hat, citing Intel, describes CPU-to-GPU ratios moving from one to eight in training toward one to one in agentic deployments, with some customers running four CPUs per GPU. Silicon Motion is positioning enterprise SSDs as a persistent tier for KV cache offload. Both point the same way: agentic cost is migrating out of the accelerator and into orchestration, memory and storage, which are budgeted and procured by different people. [7] [8]

FINDING 9 Cloudflare's margin collapse is restructuring, not agent capex, and should not be cited as evidence of agent economics.

FACT Cloudflare's GAAP operating margin moved from negative 13.1 percent to negative 29.6 percent on revenue of 696.1 million dollars, up 36 percent. A 150.7 million dollar restructuring charge accounts for roughly 21.6 points of that 16.5-point swing. Excluding it, operating margin improved year over year. The agentic build-out is visible in Cloudflare's product line, not in its margin. [1]

FINDING 10 Three of the most-circulated numbers in this debate do not survive sourcing.

ASSESSMENT The seventeen-times multiplier traces to an illustrative fraud-detection example, not a measured comparison against retrieval. The three-cents-to-fifteen-cents transaction cost has no locatable primary source. The 4.7x six-month increase attributed to Apptio in federal deployments appears to originate in an April 2026 startup blog as a hypothetical. Each is examined below. [18] [19]

The two curves

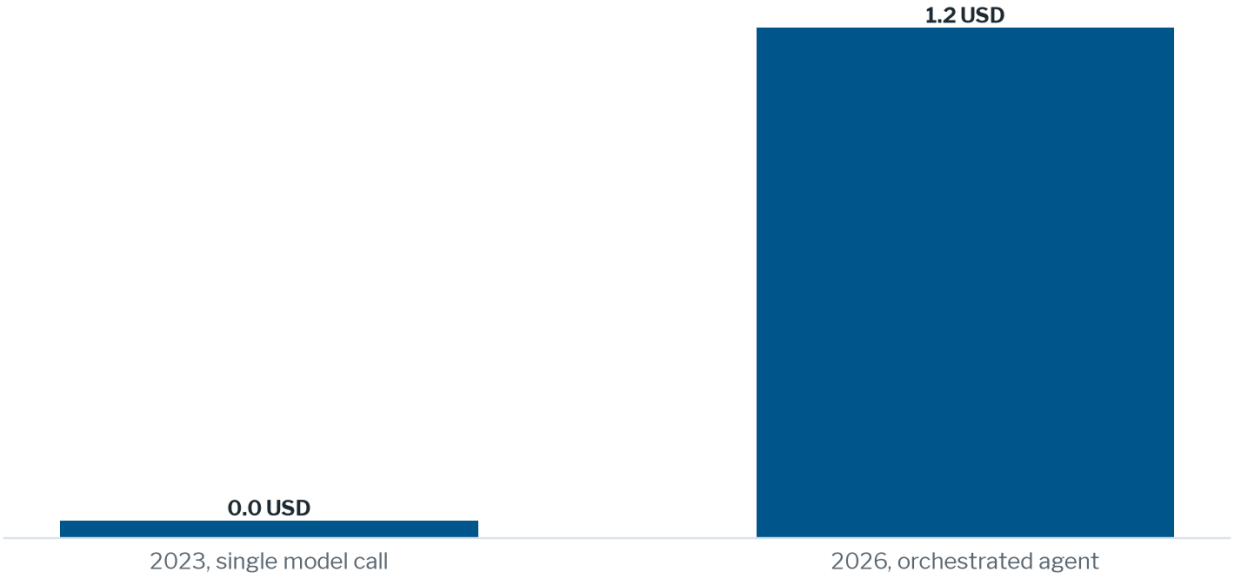
FACT Epoch AI, which tracks inference pricing against fixed capability milestones rather than against model names, found the price of reaching a given benchmark score falling at rates between nine and nine hundred times per year across six benchmarks. The price of matching GPT-4 on PhD-level science questions fell roughly forty times per year. Epoch also recorded the important caveat: the fastest declines in that range occurred in the most recent year measured, so they are the least likely to persist. [9]

ASSESSMENT That caveat is doing more work than the headline. A buyer planning a three-year budget against the idea that prices fall by an order of magnitude annually has extrapolated from the fastest and youngest part of a distribution. The measurement is also anchored to fixed

capability, which means it tracks the falling price of yesterday's intelligence. Agents are not built on yesterday's intelligence. They are built on whatever currently reasons well enough to plan a multi-step task without derailing, and that tier is repriced by capability competition rather than by commoditisation. [9]

THIRD-PARTY ESTIMATE Against that, EY's practice put the cost of a customer service interaction at four cents in 2023 using a single model call, and one dollar twenty in 2026 using an orchestrated agent, roughly thirty times higher. EY also reported that nearly a third of developers exhaust their monthly token allotment early. These are consulting observations rather than an audited dataset, and the two endpoints are not the same product doing the same job, which is precisely the point: the job got redefined and the cost followed the redefinition, not the price list. [10]

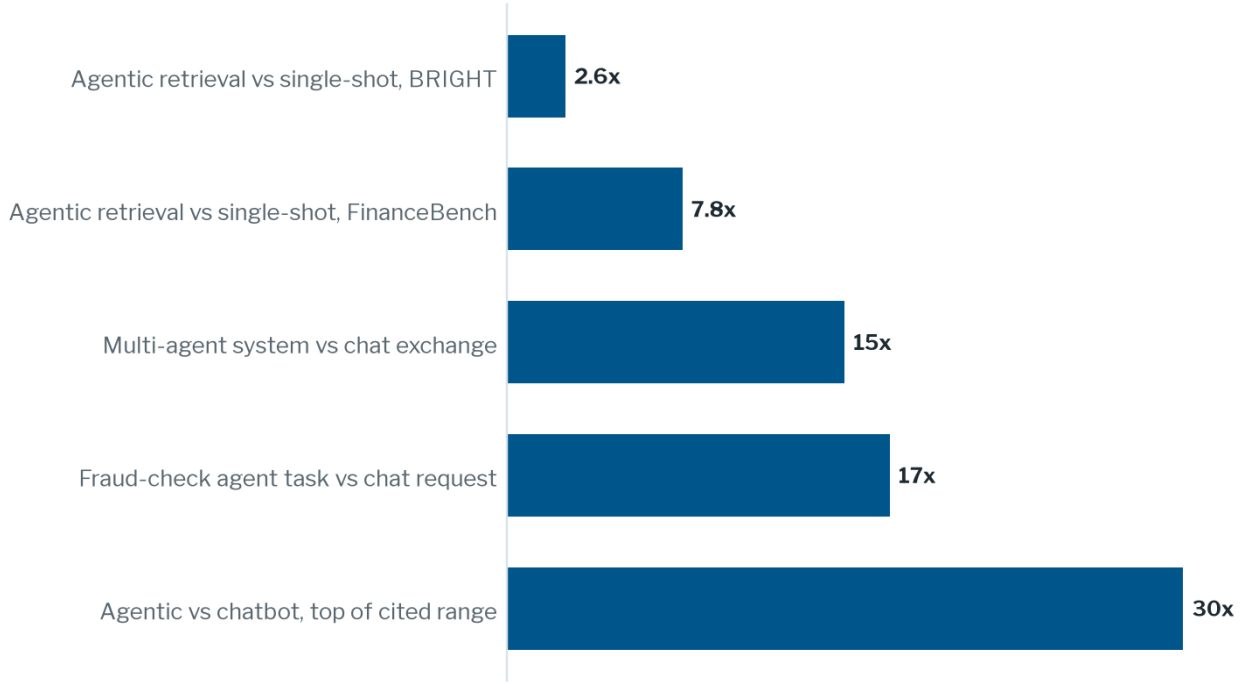
FIGURE 1 Cost of one customer service interaction, as the unit of work was redefined



