David Soden

MARKET ASSESSMENT | AUGUST 2026

Nvidia Says 450x. Your Invoice Says 3x.

AgentX, Vera Rubin NVL72, Groq 3 LPX and Crescent Island read against the consumption contracts buyers are actually signing

David Soden

Independent analyst and practitioner, enterprise automation and applied AI

davidsoden.com/market-assessment

REPORT DS-2026-E67B · 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 report follows one claim from the silicon to the invoice. The claim is that a new generation of inference hardware makes agentic workloads dramatically cheaper to serve. The question is whether any of that reaches the organisation paying the bill, and if not, where it stops.

Four hardware positions are in scope: the SemiAnalysis AgentX benchmark and its methodology, Nvidia's Vera Rubin NVL72 throughput-per-megawatt claims, the Nvidia Groq 3 LPX token-generation claims, and Intel's tokens-per-watt framing for Crescent Island. Against those sit the commercial structures buyers sign: published per-million-token API pricing from Anthropic and OpenAI, published per-event agent pricing from Microsoft, Salesforce and Intercom, and the contract risk categories named by Info-Tech Research Group. Training economics, model quality, chip supply and equity valuation are all out of scope.

Three things about this category's published numbers need discounting before any of them are used. First, every headline efficiency figure in it is a vendor figure, measured by the vendor, on a configuration the vendor selected, and in two of the four cases the vendor does not disclose the batch size or concurrency the number was produced at. Second, the multiples are quoted against different baselines using different benchmarks, so they cannot be compared to each other without saying so out loud. Third, the unit of account itself is unstable: the token is defined by a tokenizer the vendor controls and has changed size within the period this report covers. Where a number cannot be checked, it is labeled as such rather than smoothed over.

Everything below carries one of six labels. The label is the first thing to read, because it says how much weight the sentence can take.

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 the conclusion that fails if it turns out to be wrong.

1. Published list prices approximate what buyers pay. Large buyers negotiate discounts that are not disclosed, and Anthropic, OpenAI and Microsoft all say so on their own pricing pages. If enterprise discounting has moved substantially faster than list pricing, the pass-through gap measured here is overstated and the second finding weakens. It does not disappear, because the vendor efficiency multiples are also stated at list.

2. Nvidia's two generational efficiency claims are measurable against each other. The 15x Hopper comparison and the 30x GB300 comparison are quoted in the same blog post but were not necessarily produced on the same workload, and the older figure predates AgentX. If they are not compatible, the compounded figure in the title is wrong and should be read as directional rather than arithmetic. The direction is not in dispute; only its size.

3. Agentic token consumption keeps growing faster than per-token prices fall. This is Gartner's position and it matches the observed shape of agent workloads. If consumption flattens, the deflation eventually shows up on the bill and the pass-through problem becomes a timing problem rather than a structural one.

4. The platform layer between the model and the buyer is where the margin accrues. This is inferred from published per-event pricing, not from anyone's disclosed margins, which are not broken out at this level anywhere in the public record. If platform gross margin on agent events turns out to be thin, the retention is happening at the model layer instead and the recommendations change target but not substance.

5. The disclosed billing-error rate is roughly representative. One audit firm's figures across sixty companies are the only quantified evidence in the public record, and that firm sells the audit. If the real rate is materially lower, the audit-rights argument rests on principle rather than on recovery, which is a weaker but still sufficient case.

The call

ASSESSMENT The efficiency gain in agentic inference silicon is real, and almost none of it reaches the buyer. Nvidia's own generation claims compound to two orders of magnitude more throughput per megawatt since Hopper, while the flagship tier of published token pricing has moved a fraction of that and its top end has risen. The gain is retained where the meter sits, at the platform layer, and agent workloads consume enough extra tokens to absorb what does get through. Do not sign a consumption contract without a defined billable unit, a hard cap and an audit clause. I would move if a major platform published per-event metering a buyer could reconcile against provider invoices.

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 four questions in the brief resolve into one finding with three supports. The finding is that the pass-through is small and the reason is structural rather than temporary.

FINDING 1 AgentX measures the thing that costs money, and no two vendor claims were run on it the same way.

> FACT AgentX is the first open, multi-turn agentic benchmark at long context, released under Apache 2.0 and built from 393 anonymised Claude Code traces with prefix reuse preserved. Its input distribution reaches a 99th percentile of 675,000 tokens against an 8.6k output p99. Fixed-sequence benchmarks in the 8k/1k class cannot see prefix reuse, KV cache lifecycle or sub-agent bursts, which is where agentic cost actually lives. Nvidia measured Vera Rubin NVL72 on AgentX and says the result is pending SemiAnalysis review. It measured Groq 3 LPX on a different benchmark, a different model and a tenth of the context. [1] [2] [3] [4]

FINDING 2 Nvidia's own multiples compound to roughly 450x. The flagship price sheet moved 3x.

> ASSESSMENT Nvidia states GB300 NVL72 delivers up to 15x better throughput per megawatt than Hopper and Vera Rubin NVL72 up to 30x better than GB300. Multiplied, that is about 450x since Hopper, a figure Nvidia does not itself publish. Over a comparable window

Anthropic's flagship went from $15 and $75 per million input and output tokens on Opus 4.1 to $5 and $25 on Opus 5, a clean 3x. The top of OpenAI's published sheet moved the other way: GPT-5.5 Pro lists at $30 and $180. [3] [9] [10]

FINDING 3 The token is not a fixed unit, and the vendor controls its size.

FACT Anthropic's own pricing documentation states that Claude 4.7 and later models use a newer tokenizer that produces approximately 30 percent more tokens for the same text. A 30 percent per-token price cut across that boundary is exactly cancelled at the invoice. No audit clause written against a token count can survive a change in what a token is unless it also fixes the tokenizer or measures in characters. [9]

FINDING 4 Copilot Studio's reasoning meter prices out near $100 per million tokens.

ASSESSMENT Microsoft's published rate card bills premium text and generative AI tools at 10 Copilot Credits per 1,000 tokens, and pay-as-you-go credits list at one cent each. That is $100 per million tokens, sitting on top of a separate per-event feature rate that is charged as well. Frontier model list pricing at the same moment runs $20 to $25 per million output tokens. The gap between those two numbers is where the hardware gain is being retained. [10] [11] [12]

FINDING 5 One prompt now generates a dozen distinct billable events.

FACT Declaring a browser toolset adds about 6,600 input tokens to every request, not once per session. Anthropic bills managed agent sessions at $0.08 per session-hour on top of tokens, web search at $10 per 1,000 searches, and cache writes at 1.25x or 2x base input. Microsoft bills 1 credit for a classic answer, 2 for a generative answer, 5 for an agent action and 10 for tenant graph grounding. Salesforce bills $0.10 per action. Intercom bills $0.99 per outcome on top of seats. None of these existed in a per-seat contract. [9] [11] [13] [14]

FINDING 6 Info-Tech names limited audit rights as a category, and no major vendor publishes a reconcilable meter.

FACT Info-Tech Research Group names five risk categories in agentic AI contracts: unclear billing definitions, hidden cost drivers, weak financial controls, vendor-controlled pricing changes, and lock-in and accountability gaps. Its principal research director says clear definitions, consumption caps, audit rights, pricing-change protections and dispute mechanisms must be negotiated before signing. On the vendor side, Anthropic's marketplace billing converts token usage into Claude Consumption Units and the buyer's AWS bill shows a single line item. [9] [15]

FINDING 7 The only quantified error rate in the public record is about five percent of reviewed spend.

THIRD-PARTY ESTIMATE An audit firm reports reviewing $34 million of AI spend across 60 companies since March 2026 and identifying close to $1.7 million in mistaken overcharges, roughly 80 percent of which was credited back. Its named error patterns include failed requests still charged, duplicate charges from agent retry storms, and billing during provider outages. The firm sells the audit tool, so treat the rate as an upper-bound marketing figure rather than a measurement. The error patterns are more useful than the percentage. [19] [20]

FINDING 8 Gartner's position and mine agree on the mechanism and differ on the framing.