THIRD-PARTY ESTIMATE EY practice figures, June 2026. Not an audited dataset and not a like-for-like comparison: the 2026 interaction resolves more of the request without a human. That is the mechanism, not a flaw in the number. Per-token prices fell steeply across the same period. [10]

ASSESSMENT Put those two measurements together and the arithmetic is unavoidable. If price per token fell by a large multiple while cost per completed interaction rose by roughly thirty times, tokens consumed per interaction rose by the product of the two. Nothing in the published record contradicts this. What the published record cannot do is pin the multiplier down, because the available figures measure different things against different baselines. [9] [10]

FIGURE 2 **Circulating token multipliers, and the baselines they are measured against**



THIRD-PARTY ESTIMATE These are not five measurements of one quantity. The first two come from a preprint measuring token counts on named benchmarks; the third is attributed to Anthropic in secondary reporting without a direct link; the fourth is an illustrative example; the fifth is the top of a range quoted secondhand from Gartner. Read the spread as a measure of how unstandardised this category is, not as a distribution around a true value. [17] [18] [20]

ASSESSMENT The honest summary is that agentic patterns consume several times more tokens per unit of work than single-shot retrieval, that the multiplier is workload-specific, and that anyone quoting a single number for it is quoting an anecdote. The two ends of the published range, 2.6 times on a general benchmark and 7.8 times on a finance corpus, are the most defensible figures available because they come from disclosed benchmarks with stated token counts. They also happen to be the lowest, which is worth noticing. [17]

Which numbers survived sourcing

Eight claims framed this research. What follows is what each turned out to be. It is reproduced in full because a reader deciding how much weight to put on the argument needs to see the same audit the argument was built on.

Claim as it circulates	What the primary sources show	Verdict
Agentic reasoning loops consume up to 17x the tokens of retrieval-augmented generation	Traces to an illustrative fraud-detection example comparing an 800-token chat request with a 13,500-token agent task. That is agent versus chat, not agent versus RAG. Published benchmark work puts agentic retrieval at 2.6x to 7.8x single-shot retrieval.	Does not survive as stated. Direction supported, magnitude unsourced, comparison misattributed.

Claim as it circulates	What the primary sources show	Verdict
Average transaction cost moves from roughly three cents to fifteen cents	No primary source located for either endpoint. The nearest sourced pair is EY's four cents in 2023 and one dollar twenty in 2026, which measures a different unit over a different period.	Not verified. Replaced with the sourced pair.
Gartner projects one third of enterprise software includes agentic AI by 2028	Gartner's published predictions state 33 percent of enterprise software applications by 2028, up from less than 1 percent in 2024, alongside 15 percent of day-to-day work decisions made autonomously.	Survives. Analyst forecast, proprietary methodology.
DeepSeek raised V4-Pro from 0.36 cents per million to 2.20 off-peak and 4.40 peak	Output tokens moved from 0.87 dollars per million flat to 1.98 off-peak and 3.96 at peak, effective 16 August 2026. Peak windows are 01:00 to 04:00 and 06:00 to 10:00 UTC. V4-Flash moved from 0.28 to 0.66 and 1.32.	Wrong figures, right event. Corrected throughout.
Apptio reported a 4.7x increase in model costs over six months in federal deployments	No Apptio or IBM publication containing this figure was located. The number appears in an April 2026 startup blog as a hypothetical monthly bill growing from month six to month eighteen, presented as the author's own illustration.	Does not survive. Dropped from the analysis.
Cloudflare Q2 CY2026 operating margin at -29.6 percent from -13.1 percent	Confirmed verbatim in the filed release. These are GAAP figures and include a 150.7 million dollar restructuring charge, roughly 21.6 points of margin.	Survives, with a caveat that changes its meaning.
ServiceNow reports 59 percent using agentic AI and 9 percent reaching autonomous workflows	Confirmed in the published index, drawn from 4,500 senior leaders across 19 countries and 12 industries.	Survives. Vendor-run, self-reported, self-selected sample.
Red Hat projects CPU-to-GPU ratios shifting from 1:8 toward 1:1 or 4:1	Red Hat published this on 6 August 2026 but explicitly attributes the projection to Intel rather than to its own measurement.	Survives with attribution corrected.

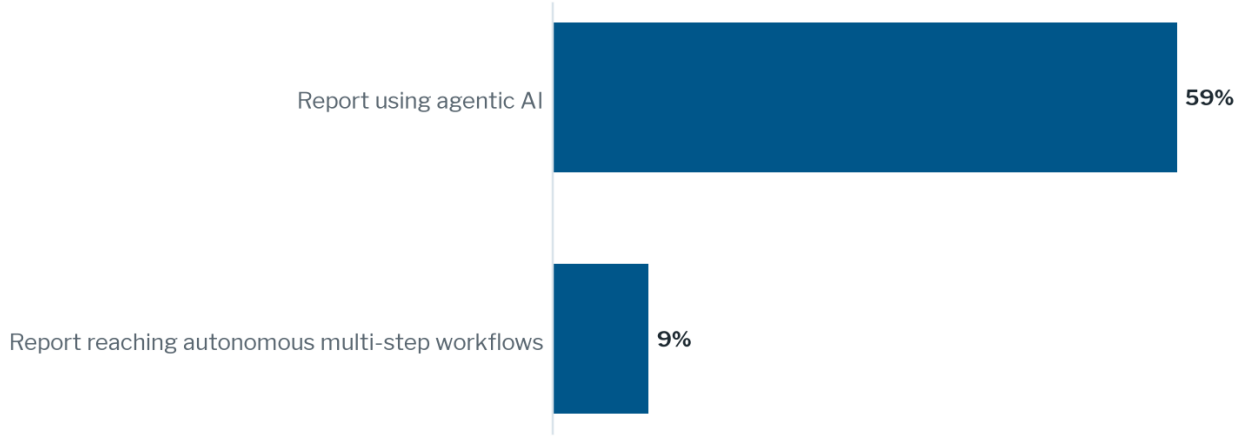
ASSESSMENT Two of the three failures share a pattern worth naming. A plausible figure appears in a blog as an illustration, loses its hedging when quoted, acquires an institutional attribution it never had, and then circulates as a measured finding. The Apptio case is the clearest: a hypothetical bill in a founder's blog post became a federal deployment statistic from an IBM company. Anyone building a budget model on figures gathered from secondary coverage should assume some proportion of it has been through this laundry. [19]

What actually binds

VENDOR CLAIM ServiceNow's Enterprise AI Maturity Index, drawn from 4,500 senior leaders across 19 countries and 12 industries, reports 59 percent of organisations using agentic AI and 9 percent reaching autonomous multi-step workflows. Average maturity scored 51, up from 35. Reported AI spend rose 110 percent year over year with a further 81 percent expected by 2027.

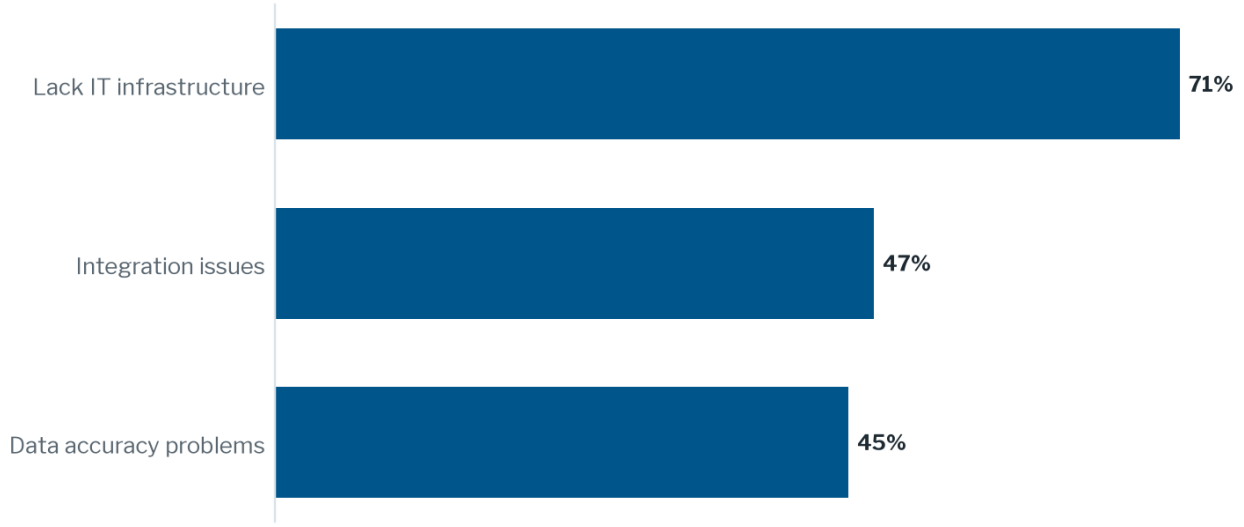
The named barriers are IT infrastructure at 71 percent, integration at 47 percent, and data accuracy at 45 percent. [6]

FIGURE 3 Adoption against autonomy: the fifty-point gap



VENDOR CLAIM ServiceNow Enterprise AI Maturity Index 2026, n=4,500 senior leaders. Vendor-commissioned, self-reported, self-selected respondents, and ServiceNow sells to both ends of this gap. Neither figure defines its threshold publicly, so using may include a single pilot. [6]

FIGURE 4 Barriers as respondents named them. Cost is not among them.



VENDOR CLAIM Same survey and same limitations. This chart is the strongest published evidence against this report's central argument and is included for that reason. Respondents were not asked to rank cost per completed task, so its absence is partly an artefact of instrument design. [6]

ASSESSMENT Taken at face value, that barrier list refutes the argument this report was asked to test. Cost is not named. Infrastructure, integration and data quality are. If the question is what executives say stops them, the conventional explanation wins on the evidence and this report should concede it. [6]

ASSESSMENT The concession is narrower than it looks, for three reasons. First, attribution surveys measure what respondents will say, and cost is the one blocker with a named owner and a career consequence. Infrastructure is nobody's fault. Second, the 71 percent citing

infrastructure is not evidence against a cost argument, because the reason agentic workloads strand on existing infrastructure is that they need more of it. Third, and decisively, the population inside the fifty-point gap has already passed the data quality gate. They are running agents in production. Whatever stops them at step three is not the thing that would have stopped them at step one. [6]

HYPOTHESIS The proposition offered for testing is that what separates the 59 percent from the 9 percent is the variance of cost per completed task rather than its level. A task whose cost is reliably fifteen cents can be funded. A task whose cost is fifteen cents at the median and eleven dollars at the ninety-ninth percentile, because a retry loop can run forty iterations against a tool that is timing out, cannot be underwritten at all. This would be confirmed by evidence that organisations reaching autonomous workflows differ from those that stall primarily in having deployed hard spend ceilings, and refuted by evidence that they differ primarily in data readiness scores. [6]

FACT The one well-documented natural experiment supports the hypothesis. Uber's engineering organisation consumed its entire 2026 budget for Anthropic's Claude Code by April, four months into the year, after an internal push that ranked staff on a usage leaderboard. The fix was a hard limit of 1,500 dollars per employee per month per AI coding tool. Uber's CTO subsequently said the number of employees on advanced AI tools had more than quadrupled while total AI cost was falling. [11]

ASSESSMENT That last detail is the finding. Quadrupling users while cutting spend is not what happens when a price falls. It is what happens when an unbounded distribution gets a ceiling and the workloads redistribute below it. The binding constraint at Uber was never the price of a token, and no reduction in token price would have prevented the April blowout, because the mechanism was consumption expanding to fill an unpriced allowance. Procurement has a name for this and it is not a cost problem, it is a commitment problem. [11]

THIRD-PARTY ESTIMATE Gartner's forecast that more than 40 percent of agentic AI projects will be cancelled by the end of 2027 lists escalating costs first among its reasons, ahead of unclear business value and inadequate risk controls. Analyst cancellation forecasts are unfalsifiable in practice, since a project quietly rescoped is not recorded as cancelled, so this is corroboration rather than proof. It is notable mainly because the ordering contradicts the barrier surveys from the same period. [4]

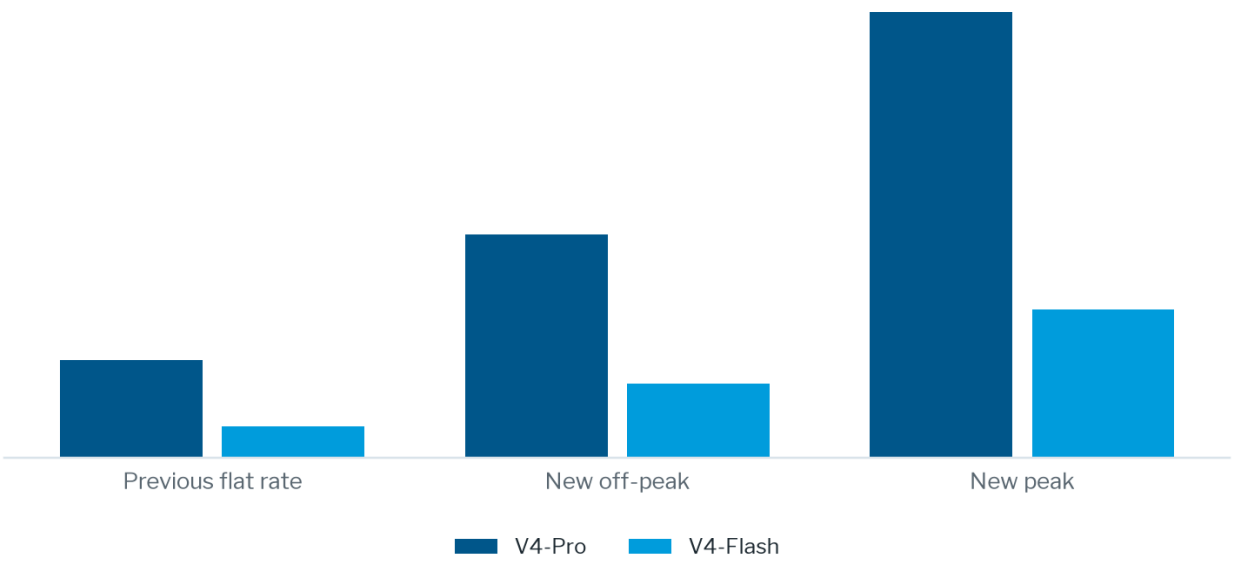
ASSESSMENT So the argument as originally put is half right. Cost per completed task is the mechanism, and per-task token growth is outrunning per-token price decline by a wide margin. But cost per task is not binding in the sense of a level nobody can afford. It binds because it cannot be forecast, and an enterprise cannot approve a business case against a cost it cannot bound. That is a different problem with a different fix, and the fix is contractual rather than technical. [11]