> THIRD-PARTY ESTIMATE Gartner predicts inference cost per agentic workflow rises more than fivefold through 2028 while token costs fall 95 percent by 2030, and calls the gap an inference paradox driven by a token-deflation illusion. That is a consumption argument. This report adds a pricing argument on top of it: even holding consumption flat, the published per-token pass-through is a small fraction of the claimed hardware gain, and the per-event meters above the token are where the difference sits. [16] [17]

The benchmark: what AgentX measures, and on whose configuration

What the benchmark does differently

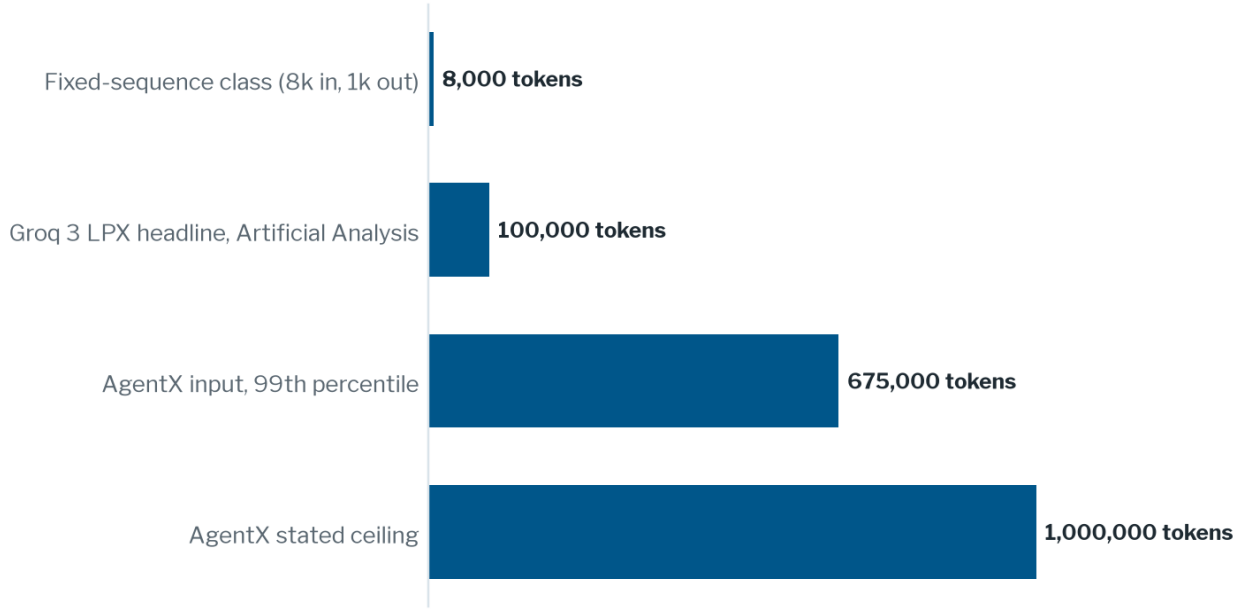
FACT AgentX is published by SemiAnalysis as the first fully open source, multi-turn agentic coding inference benchmark at one million context, released under Apache 2.0 with the code in a public GitHub repository, a public results database served over REST, and GitHub Actions CI provenance. It is built from a dataset of 393 internal Claude Code traces, anonymised with prefix reuse patterns preserved, and SemiAnalysis describes the dataset as costing three million dollars to produce. [1] [2]

FACT The workload captures four properties a fixed-sequence benchmark cannot. Sessions run to tens or hundreds of turns. Context accumulates through system prompts and tool definitions. Prefix reuse is high enough that most context can be served from KV cache, and SemiAnalysis reports 95 percent or better cache hit rates on representative workloads. Sub-agents spawn in bursts, which makes the cache access pattern bursty rather than steady. [1]

ASSESSMENT That last property is the one that matters commercially. A benchmark in the 8k/1k or 1k1k class measures a single-turn request where prefix reuse is irrelevant by construction. It therefore measures kernel throughput. AgentX measures a systems problem: KV cache lifecycle, hybrid-attention cache correctness, CPU KV offload, transfer progress and routing affinity. Cost per token in production is decided by those, not by kernel throughput, which is why a hardware claim measured on a fixed sequence tells a buyer almost nothing about what their agent fleet will cost. [1]

FACT The measurement infrastructure is substantial: roughly two megawatts of continuously operated compute across more than a thousand chips, covering GB300 NVL72, GB200 NVL72, B300, B200 and H200 on the Nvidia side and MI355X, MI325 and MI300X on the AMD side, with Rubin, TPUs and MI455X listed as incoming. Models under test include DeepSeek V4 Pro at 1.6 trillion parameters with 49 billion active, Kimi K3 at 2.8 trillion, MiniMax M3 at 432 billion, Qwen 3.5 at 397 billion and GLM 5.3. [1] [2]

FIGURE 1 **Every headline claim was measured at a different context length**



FACT Input context each figure was produced at, from the published methodology or press release. These are not four measurements of one thing; they are four different measurements. The Groq 3 LPX and fixed-sequence bars also use different models and different hardware from the AgentX bars, so the ordering carries no performance meaning. [1] [4] [5]

The four claims side by side

VENDOR CLAIM Nvidia states that Vera Rubin NVL72 delivers up to 30x higher throughput per megawatt than GB300 NVL72 on agentic workloads, specifically on DeepSeek V4 Pro, and up to 35x lower cost per million tokens than GB300 NVL72. It states GB300 NVL72 itself delivers up to 15x better throughput per megawatt than Hopper. It also states that DSX MaxLPS power management provisions up to 40 percent more GPUs inside the same megawatt budget. The measurement is attributed to the SemiAnalysis AgentX workload, and Nvidia notes the results are currently pending SemiAnalysis review. [3]

ASSESSMENT Pending review is the operative phrase. The benchmark's whole design argument is that agentic performance is a systems problem where the software stack moves weekly, which is why SemiAnalysis runs it continuously rather than at a fixed point. A vendor-run result on an unreviewed configuration against a moving software baseline is a marketing number until the benchmark's maintainer publishes it. SemiAnalysis has said an AgentX update with refreshed Nvidia and AMD results is due within three to four weeks of 24 August 2026. That publication is the single most useful thing a buyer in this category can wait for. [1] [3]

FACT The Groq 3 LPX claims come from a different measurement entirely. Nvidia reports 3,431 output tokens per second on Gemma 4 31B at 100,000 tokens of context, measured by Artificial Analysis, and 3,382 tokens per second at 10,000 tokens of context, plus a median 4,767 tokens per second on SPEED-Bench coding tasks with a P80 of 5,520. The headline comparison is 4x faster responsiveness than the nearest alternative platform, a platform the press release does not name. [4] [5]

FACT Nvidia's own technical documentation confirms these are aggregate throughput figures rather than per-user rates, and does not disclose the batch size or the number of concurrent users behind them. The rack configuration is 256 LP30 accelerators carrying 128 GB of SRAM collectively, sixteen LPUs per 2U tray, 96 chip-to-chip links per chip at 112 Gbps, and 400Gb Ethernet at the front end. Power draw per rack is not published. [5] [6]

ASSESSMENT Aggregate throughput with undisclosed concurrency is not a number a capacity planner can use. Two systems can report identical aggregate tokens per second while one serves eight users at 400 tokens each and the other serves four hundred users at eight tokens each, and only one of those is an agentic platform. The absence of batch size in a claim explicitly framed around interactivity is the disclosure gap that matters most in this release. [5]

FACT Independent reporting has already contested the 4x figure. The comparator is Cerebras at 882 tokens per second, and Cerebras runs the model on one or two accelerators where the Nvidia configuration needs at least 64. Each LPU carries roughly 500 MB of memory against 288 GB on a Rubin GPU, so models split across accelerators over Ethernet. Gemma 4 31B is a dense model that fits in a single rack, which is the favourable case; DeepSeek V3 on the same architecture is reported to require 1,342 accelerators, a little over five racks. The comparison also excludes Cerebras' current CS-4 generation, per-chip efficiency, power and cost. [7]

Claim	Benchmark and who ran it	Model and context	Configuration disclosed	What it does not establish
Vera Rubin NVL72: up to 30x throughput per MW vs GB300	AgentX, run by Nvidia, pending SemiAnalysis review	DeepSeek V4 Pro; context not stated	NVL72 scale-up domain; rack power not published	Cost per token to the buyer, or the result under an independent run
Vera Rubin NVL72: up to 35x lower cost per million tokens vs GB300	AgentX, run by Nvidia, pending review	DeepSeek V4 Pro; context not stated	No TCO inputs, rack price or utilisation assumption published	Anything about price, which is set commercially and not by cost
GB300 NVL72: up to 15x throughput per MW vs Hopper	Not attributed to AgentX in the same post	Not stated	Not stated	Compatibility with the 30x figure, and therefore the compounded number
Groq 3 LPX: 3,431 output tokens per second	Artificial Analysis, endpoint operated by Nvidia	Gemma 4 31B at 100,000 tokens	256 LP30s, 128 GB SRAM total; batch size and concurrency not disclosed	Per-user interactivity, tokens per watt, or behaviour on a large MoE model
Groq 3 LPX: 4x faster than the nearest alternative	Same, comparator identified only in press coverage	Same dense 31B model	Comparator hardware count not normalised	A like-for-like comparison; the comparator uses 1 to 2 chips against at least 64

Claim	Benchmark and who ran it	Model and context	Configuration disclosed	What it does not establish
Intel Crescent Island: optimised for tokens per watt	No benchmark cited	None stated	160 GB to 480 GB LPDDR5X, 350W TDP, air-cooled PCIe	Any tokens-per-watt value at all, because none is published

Intel's framing without a figure

FACT Intel presented Crescent Island at Hot Chips 2026 as an inference part for long-context agentic work: 32 Xe cores with 256 XMX engines, third-generation Xe Matrix Extensions on a 16-deep systolic array, FP4, MXFP4 and FP64 support, 32 MB of unified L2, and 160 GB of LPDDR5X in the standard configuration rising to 480 GB in an ODM variant. Thermal design power is 350W in an air-cooled PCIe form factor, with idle states at 50W or below and roughly 10W at the lowest. Intel frames the value in tokens per watt and cites no tokens-per-watt number. [8]

FACT Memory bandwidth is not officially specified. The only figure in circulation is a reader estimate of roughly 1.3 TB/s derived from LPDDR5X package arithmetic, which is second-hand and unverified. For reference, published bandwidth on an AMD Instinct MI350P runs 3.6 to 4 TB/s and an Nvidia RTX Pro 6000 Blackwell 1.6 to 1.79 TB/s. No conclusion in this report rests on the 1.3 TB/s figure. [8]

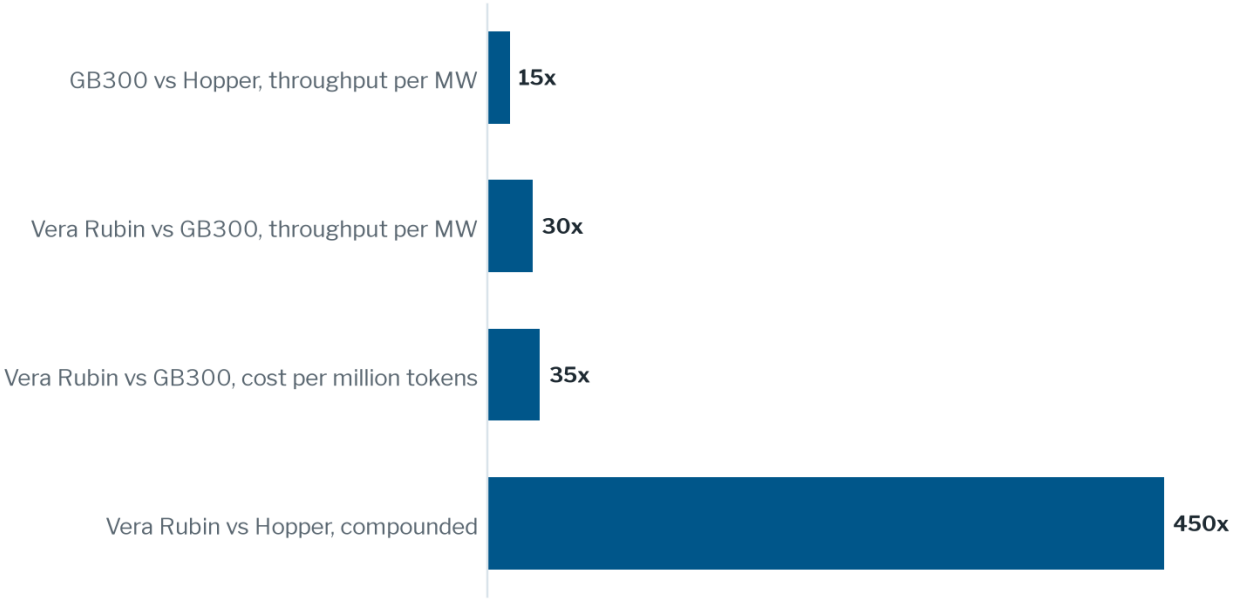
ASSESSMENT Intel is making the correct argument and declining to score itself on it. Tokens per watt is the right buyer-side metric, better than tokens per second and better than throughput per rack, because it is the one that maps to an operating cost a data centre can forecast. Choosing that framing while publishing no value for it, and no bandwidth figure, means the part cannot be compared to anything. The genuinely differentiated claim is capacity: FP8 weights and KV cache on a single card, which is a real answer to long-context agentic serving and needs no benchmark to be interesting. Intel would be better served publishing one honest tokens-per-watt figure on a named model than repeating the framing. [8]

The pass-through arithmetic

What the vendor claims

ASSESSMENT Taking Nvidia's two throughput-per-megawatt claims at face value and multiplying them gives about 450x since Hopper. Nvidia does not publish that compounded figure and the two comparisons may not share a workload, so treat it as directional. Even discounted hard, to say a tenth of it, the claimed improvement is an order of magnitude larger than anything visible in published token pricing over the same period. [3]

FIGURE 2 **Nvidia's stated efficiency multiples, generation over generation**



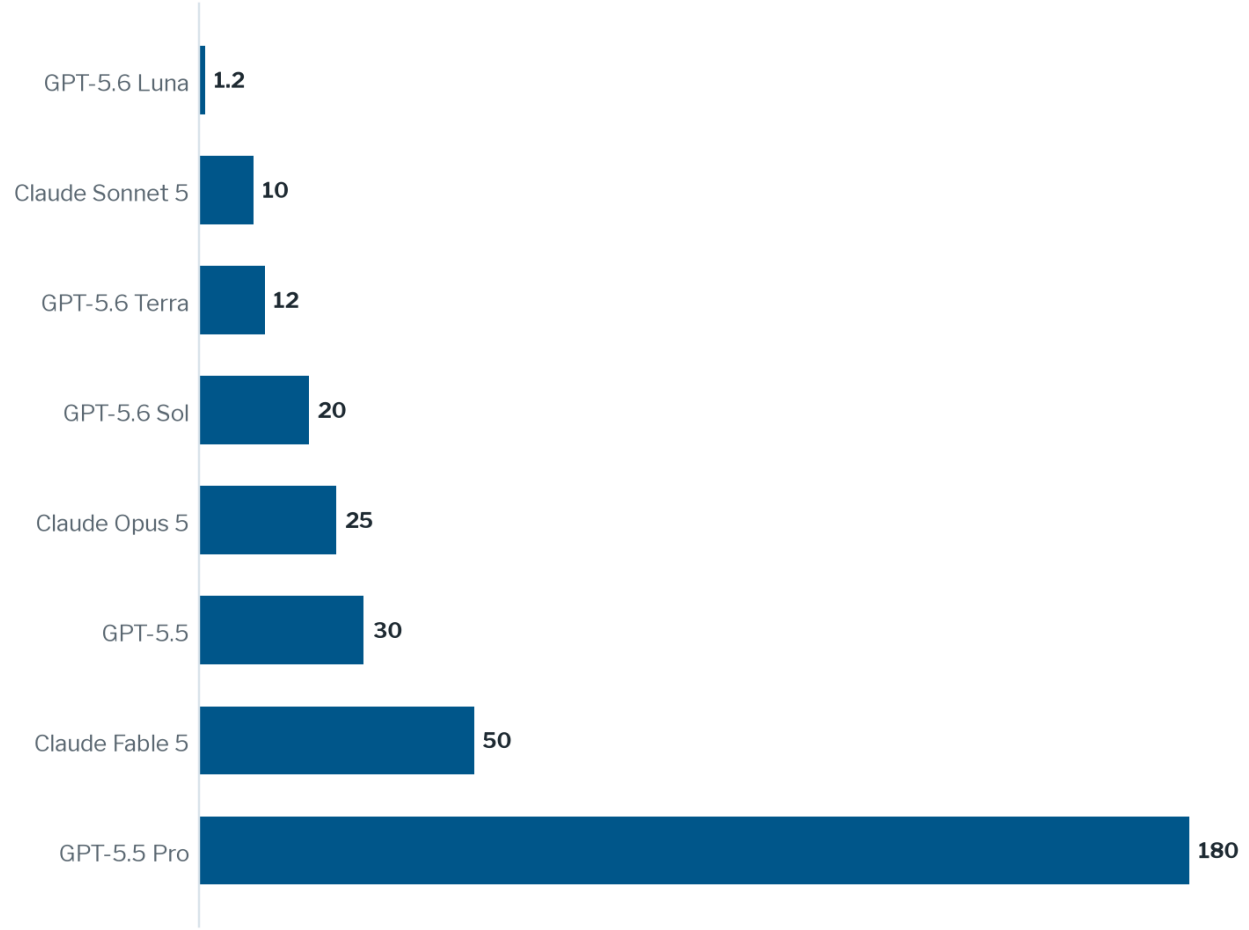
VENDOR CLAIM The first three are Nvidia's own figures, stated as "up to" and measured by Nvidia. The fourth is this report's multiplication of the first two; Nvidia does not publish it and the two inputs are not confirmed to share a workload. None has been independently reproduced. [3]