The price floor stopped falling

FACT Effective 16 August 2026, DeepSeek moved V4-Pro output pricing from 0.87 dollars per million tokens flat to 3.96 at peak and 1.98 off-peak. V4-Flash output moved from 0.28 to 1.32

and 0.66. Peak windows are 01:00 to 04:00 and 06:00 to 10:00 UTC. The company's stated rationale was to allocate resources more reasonably and to encourage developers to shift work to less congested periods. [2] [3]

FIGURE 5 DeepSeek output token pricing before and after 16 August 2026



FACT Output tokens only, in dollars per million. Input pricing changed on the same schedule. Peak is 01:00 to 04:00 and 06:00 to 10:00 UTC. Reported by Fortune from the company's announcement and reflected on the vendor's own rate card. [2] [3]

ASSESSMENT The increase is the headline and the tiering is the news. A provider raises a flat rate when its unit costs rise or its margin strategy changes. A provider introduces time-of-day pricing when its capacity binds at particular hours and it wants demand moved rather than revenue raised. Those are different economic conditions with different implications, and only the second one says anything about the wider market. [2]

ASSESSMENT This matters disproportionately because of who did it. DeepSeek's position in the market has been price. A provider whose entire competitive claim is that inference is cheap does not introduce congestion pricing casually, because doing so concedes that its capacity is finite in exactly the way its positioning implied it was not. Read that way, the tiering is a stronger signal about serving capacity than the price increase is about cost. [2] [3]

HYPOTHESIS The testable proposition is that time-of-day inference pricing spreads to at least one more provider with meaningful share within four quarters, and that when it does, it appears first on reasoning-tier models rather than on the cheap tier. Confirmed by any second frontier or near-frontier provider publishing peak and off-peak rates. Refuted if DeepSeek's tiering is withdrawn or remains unique through mid-2027, in which case it was a commercial decision by one company under margin pressure and carries no market signal at all. [3]

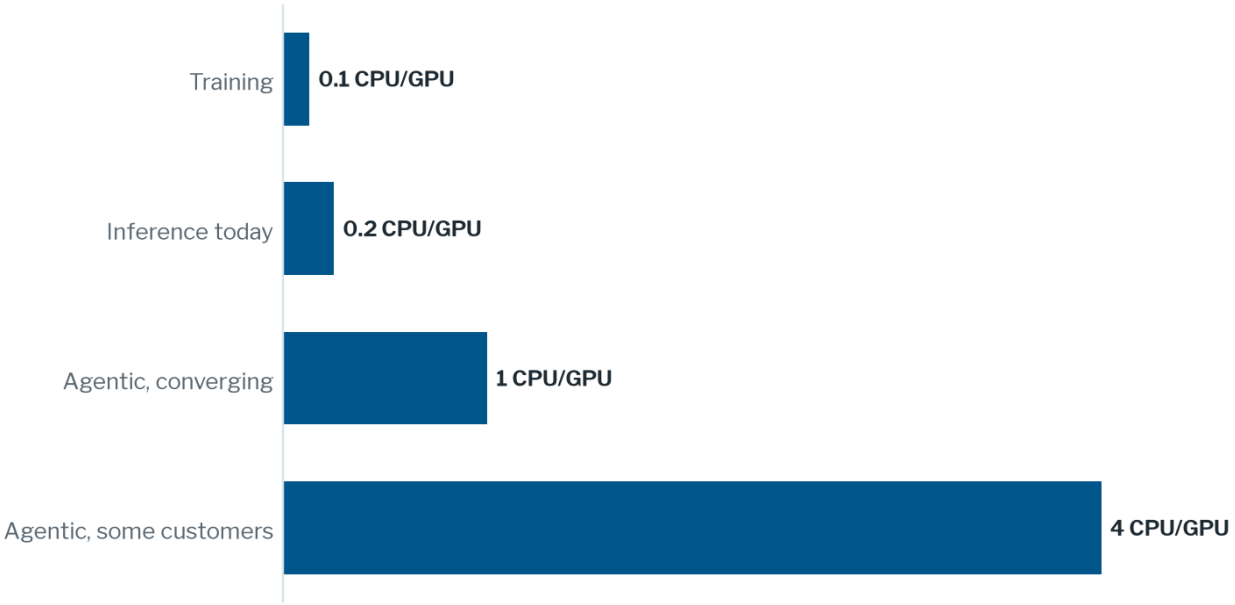
ASSESSMENT For a CFO the practical consequence is narrower than the debate. Time-of-day pricing converts a portion of AI spend into something schedulable. Batch-eligible agent work, which is most reconciliation, enrichment, summarisation and back-office review, can be moved

off peak for roughly half the token cost. Interactive agent work cannot. That split is now a real line in an architecture decision and it did not exist in July 2026. [2] [3]

The infrastructure bill reweights away from the GPU

VENDOR CLAIM Red Hat, in a post dated 6 August 2026, states that Intel has called out CPU-to-GPU ratios moving from one to eight in training workloads toward one to one, and in some agentic deployments four CPUs deployed per GPU. Red Hat also cites research finding that CPU-side tool processing accounts for 50 to 90 percent of end-to-end latency in agentic workloads. The projection is Intel's; Red Hat is reporting it, and Intel sells CPUs. [7]

FIGURE 6 **CPUs deployed per GPU, by workload type**



VENDOR CLAIM Ratios published by Red Hat and attributed to Intel, whose commercial interest in a CPU-heavy future is direct. Treat the direction as credible and the endpoints as advocacy. No independent measurement of deployed ratios was located. [7]

VENDOR CLAIM Silicon Motion is positioning enterprise SSDs as a persistent memory tier for KV cache offload, with its MonTitan reference design kit and PerformaShape QoS engine aimed at what it describes as always-on agentic workloads. Its SM8466 PCIe Gen6 controller is specified at up to 28 GB/s sequential and 7 million random IOPS. This is a component vendor describing a market it would like to exist, and no deployment data accompanies it. [8]

ASSESSMENT Strip the vendor framing and both point at the same structural change. In single-shot inference the accelerator is the cost. In an agentic loop the model is one component of a longer sequence dominated by tool calls, state handling and context replay, so cost migrates into orchestration, memory bandwidth and storage. The KV cache is the specific mechanism: context that must persist across loop iterations either stays in expensive HBM, gets recomputed at full token cost, or moves down a tier. [7] [8]

ASSESSMENT The budgeting consequence is more awkward than the engineering one. GPU capacity is a named, tracked, board-visible line. CPU cores, memory and enterprise SSD are absorbed into general infrastructure and refreshed on a different cycle by a different team. A cost shift from the first category into the second is close to invisible in most technology business management practice, which means agentic infrastructure cost will be systematically under-attributed to AI programmes for at least a budget cycle. [7] [8]