What the price sheet did

FACT Anthropic's published pricing shows the flagship tier moving from $15 per million input tokens and $75 per million output tokens on Claude Opus 4.1 to $5 and $25 on Claude Opus 5. That is a factor of three on both sides, over roughly the window in which Nvidia claims 450x. Claude Sonnet 5 lists at $2 and $10, and Anthropic has confirmed that the introductory rate is now the standard rate rather than rising to $3 and $15 as originally scheduled. [9]

FACT OpenAI's published sheet tells the same story with a sharper edge. The floor has collapsed: GPT-5.6 Luna lists at $0.20 input and $1.20 output per million tokens. The ceiling has risen: GPT-5.5 Pro lists at $30 and $180, and o1-pro at $150 and $600. The original GPT-4 launched in March 2023 at $30 and $60 and is no longer on the sheet, a figure this report treats as second-hand because it can no longer be verified against OpenAI's own page. [10] [23]

FIGURE 3 **Published output price per million tokens, US dollars, August 2026**



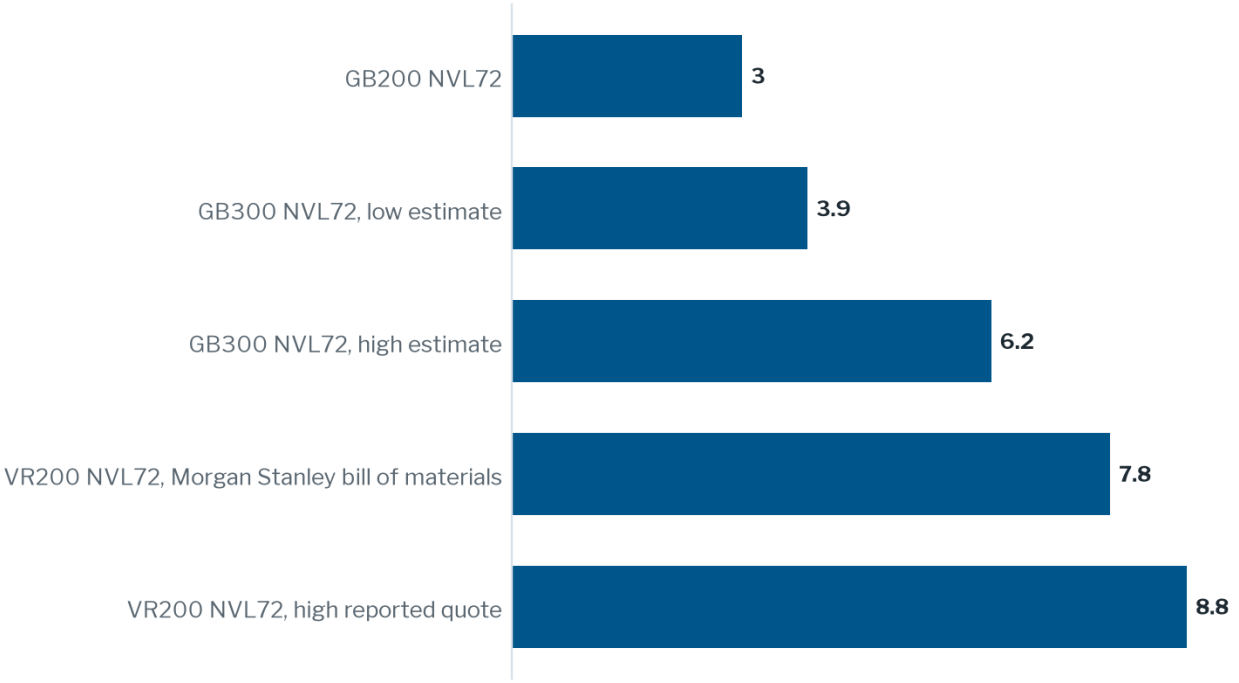
FACT List output pricing read from each vendor's own pricing documentation. Undiscounted, and large buyers negotiate. The 150-fold spread across one page is the point: the floor fell far faster than the ceiling, and the ceiling rose above the March 2023 GPT-4 output price of $60. [9] [10]

ASSESSMENT The frontier tier is what an agentic workload actually runs on, because agents are the workload that needs the strongest model, and that is the tier where the price has moved least. Cheap models got much cheaper, which is a genuine and useful deflation for classification, extraction and summarisation. It is close to irrelevant to a coding agent holding 675,000 tokens of context. Quoting category-wide token deflation at an agent buyer is quoting the wrong tier at them. [9] [10]

What a rack costs

THIRD-PARTY ESTIMATE Analyst estimates put a GB200 NVL72 rack near $3 million, a GB300 NVL72 rack somewhere between $3.7 million and $6.5 million depending on whose estimate and which configuration, and a VR200 NVL72 rack at roughly $7.8 million on a Morgan Stanley bill of materials, with quotes reported as high as $8.8 million. Memory is around $2 million of the Vera Rubin figure, about a quarter of the rack. [21]

FIGURE 4 **Estimated cost of one NVL72-class rack, millions of US dollars**



THIRD-PARTY ESTIMATE Third-party analyst estimates, not list prices; Nvidia does not publish rack pricing. The near-twofold spread between two estimates of the same GB300 rack is itself the finding: nobody outside the supply chain knows this number, which means no buyer can independently check a cost-per-token claim built on it. [21]

ASSESSMENT Throughput per megawatt and cost per token are different quantities, and the gap between them is capital. If a rack roughly doubles in price each generation while delivering a large multiple of throughput per megawatt, the efficiency gain is real for anyone whose binding constraint is power, and much smaller for anyone whose binding constraint is capital or utilisation. Most operators outside the largest clouds are constrained by utilisation. A 30x throughput-per-megawatt part running at 20 percent utilisation is not a 30x cost improvement, and none of the published claims state a utilisation assumption. [3] [21]

Where the gain stops

FACT Microsoft publishes a rate card for Copilot Studio that prices tokens in three tiers through its credit currency. Basic text and generative AI tools bill at 0.1 Copilot Credits per 1,000 tokens, standard at 1.5, and premium, which covers advanced reasoning, at 10 credits per 1,000 tokens. Pay-as-you-go credits list at one cent each, and prepaid packs run $200 for 25,000 credits, the same rate. [11] [12]

ASSESSMENT Converted at the published credit price, those three meters are $1, $15 and $100 per million tokens. The premium tier sits four to five times above frontier model list output pricing and roughly twenty times above frontier input pricing, and Microsoft charges it on top of a separate per-event feature rate for the same operation. Its own documentation states the rule plainly: total cost equals the feature rate for the operation plus the premium token rate for the reasoning model's token usage. That is the clearest disclosed instance in this evidence base of hardware and model deflation being retained above the buyer. [10] [11]

Copilot Studio token meter	Credits per 1,000 tokens	Dollars per million tokens at $0.01 per credit
Text and generative AI tools, basic	0.1	$1.00
Text and generative AI tools, standard	1.5	$15.00
Text and generative AI tools, premium (reasoning)	10	$100.00
Memo: Claude Opus 5 output, list	n/a	$25.00
Memo: GPT-5.6 Sol output, list	n/a	$20.00

ASSESSMENT Two caveats keep this honest. The credit meter is not a pure token resale; it bundles orchestration, connectors, governance and support, and a buyer choosing Copilot Studio is buying those. And a credit is not necessarily counted against the same tokenizer the model vendor uses, which is the next section's problem. Neither caveat closes a gap of that size. A buyer paying the premium meter should know they are paying roughly four times frontier list for the model call and should price the orchestration accordingly. [10] [11]

The token is not a fixed unit

FACT Anthropic's pricing documentation states that Claude 4.7 and later models, along with Claude Mythos Preview, use a newer tokenizer, and that this tokenizer produces approximately 30 percent more tokens for the same text. The exact increase depends on content and workload shape. Models at Sonnet 4.6 and earlier use the previous tokenizer. [9]

ASSESSMENT This is the most consequential disclosure in the entire evidence base and it appears as a footnote on a pricing page. Price per token and tokens per unit of work are two variables, and only one of them is in the contract. A buyer who negotiated a 30 percent per-token reduction and then upgraded across that tokenizer boundary paid exactly what they paid before for exactly the same work. Every audit clause written against a token count inherits this problem. The clause has to fix the tokenizer version, or meter in characters or requests, or it measures nothing. [9]

FACT Several other published multipliers move the effective rate without changing the headline price. Anthropic bills a five-minute cache write at 1.25x base input and a one-hour write at 2x, with cache reads at 0.1x. Pinning inference to US-only geography applies a 1.1x multiplier across input, output, cache writes and cache reads. Fast mode on Opus 5 doubles both sides to $10 and $50. Regional and multi-region endpoints on Bedrock and Google Cloud carry a 10 percent premium. These stack. [9]

FACT On marketplace billing the unit changes again. Anthropic rates usage in dollars at standard rates, applies any negotiated discount, converts to Claude Consumption Units at one cent each, and reports the quantity hourly to AWS Marketplace or Azure Marketplace. Anthropic's own

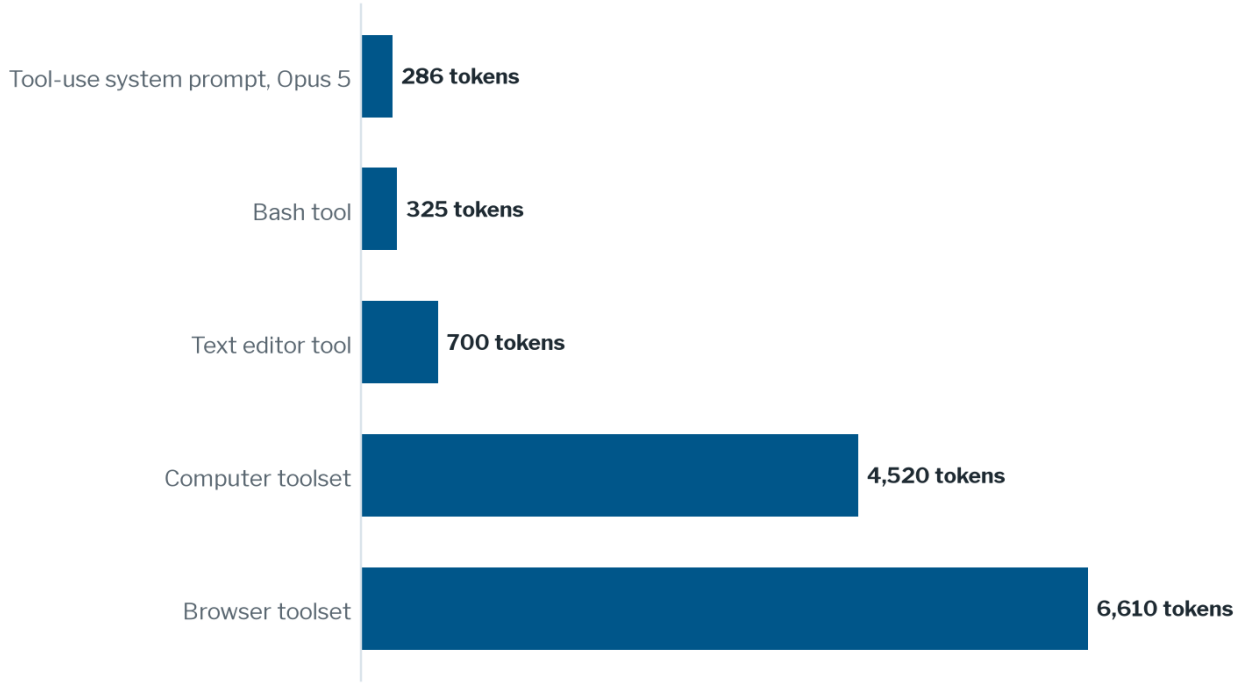
documentation states that the customer's AWS bill shows a single CCU line item, with the breakdown available in a separate console. [9]

ASSESSMENT A buyer procuring through a marketplace therefore receives an invoice denominated in a unit that exists solely for invoicing, derived from a token count produced by a tokenizer the vendor versions, aggregated into one line. Three transformations sit between the work performed and the number the finance team reconciles. That is not evidence of bad faith; the console breakdown exists and is detailed. It is evidence that the artefact the buyer has audit rights over, if they have any, is several steps removed from the thing being measured. [9]

What an agentic session bills that a seat never did

FACT Tool definitions are billed as input tokens on every request, not once per session. Anthropic publishes the counts: the tool-use system prompt runs 286 to 406 tokens on Opus 5 and 675 to 804 on Opus 4.7; the bash tool adds 325; the text editor tool 700; the computer toolset roughly 4,520; and the browser toolset roughly 6,610, rising by about 880 if all four optional members are enabled. [9]

FIGURE 5 **Input tokens a tool definition adds to every request**



FACT Published token counts from the vendor's own pricing documentation, measured on Claude Opus 5 where the figure is model-specific. These are charged per request, so a fifty-turn session declaring a browser toolset carries the 6,610 tokens fifty times. Prompt caching reduces the rate to 0.1x on a hit but does not remove the line. [9]

ASSESSMENT That arithmetic changes the shape of an agent bill. A fifty-turn session with a browser toolset declared carries roughly 330,000 input tokens of tool definitions alone before a single word of user content, task context or tool output. Caching cuts the rate to a tenth on a hit, which is exactly why cache hit rate is the metric AgentX was built to measure and exactly why a

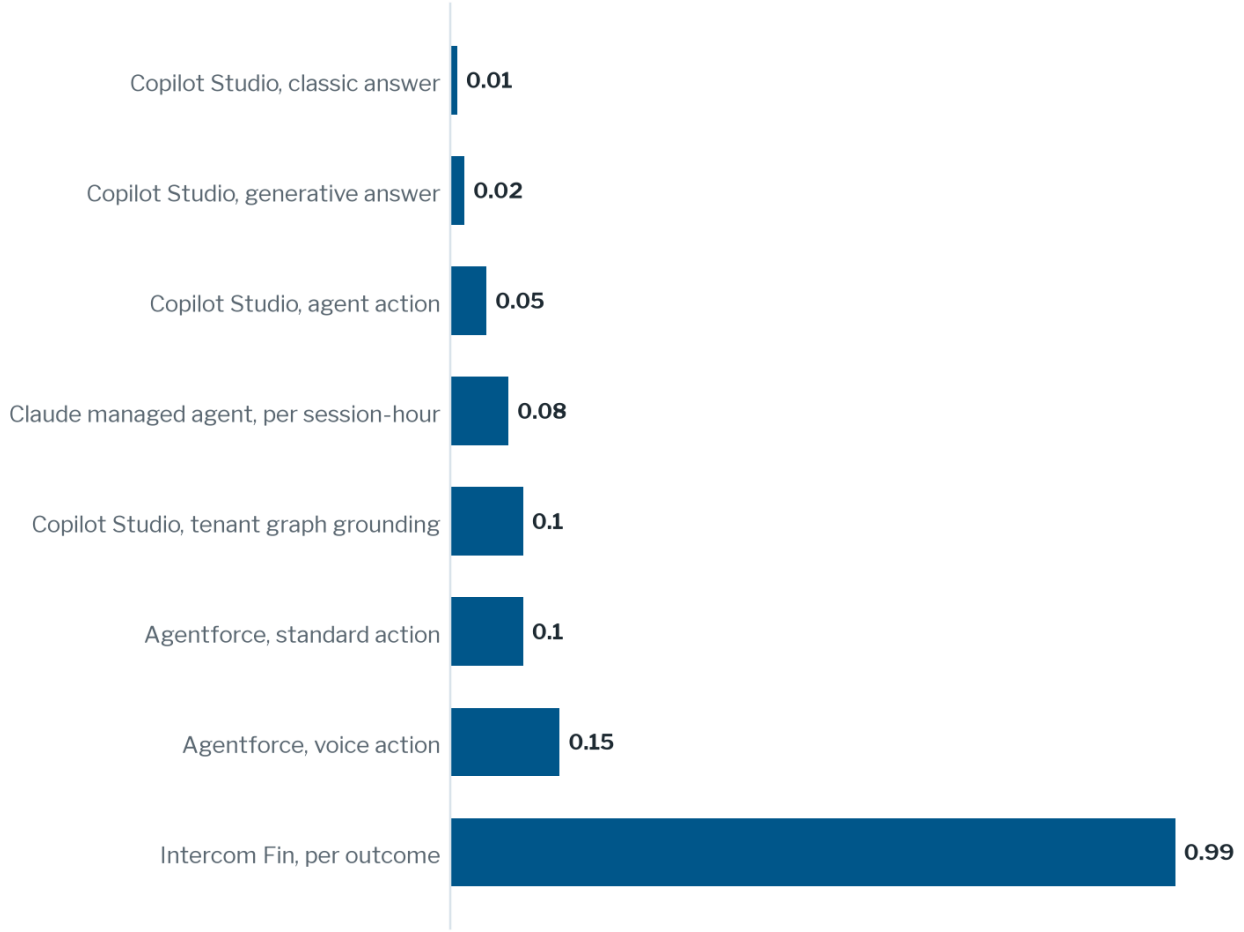
buyer cannot verify their own bill without it. Cache hit rate is reported by the provider and is not independently checkable by the customer. [1] [9]