Pricing model fragmentation

FACT CodeRabbit prices its Security Agent at 40 dollars per seat per month, with full-codebase scans metered separately and volume options available. A per-seat headline with a metered exception attached is the transitional form: the vendor keeps the predictable unit for the buyer and carves out the workload it cannot cost. [13]

VENDOR CLAIM ChurnZero sells Agentic Essentials as a single annual subscription with a flat fee and a set allotment of credits, with every agent drawing from one shared pool across roughly twenty agents. The company's stated rationale is that usage-based pricing across the industry has made costs hard to predict. That is a vendor naming the variance problem in a press release and selling the cap as the product. [14] [15]

THIRD-PARTY ESTIMATE Salesforce is reported to run three Agentforce pricing models concurrently: two dollars per conversation, Flex Credits at 500 dollars per 100,000 credits with a standard action at 20 credits and a voice action at 30, and per-user licensing from 125 dollars per user per month. These figures come from third-party pricing guides and commentary rather than a retrievable vendor rate card, so treat the structure as well attested and the specific numbers as indicative. [21] [12]

Vendor and product	Billing unit	Stated price	What the buyer cannot predict
CodeRabbit Security Agent	Per seat, with a metered carve-out	40 USD per seat per month, full-codebase scans metered separately	The scan line. Seat cost is fixed; the exception is where the variance was moved, not removed.
ChurnZero Agentic Essentials	Shared credit pool	Flat annual fee with a set credit allotment across roughly 20 agents	What a credit buys in practice. A pool shared across agents means one badly designed agent can starve the rest.
Salesforce Agentforce	Three units in parallel	2 USD per conversation, or 500 USD per 100,000 credits, or from 125 USD per user per month	Which unit is cheapest for a given workload, which is the calculation the vendor has effectively handed to the customer.
Frontier model APIs, bought direct	Per token, now time-tiered	DeepSeek V4-Pro output at 1.98 off-peak and 3.96 at peak per million	Tokens per completed task, which no provider quotes and no buyer can forecast before building the agent.

ASSESSMENT Read across that table and the pattern is that every vendor has placed the variance somewhere, and no vendor has removed it. Per-seat pricing puts it on the vendor's gross margin. Credit pools put it on the customer's capacity planning. Per-transaction pricing puts it on the customer's monthly bill. Nobody has yet built an agent product whose cost per completed task is predictable enough that the billing unit is obvious, which is the actual state of the market beneath the pricing debate. [13] [14] [21]

FACT Meanwhile the settlement rails are being built ahead of the demand. Cloudflare launched cloudflare.pay handles and programmable virtual wallets on 4 August 2026, letting an account create wallets for its agents to buy APIs, MCP tools and content, on top of its x402 Monetization Gateway. On the same quarter's earnings call, CEO Matthew Prince said more than half of traffic across Cloudflare's network was non-human for the first time, with daily AI agent requests up 1,700 percent between 1 June 2025 and 31 May 2026. [16] [22] [24]

ASSESSMENT Non-human traffic crossing half is a real threshold and it is being cited loosely. Crawlers, scrapers, health checks and API integrations have been most of the request volume on many networks for years; what changed is the composition and the growth rate, and 1,700 percent on daily agent requests is the figure that carries information. The infrastructure conclusion stands regardless: per-transaction settlement for machine buyers is being built now, which means the metering primitives for agent-unit pricing will exist before most software vendors have decided whether to use them. [16] [22]

Repricing scenarios: if agents replace seats as the billing unit

Four scenarios for how enterprise software gets repriced over the next eight quarters. Probabilities are subjective and are the least reliable content in this report. The earliest visible signs are the most useful, because each can be observed without private information.

SCENARIO A Hybrid floor plus meter 45 percent

A platform fee sets a revenue floor at roughly seat-economics level, with agent consumption metered above it and customer-settable caps as a named feature. The dominant form because it preserves the vendor's forecastable base while moving the tail to the customer.

Consequence. Revenue recognition splits. The platform fee stays ratable as a stand-ready obligation; the metered portion is recognised as consumed. A growing share of revenue becomes non-ratable, so remaining performance obligations cover less of forward revenue and the historical RPO-to-revenue relationship stops being a reliable forward indicator. Gross margin stabilises in the low seventies rather than the high seventies, because inference cost of revenue now scales with a revenue line instead of sitting inside a fixed seat price.

Earliest visible sign. A top-ten application vendor discloses metered agent revenue as a separate line, or restates its RPO commentary to distinguish committed from consumption revenue.

SCENARIO B Seat pricing survives, autonomy is throttled to fit 25 percent

Vendors hold per-seat pricing and cap agent behaviour at whatever depth the seat price supports. Marketed as governance and predictability. Genuinely popular with procurement and quietly limiting on capability.

Consequence. Revenue recognition is unchanged and ratable, RPO stays intact, and reported metrics keep working. Gross margin is protected but the product ships deliberately less autonomous than the demo, and the vendor carries the risk that a consumption-priced competitor ships depth it cannot match. This is the comfortable scenario for public SaaS and the dangerous one for its five-year position.

Earliest visible sign. Product pages begin advertising monthly action or run limits per seat rather than capability, and tiering language shifts from what the agent can do to how often it may do it.

SCENARIO C Outcome pricing 20 percent

Vendors charge per resolved ticket, per reconciled invoice, per closed candidate, and absorb token variance themselves. Attractive to buyers because it converts an unbounded cost into a unit price tied to value delivered.

Consequence. The hardest accounting treatment of the four. Consideration is variable and must be constrained under ASC 606 until a significant revenue reversal is not probable, which pushes recognition later and makes it lumpier. Gross margin becomes volatile in both directions, because the vendor now owns the entire right tail of the cost distribution including retries on tasks that never resolve. Expect at least one vendor to discover its outcome definition was too generous.

Earliest visible sign. A vendor takes a revenue true-up or reverses previously recognised variable consideration, or quietly redefines what counts as a resolved outcome.

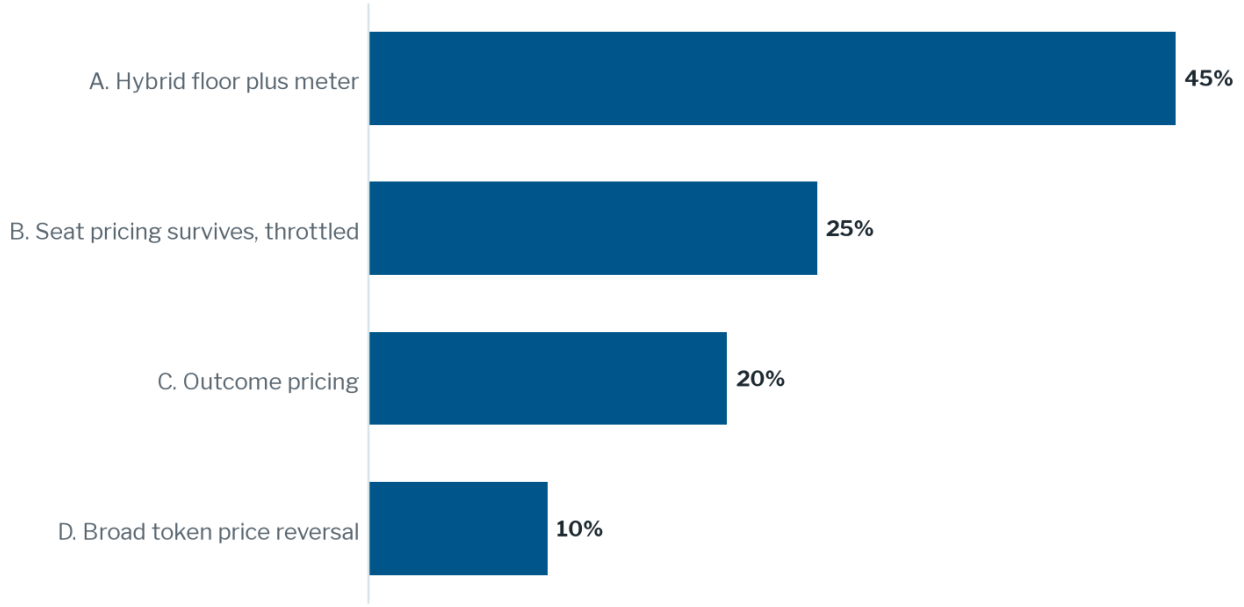
SCENARIO D Broad token price reversal forces repricing everywhere 10 percent