FACT Beyond tokens, providers bill events that have no per-seat analogue. Anthropic charges $0.08 per session-hour for managed agent sessions, metered to the millisecond while the session is running, on top of all token charges, and states that the batch discount does not apply to sessions. Web search costs $10 per 1,000 searches. Code execution runs $0.05 per container-hour beyond 1,550 free hours a month, with a five-minute minimum, and is billed even when the tool is not called if files were included in the request. [9]

FACT Microsoft's rate card bills 1 credit for a classic answer, 2 for a generative answer, 5 for an agent action, 10 for tenant graph grounding, 13 per 100 agent flow actions and 8 per page of content processing. Computer-using agents bill at the agent action rate and are explicitly excluded from the Microsoft 365 Copilot licence inclusion. Voice runs 10, 35 or 75 credits per minute by tier. Unused prepaid credits do not carry over month to month. [11]

VENDOR CLAIM Salesforce prices Agentforce actions at 20 Flex Credits, or $0.10, with voice actions at 30 credits or $0.15, credits sold in packs of 100,000 for $500, and one action covering up to 10,000 tokens. Its own worked example puts a single sales agent interaction that answers a product question and books a meeting at three to six actions, or $0.30 to $0.60. Intercom bills Fin at $0.99 per outcome on top of seats priced at $29, $85 and $132 per seat per month, with an outcome recorded when a customer confirms resolution, stops asking, or Fin completes a procedure including a handoff. [13] [14]

FIGURE 6 **Published price of one agent event in US dollars, across four vendors**



FACT List prices from each vendor's own documentation, converted to dollars at the published credit rate where the vendor meters in credits. The events are not equivalent units of work and the chart makes no claim that they are; the point is that eight distinct billable events now exist where a per-seat contract had one line. [9] [11] [13] [14]

Billable event	Billing unit	Published price	Existed under per-seat?
Input and output tokens	per million tokens	$0.20 to $180 by model and tier	No
Cache write	per million tokens	1.25x or 2x base input	No
Cache read	per million tokens	0.1x base input	No
Tool definition overhead	input tokens per request	286 to 6,610 tokens	No
Server-side tool call	per search	$10 per 1,000 web searches	No
Session runtime	per session-hour	$0.08, metered to the millisecond	No
Container execution	per container-hour	$0.05 beyond 1,550 free hours	No

Billable event	Billing unit	Published price	Existed under per-seat?
Data residency	multiplier on all token categories	1.1x for US-only inference	No
Speed tier	multiplier on input and output	2x for fast mode on Opus 5	No
Agent action	per action	$0.05 to $0.10	No
Grounding retrieval	per grounded response	$0.10	No
Resolved outcome	per outcome	$0.99	No
Named user	per seat per month	$29 to $132	Yes, and still charged

ASSESSMENT The last row is the one procurement misses. Every vendor in this sample charges consumption in addition to seats, not instead of them. The per-seat model has not been replaced by consumption; it has been supplemented by it. A renewal that reads as flat on the seat line can carry an unbounded second line beneath it, and the seat count no longer bounds the exposure because an autonomous agent consumes without a human present. [11] [13] [14]

FACT The consumption is bursty in a way seats never were. Gartner estimates agents require 5 to 30 times more tokens than an equivalent chatbot task, and that a task costing a chatbot one cent can cost an agent up to $1.50. Reported incidents match the shape: Amazon is reported to have spent $500 million on AI in a single month after rolling out access without usage caps, Uber to have consumed its entire 2026 AI coding budget in four months, and a mid-market software chief executive to have run up a $1,000 charge over one weekend debugging code, auto-charged to a card on auto-renew. A survey cited alongside those reports puts 68 percent of US companies over budget on AI initiatives, with a third saying it happens mostly or always. [17] [18]

ASSESSMENT Those incident figures are reported second-hand through a single article and should not be treated as measurements. What they establish is not the magnitude but the mechanism: in each case the overspend was discovered after the fact because nothing in the billing path stopped it in real time. That is a control failure, not a pricing failure, and it is fixable in the contract. [18]

Audit rights, and what the disclosed record shows

FACT Info-Tech Research Group names five risk categories for agentic AI agreements: unclear billing definitions, hidden cost drivers, weak financial controls, vendor-controlled pricing changes, and lock-in and accountability gaps. It observes that a single user prompt can trigger tool calls, recursive reasoning loops, background API calls, model-tier escalation, retries and multi-agent orchestration, each billable. It names limited audit rights as a weakness in current agreements, and its principal research director states that clear definitions, consumption caps, audit rights, pricing-change protections and dispute mechanisms must be negotiated before signing. [15]

ASSESSMENT Info-Tech is describing the right five things and the framing has one gap. Four of the five are about what the vendor may do. The fifth, weak financial controls, is about what the buyer failed to build. The evidence in this report suggests the binding constraint is measurement rather than either: a buyer with a cap and an audit clause still cannot verify a token count they have no independent way to produce. Caps are enforceable because the vendor meters them. Audits are not, unless the clause specifies what is being audited against. [15]

THIRD-PARTY ESTIMATE The one quantified data point in the public record comes from an audit vendor. It reports reviewing $34 million of AI spend across 60 companies since March 2026 and identifying close to $1.7 million in mistaken overcharges, roughly five percent of the spend reviewed, with about 80 percent credited back by providers including AWS, Google Cloud, Microsoft Azure, Anthropic and OpenAI. Named customers include Panasonic, HP and Honda. [19] [20]

ASSESSMENT Discount the five percent. The firm sells the audit, the sample is self-selected toward companies who suspected a problem, and no methodology is published. What survives that discounting is the taxonomy, which is specific enough to be checkable and matches the architecture: failed requests still charged, duplicate charges from agent retry storms, billing during provider outages, premium rates applied to cheaper or older models, and orchestration errors that fan one request out to several models at once. Its founder's framing is the more useful disclosure: customers often lack independent visibility into which model handled a request, how it was routed, or whether it was cached, retried or deduplicated before it appeared on the bill. [19]

ASSESSMENT One further item belongs in the record and is deliberately kept out of the argument. A July 2026 report states that Anthropic confirmed a $16.6 million billing error. It was seen only as a headline and the article could not be retrieved, so it is second-hand and unverified, it is cited here so a reader can check it themselves, and no finding, chart or recommendation in this document rests on it. If it is accurate, it strengthens the case for reconcilable metering considerably, which is precisely why it should not be used until someone has read the underlying disclosure. [22]

ASSESSMENT That sentence is the answer to the fourth question in the brief. No major provider in this evidence base publishes a metering artefact a buyer could reconcile against an invoice. Anthropic publishes per-request usage in the API response, including cache read and write token counts and server tool use, which is more than most, but the buyer's own logs are the only record and the vendor's meter is authoritative in any dispute. An audit right over a number only one party can produce is a right to be shown the vendor's arithmetic, not to check it. [9] [19]

HYPOTHESIS A billing dispute in this category will be resolved by contract language rather than by measurement, because neither party can produce an independent count. This predicts that the first published enterprise dispute over agent token billing turns on the definition of a billable event, not on a discrepancy in the count. It would be refuted by a dispute resolved through a third-party reconciliation of provider-side and customer-side logs, which is the outcome that would also make audit clauses meaningful.

FACT One vendor does ship an enforceable control. Microsoft's published enforcement policy triggers at 125 percent of prepaid capacity, disables custom agents, and returns a user-facing

message reading that there is a billing issue or that the agent has reached its usage limit. In-flight conversations complete. Administrators are notified by email and in the admin centre. Per-agent monthly consumption limits can be set before enforcement triggers. Agent flow enforcement blocks new runs while leaving the rest of the agent working. [11]

ASSESSMENT That is the model to demand contractually from every other vendor. It is a hard cap, it fails closed, it fails gracefully mid-conversation, it notifies a named human, and the buyer can set it below the contractual ceiling. Info-Tech's recommendation that caps be mandatory rather than advisory is already implemented by one large vendor, which removes the usual objection that it cannot be built. A vendor declining a hard cap in 2026 is declining a control that a competitor ships by default. [11] [15]

Info-Tech risk category	The term that closes it	What it costs to skip
Unclear billing definitions	Define the billable event in the contract, name the tokenizer version, and require notice before it changes	A 30 percent tokenizer change cancels a 30 percent price cut with no breach
Hidden cost drivers	Enumerate every meter: tokens, cache writes, tool definitions, server tool calls, session-hours, container-hours, multipliers	Meters you did not enumerate are meters you cannot dispute
Weak financial controls	Hard cap that fails closed at a buyer-set threshold, per-agent sub-caps, notification to a named owner	Discovery happens at the invoice, which is 30 days after the spend
Vendor-controlled pricing changes	Price hold for the term, notice period, and the right to exit at the old price if a meter rate changes	A rate change on one meter reprices the contract without touching the headline
Lock-in and accountability gaps	Export of usage telemetry in a machine-readable form, audit right against a named artefact, dispute path with a clock	An audit right over a number only the vendor can produce is not an audit right

Scenarios to 2028

Point forecasts do not survive in a category where a benchmark's own maintainer republishes results every three to four weeks because the software baseline moves. Three scenarios, with subjective probabilities and a first visible sign for each. The probabilities are the least reliable content in this document. The signs are the most useful, because any reader can watch for them without private information.

SCENARIO A Metering standardises 20 percent

> A credible per-event metering standard emerges, whether from a cloud marketplace, a procurement consortium or a regulator, and providers publish reconcilable usage artefacts.

Consequence. Audit clauses become enforceable, billing disputes get resolved on evidence, and pass-through becomes measurable rather than assumed.

Earliest visible sign. A hyperscaler marketplace publishes a token-usage export schema, or two providers agree on a common definition of a billable agent event.

SCENARIO B The platform layer keeps the gain 55 percent

Hardware efficiency keeps improving, model list prices keep drifting down at the low end, and the orchestration layer above the model captures the difference through per-event and premium-token meters.

Consequence. Buyer cost per unit of agent work stays roughly flat or rises while every vendor in the chain truthfully reports getting more efficient.

Earliest visible sign. A second major platform introduces a reasoning-tier token meter priced well above frontier model list, or an existing per-event rate rises while the headline seat price holds.

SCENARIO C Buyers force outcome pricing 25 percent

Enough budget overruns land on enough CFOs that buyers refuse consumption terms and push risk back onto vendors through per-outcome or per-resolution pricing.

Consequence. Vendors absorb the token variance, which advantages those with the best inference economics and squeezes resellers with no infrastructure of their own.

Earliest visible sign. A large enterprise publicly renegotiates from consumption to outcome pricing, or a second major agent vendor follows Intercom onto a per-outcome list price.

Leading indicators

Six observable events. Each is public, dated and checkable without access to anyone's contract.

If this happens	It means	Moves toward	Where to watch
SemiAnalysis publishes reviewed Vera Rubin AgentX results	The 30x claim is either confirmed on an independent run or quietly restated	A or B depending on the number	inferencex.semianalysis.com and the promised AgentX update
A vendor publishes a tokens-per-watt figure on a named model	Intel's framing becomes a comparable metric instead of positioning	A	Intel Crescent Island launch materials
Groq 3 LPX figures are restated with batch size and concurrency	Aggregate throughput claims become usable for capacity planning	A	Nvidia technical blog and Artificial Analysis methodology notes
A frontier model launches above the current top of the price sheet	The ceiling continues to rise while the floor falls	B	OpenAI and Anthropic pricing pages, checked on launch day

If this happens	It means	Moves toward	Where to watch
A second vendor ships a hard cap that fails closed by default	Financial control stops being a differentiator and becomes table stakes	A or C	Vendor documentation, not press releases
A named enterprise moves a large agent contract to per-outcome pricing	Token variance shifts from buyer to vendor	C	Earnings calls and customer case studies from agent vendors

Implications

For CFOs signing consumption-based AI contracts

ASSESSMENT Model the downside on tokens per unit of work, not on price per token, and assume the first stays volatile while the second falls slowly at the tier your agents actually use. The deflation being quoted at you is real and mostly belongs to the cheap tier. Treat any AI line without a hard cap as an uncapped liability, because the reported failures all share the same shape: nothing in the billing path stopped the spend and discovery happened at the invoice. Require the cap to fail closed at a threshold you set, and require a named human to be notified. One large vendor already ships exactly that, so the capability objection is closed. [11] [17] [18]

ASSESSMENT On accounting, the seat line no longer bounds the exposure. Every vendor in this sample charges consumption in addition to seats, and an autonomous agent consumes with no human present, so headcount-based forecasting has stopped working for this category. Forecast on workflow volume and a per-workflow token band, and hold a variance reserve sized to the fat tail rather than the median, because retry behaviour puts the cost distribution's weight in sessions that run many times the median. [9] [11] [13]

For procurement leads negotiating agent billing terms

ASSESSMENT Four clauses, in priority order. Define the billable event, including the tokenizer version and a notice requirement before it changes, because a tokenizer that produces 30 percent more tokens for the same text will silently reverse a 30 percent price concession. Enumerate every meter in a schedule, not just the token rate, because the ones you did not name are the ones you cannot dispute. Take the hard cap. Then negotiate the audit clause against a named artefact rather than against a token count, since the count is a number only the vendor can produce. [9] [11] [15]

ASSESSMENT Two negotiating levers are undervalued. Price holds matter more than discounts here, because the rate card has more lines than the contract and any one of them can move; a hold on all meters for the term is worth more than a few points off the token rate. And the right to exit at the old price if any meter changes converts a unilateral pricing right into a bilateral one. Info-Tech's list of caps, throttles, kill switches, dispute paths and pricing-change protections is the right checklist; the addition is to insist that each one names the artefact it operates on. [15]

For infrastructure architects sizing inference capacity

ASSESSMENT Do not size from any published throughput number in this report. Aggregate tokens per second with undisclosed batch size and concurrency cannot be converted into a user count, and Nvidia's own documentation confirms the Groq 3 LPX figures are aggregate. Size instead from AgentX-shaped inputs: a 675,000-token input p99 against an 8.6k output p99, a KV cache hit rate you have measured on your own traffic rather than inherited, and a sub-agent burst factor. Cache hit rate is the variable that decides your cost, which is why it is the variable the benchmark was designed around. [1] [5]

ASSESSMENT On part selection, the honest comparison is capacity per dollar at your context length, not throughput per megawatt. Intel's Crescent Island argument, FP8 weights and KV cache on a single card at 160 GB to 480 GB, is the most relevant claim in the set for long-context serving even though Intel publishes neither a bandwidth figure nor the tokens-per-watt number it frames itself on. The disaggregated pattern Nvidia describes, GPUs for prefill and SRAM parts for decode, is architecturally sound and separates two workloads with genuinely different memory profiles. Both deserve evaluation on your own traces. Neither deserves procurement on a press release. [5] [8]

For vendor pricing strategists setting token rates

ASSESSMENT The per-event meter is working commercially and is accumulating a specific reputational liability. Charging a premium reasoning tier at roughly $100 per million tokens on top of a per-event feature rate, while frontier list runs $20 to $25, is defensible as bundled orchestration and indefensible if a buyer discovers the multiple by doing the arithmetic themselves rather than reading it from you. The buyer who works it out from your own published rate card, as this report did, does not conclude they were sold orchestration. They conclude the disclosure was structured to be tedious. [10] [11]

ASSESSMENT There is a real opening for whoever moves first on metering. Publish a reconcilable usage artefact, hold meter rates for the contract term, and ship a buyer-set hard cap that fails closed, and you convert the strongest current objection into a differentiator while your competitors are still arguing that caps are operationally hard. The reported case of a chief executive running up a weekend charge with no dispute path is a churn event and a story that travels. Outcome pricing is the harder version of the same move: it wins the deals where the buyer refuses variance, and it only works if your inference economics are genuinely better than your competitors', which is testable on AgentX. [1] [14] [18]

What this document cannot answer

Six questions matter to this decision and cannot be settled from public sources. Each needs primary research.