Sustained inference price increases across two or more providers with meaningful share, following DeepSeek. Not a spike but a durable reset of the floor as capacity binds.

Consequence. Mid-contract economics break. Multi-year fixed-price agreements signed against an assumption of falling input costs become loss-making, and renegotiation is a contract modification whose treatment depends on whether added scope is priced at standalone value. Gross margin compresses sharply and immediately for any vendor that sold three-year fixed pricing on agent-heavy products in 2025 and 2026. The buyer-side consequence is a wave of price-increase notices framed as model cost pass-through.

Earliest visible sign. A second provider with meaningful share publishes peak and off-peak rates, or a public SaaS vendor names model input cost as a factor in a guidance revision.

FIGURE 7 **Scenario probabilities to mid-2028**



SCENARIO Subjective probabilities from the author, summing to 100. These are the least reliable content in this report and are included because refusing to quantify a judgment is not the same as being careful. The tripwires beneath each scenario are the part worth monitoring.

What the billing unit does to the accounts

The revenue recognition consequences are not a footnote. Changing the billing unit changes which ASC 606 mechanics apply, and that changes the metrics investors have spent a decade learning to read.

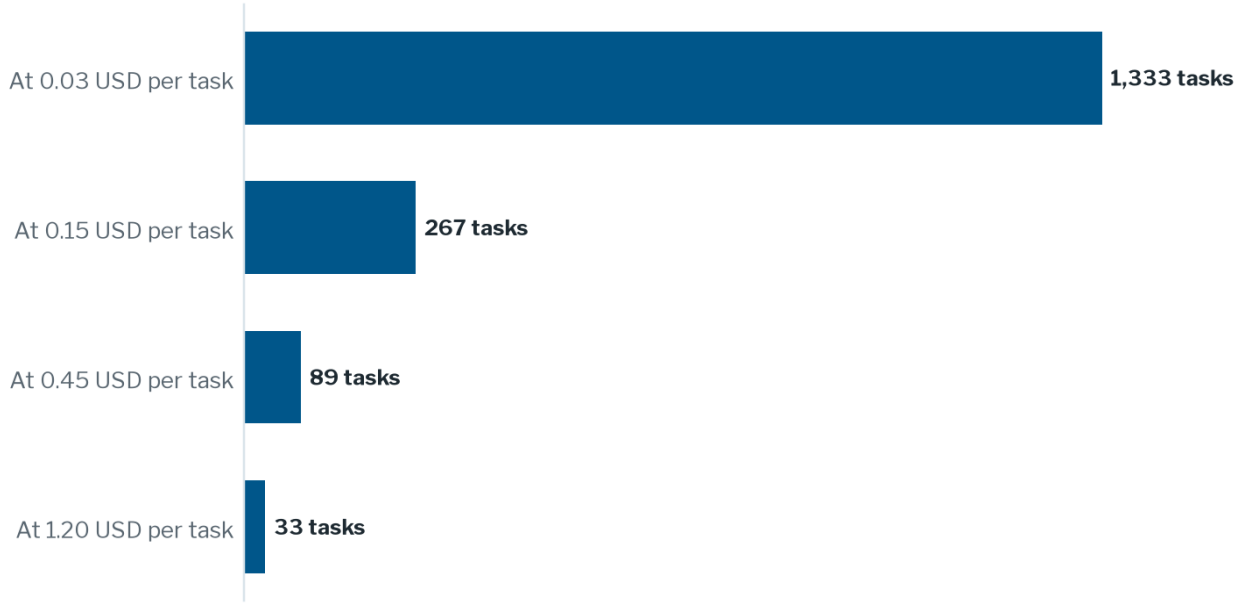
Billing unit	Recognition treatment	Effect on deferred revenue and RPO	Gross margin behaviour
Per seat	Ratable over the contract term as a stand-ready obligation. Fully predictable.	Deferred revenue and RPO behave as they always have and cover most of forward revenue.	Vendor absorbs all inference cost variance. Margin degrades quietly as agent depth increases, with no offsetting revenue line.
Credit pool	Typically ratable if credits expire with the term, otherwise recognised on redemption with breakage estimated.	Largely preserved, which is why finance teams like it. Breakage estimation is the audit exposure.	Variance is transferred to the customer's planning problem rather than eliminated. Margin is protected until customers learn to size pools accurately.
Per transaction	Recognised as consumption occurs. Variable consideration must be estimated and constrained where it is billed in arrears.	A growing share of revenue sits outside RPO. Forward visibility falls and quarterly revenue tracks customer usage seasonality.	Margin follows the markup on inference and is structurally lower but far more stable, because cost and revenue move together.
Per outcome	Variable consideration with a constraint. Recognised only	Weakest visibility of the four. Estimates carry genuine	Most volatile. The vendor owns the full cost distribution

Billing unit	Recognition treatment	Effect on deferred revenue and RPO	Gross margin behaviour
	when the outcome is reasonably assured, which is later and lumpier.	reversal risk and invite auditor scrutiny.	including unresolved retries, so a change in task mix moves margin without any change in price.

ASSESSMENT The metric consequence deserves stating plainly for anyone modelling these companies. A SaaS business moving from seats to consumption does not merely change its growth profile, it degrades the predictive power of the disclosures analysts use. Remaining performance obligations stop bounding forward revenue. Net revenue retention starts blending price, seat expansion and raw usage into one number that no longer decomposes. Anyone reading a 2028 SaaS model with 2023 metric intuitions will be wrong in a direction they cannot see. [23]

ASSESSMENT The gross margin arithmetic is less dramatic than the framing suggests, and it is worth doing rather than asserting. Take a seat-priced agent product at 40 dollars per seat per month and ask how many completed tasks that seat supports before inference cost alone exhausts the entire subscription. The answer is the vendor's real capacity headroom, and it collapses fast as tasks get deeper. [13]

FIGURE 8 Tasks per seat per month before inference cost alone consumes a 40 dollar subscription



ASSESSMENT The author's arithmetic on a 40 USD per seat per month price point, taken from CodeRabbit's published Security Agent pricing. Deliberately generous to the vendor: it assumes the entire subscription is available to cover inference and ignores all other cost of revenue, sales, and support. Real headroom is a fraction of these figures. The per-task cost points span the sourced range in this report. [13] [10]

ASSESSMENT Thirty-three tasks per seat per month is roughly one and a half per working day, and it is the ceiling under assumptions that ignore every other cost. A code review agent, a support triage agent or a reconciliation agent doing useful work runs well past that. This is why

the hybrid in Scenario A is the most likely outcome and not merely the most convenient one: at realistic task depths there is no seat price a buyer will accept that also covers the workload, and no vendor can hold a fixed price against a cost it cannot bound. [13]

Leading indicators

Each of these is observable from public information. They are ordered by how much they would move the assessment, not by how likely they are.

If this happens	What it means	Moves toward	Where to watch
A second provider with meaningful share publishes peak and off-peak inference rates	Capacity, not commercial strategy, is setting the price floor. DeepSeek was a signal rather than an outlier.	Scenario D	Provider pricing pages and changelogs. The change usually appears on the rate card before it appears in press coverage.
A public SaaS vendor discloses metered agent revenue as a separate reported line	The hybrid has become material enough to require segmentation, and the seat era is formally over in reporting terms.	Scenario A	Quarterly supplementals and 10-Q revenue disaggregation, not the earnings release headline.
A vendor reverses previously recognised variable consideration on an outcome-priced agent product	Outcome pricing was mispriced against the real cost distribution, which is the failure mode this report expects.	Away from Scenario C	Restatements, critical audit matters, and any change in the revenue recognition policy note.
Product pages start advertising monthly action limits per seat rather than capability	Vendors are protecting margin by capping autonomy, and the seat price is being defended rather than replaced.	Scenario B	Pricing and packaging pages, and the tier comparison table specifically.
Published per-task token counts for a named agentic workload fall materially two quarters running	The efficiency assumption in this report is wrong and the two curves are converging.	Refutes the call	Benchmark preprints with disclosed token counts. Vendor efficiency claims do not count.
Named enterprise buyers publish hard per-user AI spend ceilings as standard policy	The variance argument is correct and the market has settled on the contractual fix rather than a technical one.	Confirms the call	Engineering blogs and CIO interviews. Uber's disclosure is the template.
Enterprise CPU and SSD refresh cycles get pulled forward and attributed to AI programmes	Infrastructure reweighting is real and is being budgeted rather than absorbed.	Supports the infrastructure read	Server OEM and component vendor guidance, and technology business management cost model changes.

Implications

For CFOs, FP&A and technology business management

ASSESSMENT Stop modelling AI spend as a unit price multiplied by a volume forecast. The unit price is falling, the volume is not forecastable, and the product of the two has no defensible central estimate. Model it instead as a capacity commitment with a ceiling, the way a variable cost with a fat right tail is normally handled. The immediate action is to require every agent workload to carry a spend ceiling before it reaches production, and to make the ceiling a budget-holder decision rather than a platform default. Uber's outcome, more users at lower total cost, is the reference case for what ceilings do. [11]

ASSESSMENT Second, insist that AI cost reporting attribute CPU, memory and storage growth to AI programmes rather than to general infrastructure refresh. The reweighting described earlier will otherwise flatter your AI unit economics for a full budget cycle while the cost lands somewhere else in your own P&L. Any cost model that maps only GPU and API spend to AI will understate the true figure, and the gap widens as agent depth increases. [7] [8]

For CIOs, platform engineering and FinOps

ASSESSMENT The FinOps primitive that matters here is not showback, it is the enforced ceiling with a defined failure behaviour. Decide now what an agent does when it hits its budget: fail closed, degrade to a cheaper model, or queue for off-peak. That decision belongs in the platform, not in each agent, and it is far cheaper to make before a hundred agents exist than after. The DeepSeek tiering makes the third option real for the first time, and batch-eligible work moved off peak is roughly half price. [2] [3]

ASSESSMENT Second, instrument cost per completed task, including failed attempts, as a first-class metric alongside latency and success rate. Almost nobody does this, and it is the only number that makes agent designs comparable. A design that succeeds 95 percent of the time at three retries is worse economics than one that succeeds 88 percent at one retry, and no dashboard currently in wide use will tell you that. [17]

For SaaS pricing, packaging and product strategy

ASSESSMENT Pick the hybrid and pick it deliberately rather than arriving at it after two failed repricings. A platform fee that preserves a forecastable floor, metered consumption above it, and a customer-settable cap marketed as a feature rather than a limitation. Salesforce running three units concurrently is not sophistication, it is the visible cost of not choosing, and every month spent there teaches customers to arbitrage between your own price lists. [21] [12]

ASSESSMENT Resist outcome pricing until your cost distribution is instrumented at the ninety-ninth percentile rather than the mean. It is the most attractive model to buyers and the one most likely to produce a revenue reversal, because the tasks that never resolve consume tokens and generate no recognisable revenue. If you sell it anyway, define the outcome narrowly enough that an unresolvable request is excluded by the contract rather than by a later argument. [23]

For equity analysts and investors

ASSESSMENT Two specific corrections. First, do not read Cloudflare's move to a negative 29.6 percent GAAP operating margin as the cost of the agentic build-out; a 150.7 million dollar restructuring charge is roughly 21.6 points of it, and margin improved excluding it. The agentic thesis at Cloudflare is visible in product and traffic, not in that margin line. Second, expect RPO to lose its predictive relationship with forward revenue at any vendor moving to metered agent pricing, and expect net revenue retention to become a blended number that no longer decomposes into price and seats. [1] [23]

ASSESSMENT On the semiconductor and infrastructure side, the CPU-to-GPU reweighting is directionally credible and quantitatively unproven, and the only sources for it are companies that sell CPUs and storage controllers. The useful discipline is to wait for it to appear in server OEM mix rather than in component vendor positioning. The KV cache offload thesis in particular is a component vendor describing a market it wants, with no deployment data attached. [7] [8]

For IT procurement and vendor management

ASSESSMENT Three clauses are worth more than any discount you will negotiate this year. A hard spend ceiling with a defined behaviour at the ceiling. A definition of what constitutes one billable task that explicitly addresses retries and failures, because the vendor's default definition will not. And protection against mid-term input cost pass-through, which Scenario D makes a live risk rather than a theoretical one for anyone signing multi-year agreements on agent-heavy products. [2]

ASSESSMENT Where a vendor sells a shared credit pool, ask what a credit buys in tokens and whether that conversion can change during the term. A pool is a cap whose denominator the vendor controls, and a quiet repricing of the conversion rate is a price increase that does not appear as one. Ask the same question of any per-action pricing where the action is defined by the vendor. [14] [15]

What this document cannot answer

The following need primary research, and this analysis is weaker for their absence. Each is stated as the question that would actually have to be answered.

What is the distribution, not the average, of cost per completed task for a defined agentic workload in production? Every published figure is a mean or an anecdote. The argument in this report turns on the ninety-ninth percentile, and no dataset containing it was located. Answering it requires instrumented telemetry from operators willing to disclose, which currently nobody is.

Do organisations that reach autonomous multi-step workflows differ from those that stall primarily in spend controls or primarily in data readiness? This is the direct test of this report's central hypothesis and it requires a survey that asks both questions of the same respondents, segmented by deployment stage. No such instrument was found.

What share of enterprise agent token consumption is retries and failed attempts? This is the single number that would settle whether the variance problem is a design problem or a structural one. It is knowable from any operator's own logs and is published by nobody.

Is DeepSeek's tiering a capacity signal or a margin decision? Answering it needs utilisation data no provider discloses. The leading indicator table gives the observable proxy, which is whether anyone follows.

What are actual deployed CPU-to-GPU ratios in production agentic clusters? The only figures available come from Intel via Red Hat. An independent survey of deployed ratios would either confirm a real infrastructure reweighting or expose it as component vendor positioning, and the difference is worth a great deal to anyone with a semiconductor position.

Half-life of this assessment

Decaying within six months: every price in this document. The DeepSeek figures, the CodeRabbit seat price, the Salesforce credit rates and the break-even arithmetic built on them are all quotations from a market that reprices quarterly. The scenario probabilities are also short-lived, since a single second provider adopting time-of-day pricing would move Scenario D materially. Treat the numbers as of 19 August 2026 and re-source them before using any of them in a negotiation.

Decaying within twelve months: the adoption figures and the barrier rankings. ServiceNow's 59 and 9 will both move, and the gap between them is the thing to re-measure rather than either number alone. The token multiplier range will also tighten as benchmark work accumulates, and this report expects it to tighten downward as agent designs mature, which would weaken the central argument.

Decaying within twenty-four months: the pricing model fragmentation. Markets do not sustain three incompatible billing units inside one vendor for long. By late 2028 either the hybrid has consolidated or something in this report's scenario set was wrong, and either way this section will read as a snapshot of an unsettled moment.

Architectural, and unlikely to decay: that an iterative reasoning loop consumes more tokens per completed unit of work than a single retrieval call, and by a factor rather than a margin. That the cost of a task with retries has a long right tail. That a variable cost with an unbounded tail cannot be underwritten without a cap. Those follow from the shape of the computation rather than from any current price, and they will still be true when every number in this document is stale.

Limitations

This is independent work. No vendor named here commissioned, reviewed, funded, or was given advance sight of it. The author holds no position in any company named and has no commercial relationship with any of them.

The evidence base has specific weaknesses a reader should weigh. The adoption and barrier data comes from a single vendor-commissioned survey with a self-selected sample, run by a company that sells into both sides of the gap it measures. It is used because it is the largest instrument available with published figures, not because it is disinterested.

The infrastructure reweighting rests entirely on vendor sources: a Red Hat post relaying an Intel projection, and a Silicon Motion product announcement. Both companies profit if the projection is correct. No independent measurement of deployed CPU-to-GPU ratios in agentic clusters was located, and the direction is treated as credible while the magnitudes are not.

The token multiplier evidence is the weakest part of this report and the part its argument most depends on. The two figures with disclosed methodology come from a preprint that has not been peer reviewed. The remainder are illustrations from commercially interested sources. The argument is built to survive this by resting on the direction and on the EY cost-per-interaction pair rather than on any single multiplier, but a reader who rejects the multiplier evidence entirely should discount the first analysis section accordingly.

The break-even chart and the restructuring adjustment to Cloudflare's operating margin are the author's own arithmetic on published inputs, not reported figures, and are labelled as assessment rather than fact. The scenario probabilities are subjective and are the least reliable content here.

Three of the eight anchor claims that prompted this research failed verification, which raises the base rate of contamination for anything in this category sourced from secondary coverage. This report may itself contain a laundered figure that survived checking. The source list below is typed by provenance so a reader can see which claims rest on what.

Sources

[1] Cloudflare, Inc. "Cloudflare Announces Second Quarter 2026 Financial Results." 6 August 2026. *Vendor primary* cloudflare.com

[2] Fortune. "DeepSeek increases prices for AI services by multiple times." 13 August 2026. *Independent press* fortune.com

[3] DeepSeek. "API Pricing." Accessed 19 August 2026. *Vendor primary* deepseek.ai

[4] Gartner. "Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027." 25 June 2025. *Analyst firm* gartner.com

[5] Gartner. "Gartner Predicts 40% of Enterprise Apps Will Feature Task-Specific AI Agents by 2026, Up from Less Than 5% in 2025." 26 August 2025. *Analyst firm* gartner.com

[6] ServiceNow. "Enterprise AI Maturity Index 2026." Survey of 4,500 senior leaders across 19 countries and 12 industries. *Vendor primary, self-selected sample* servicenow.com

[7] Bowerman, S. and Ableidinger, G. "The CPU is back: Rethinking the CPU-GPU split for LLM inference." Red Hat, 6 August 2026. *Vendor primary, relaying an Intel projection* redhat.com

[8] Silicon Motion Technology Corporation. "Silicon Motion Unveils MonTitan SSD Reference Design Kit for AI Infrastructure at FMS 2026." August 2026. *Vendor primary* siliconmotion.com

[9] Epoch AI. "LLM inference prices have fallen rapidly but unequally across tasks." 12 March 2025. *Independent research* epoch.ai

[10] Smith, S., Wanner, S., Butler, W. and Diasio, D. "Agentic AI token costs." EY, 1 June 2026. *Professional services firm, practice observation* ey.com

[11] Fortune. "After blowing through its entire 2026 AI budget in months, Uber CTO says the tokenmaxxing era is ending." 7 August 2026. *Independent press* fortune.com

[12] Cohan, P. "As Token Costs Plunge, Enterprise AI Providers Face A New Margin Squeeze." Forbes, 28 July 2026. *Independent press, opinion contributor* forbes.com

[13] InfoWorld. "CodeRabbit targets AI-generated code overload with Agentic Change Management." August 2026. *Independent press* infoworld.com

[14] ChurnZero. "ChurnZero launches Agentic Essentials, the only customer success AI that delivers execution, intelligence and reach in one system." 2026. *Vendor primary* churnzero.com

[15] ChurnZero. "ChurnZero launches Retrospective, an AI churn analysis agent, expanding its Agentic Essentials package." PR Newswire, 12 August 2026. *Vendor primary* prnewswire.com

[16] Cloudflare. "Announcing Cloudflare Wallets: The programmable wallet for the agentic Internet." 4 August 2026. *Vendor primary* blog.cloudflare.com

[17] "AgenticRAG: Agentic Retrieval for Enterprise Knowledge Bases." arXiv preprint 2605.05538. *Preprint, not peer reviewed* arxiv.org

[18] Spheron. "Agentic AI Inference Cost: Why Agents Burn 5-30x Tokens." 2026. *Third-party commentary, commercial interest* spheron.network

[19] Dagade, G. "Your AI Costs Will 3x This Year. Here's How to Survive It." Preto.ai, 11 April 2026. *Second-hand, illustrative and unverified* preto.ai

[20] Singh, B. M. "Why Cheaper AI Tokens Are Increasing Enterprise AI Costs." The Modern Data Company, 8 July 2026. *Vendor-published commentary, commercial interest* themoderndatacompany.com

[21] SaaStr. "Salesforce now has 3 pricing models for Agentforce." 2026. *Third-party commentary, commercial interest possible* saastr.com

[22] Investing.com. "Cloudflare Q2 2026 slides: agentic internet drives 36% revenue growth." 6 August 2026. *Independent press, reporting company slides* investing.com

[23] Ordway Labs. "ASC 606 for Usage-Based Pricing: Rules, Examples and How to Comply." *Practitioner reference, commercial interest* ordwaylabs.com

[24] Converge Digest. "Cloudflare: 1,700% Increase in Agentic AI Traffic." August 2026. *Independent press* convergedigest.com

Rights and permitted use