What buyers actually pay. Every price in this report is a list price. Enterprise discounting on tokens, credits and per-event rates is not disclosed by any vendor here, and the real pass-through could be materially better or worse than the list-price comparison shows. Settling this needs a sample

of executed contracts under NDA, which is the single highest-value piece of primary research in the category.

Whether the 15x and 30x figures are compatible. Nvidia publishes both without stating whether they were produced on the same workload. If they were not, the compounded figure is not arithmetic. Answering it needs either a disclosure from Nvidia or an independent run of both generations on one benchmark.

What Vera Rubin does on an independently reviewed AgentX run. Nvidia says its results are pending SemiAnalysis review. Until that publishes, the 30x and 35x claims are unreviewed vendor numbers on a benchmark whose own maintainer has not signed off.

The batch size behind the Groq 3 LPX figures. Without it, 3,431 aggregate tokens per second cannot be turned into a per-user interactivity number, and interactivity is the entire premise of the product.

Where the margin actually sits. This report infers that the orchestration layer retains the gain, from published per-event pricing. Nobody breaks out gross margin at this level. The inference is consistent with the evidence and is not the same thing as a measurement.

Whether any enterprise contract in this category grants a meaningful audit right. Info-Tech names limited audit rights as a category and recommends negotiating them. No executed clause is in the public record, and no provider publishes the artefact such a clause would run against. A survey of what buyers have actually secured would settle it.

Half-life of this assessment

Decays within six months: every price in this document. The published sheets moved several times inside the twelve months before publication, one flagship tier fell threefold, and one vendor cancelled a scheduled increase after announcing it. The Vera Rubin and Groq 3 LPX figures also sit in this bucket, because a reviewed AgentX result is expected within weeks of publication and either confirms or restates them.

Decays within twelve months: the specific meters and multipliers. Cache write multipliers, geography and speed premiums, session-hour rates and credit conversion rates are all young enough that vendors are still discovering what the market will bear. The billable-event table should be re-derived from source rather than reused. The scenario probabilities also belong here; the scenarios themselves and their tripwires last longer.

Decays within twenty-four months: the competitive positions. Intel's part is pre-launch with no benchmark behind it, AMD's context-parallelism gap is a software problem and software gaps close, and the SRAM-decode architecture is one generation into commercial production. Any read of who wins on agentic inference economics should be treated as expired by late 2028.

Architectural, and unlikely to decay: the structural argument. As long as the vendor defines the unit, meters the usage, and issues the invoice, the buyer has no independent count, and every efficiency gain passes through a layer with both the incentive and the mechanism to retain it. That

does not depend on this generation of silicon or this year's price sheet. It changes only if metering becomes reconcilable, which is scenario A and is why it is the one worth watching.

Limitations

This is independent work. No vendor named here commissioned, reviewed, funded or was shown this report before publication, and the author holds no position in any company mentioned and has no commercial relationship with any of them.

The evidence base has four specific weaknesses worth naming. First, three sources refused automated retrieval during research: Salesforce's pricing page, Gartner's press release and OpenAI's marketing pricing page all returned HTTP 403. OpenAI's figures were taken from its developer documentation instead, which is primary. Gartner's figures were taken from Computerworld's contemporaneous coverage, which quotes the analysts directly. The Agentforce figures were taken from consistent secondary summaries and are tagged as vendor claims rather than verified fact for that reason.

Second, the March 2023 GPT-4 launch price of $30 and $60 per million tokens is no longer on OpenAI's own page and is cited from a pricing-history compilation. It is consistent across several such compilations and is treated here as second-hand. No conclusion rests on it; the Anthropic flagship comparison, which is verifiable on a single current page, carries that argument instead.

Third, one item was seen only as a headline and could not be opened: a report that Anthropic confirmed a $16.6 million billing error in July 2026. It is mentioned here for completeness and is excluded from every finding, chart and recommendation in this document. A reader should treat it as unverified.

Fourth, the secondary summaries of Intercom's Fin pricing describe additional charges for procedure handoffs, disqualifications and qualifications that do not appear on Intercom's own pricing page, which describes a single $0.99 per-outcome charge covering procedure completion including handoffs. This report uses the primary page. The discrepancy is noted because it is a good illustration of why per-event pricing is hard to audit even before a contract is signed.

Finally, the central pass-through comparison sets vendor efficiency claims against published list prices. Those are different kinds of number measured by different parties for different purposes, and the comparison is a reasoned judgment rather than a calculation. It is labeled as an assessment throughout. A reader who believes enterprise discounting has moved much faster than list pricing should discount the size of the gap while noting that the direction is not in dispute.

Sources

[1] SemiAnalysis (Cam Quilici, Bryan Shan, Alec Ibarra and others). "AgentX InferenceXv3: Does the CUDA Moat Still Hold?" 24 August 2026. *Analyst firm* semianalysis.com

[2] SemiAnalysisAI. InferenceX benchmark repository, released under Apache 2.0. Accessed 27 August 2026. *Analyst firm* github.com

[3] NVIDIA. "Vera Rubin NVL72 Efficiency for AI Agents." NVIDIA Blog, 24 August 2026. *Vendor primary* blogs.nvidia.com

[4] NVIDIA. "NVIDIA Groq 3 LPX Now in Full Production With World-Class Speed for Agentic AI." NVIDIA Newsroom, 24 August 2026. *Vendor primary* nvidianews.nvidia.com

[5] NVIDIA. "How NVIDIA Groq 3 LPX Unlocks Ultrafast Interactivity at Long Context on NVIDIA Vera Rubin." NVIDIA Technical Blog, 24 August 2026. *Vendor primary* developer.nvidia.com

[6] Harold Fritts. "NVIDIA Groq 3 LPX Enters Full Production: 3,400 Tokens per Second at 100K Context, 256 LP30s per Rack." StorageReview, 24 August 2026. *Independent press* storagereview.com

[7] Maximilian Schreiner. "Nvidia says its Groq 3 LPX is four times faster than Cerebras, but the math is more complicated." The Decoder, 25 August 2026, citing analysis in The Register. *Independent press* the-decoder.com

[8] Patrick Kennedy. "Intel Crescent Island 160GB to 480GB LPDDR5X AI GPU at Hot Chips 2026." ServeTheHome, 24 August 2026. *Independent press* servethehome.com

[9] Anthropic. "Pricing." Claude platform documentation. Accessed 27 August 2026. *Vendor primary* platform.claude.com

[10] OpenAI. "Pricing." API documentation. Accessed 27 August 2026. *Vendor primary* developers.openai.com

[11] Microsoft. "Billing rates and management, Microsoft Copilot Studio." Microsoft Learn, updated 3 August 2026. *Vendor primary* learn.microsoft.com

[12] Microsoft. "Licensing for agents, Microsoft Copilot Studio." Microsoft Learn, updated 3 August 2026. *Vendor primary* learn.microsoft.com

[13] Salesforce. "Agentforce Pricing" and "Flexible Pricing is a Full Circle Moment for Salesforce." Figures taken from consistent secondary summaries; the pages refused automated retrieval on 27 August 2026. *Vendor primary (retrieval blocked)* salesforce.com

[14] Intercom. "Pricing." Accessed 27 August 2026. *Vendor primary* intercom.com

[15] Info-Tech Research Group. "Agentic AI Contracts Expose Organizations to Runaway Costs With Limited Recourse." PR Newswire, 26 August 2026. Underlying blueprint: Negotiate Safe AI Contracts to Prevent Bill Shock. *Analyst firm* prnewswire.com

[16] Gartner. "Gartner Predicts AI Inference Costs Per Agentic Workflow Will Increase More Than Fivefold Through 2028." Press release, 17 August 2026. Page returned HTTP 403 to automated retrieval; figures taken from source 17. *Analyst firm* gartner.com

[17] Taryn Plumb. "AI inference is getting cheaper, but your agents are getting more expensive." Computerworld, 17 August 2026, quoting Gartner analysts Will Sommer and Sabine Zimmerhansl. *Independent press* computerworld.com

[18] Fortune. Report on token spend, budget overruns and the end of tokenmaxxing, 22 August 2026. *Independent press* fortune.com

[19] Vaudit. "Vaudit Launches TokenAudit to Recover Millions in Enterprise Token Spend Billing Errors From Anthropic, OpenAI, and AI Providers." BusinessWire, 30 June 2026. The firm sells the audit product described. *Vendor primary (commercial interest)* businesswire.com

[20] Tim Keary. "After 'Tokenmaxxing', Token Spend Has Become The New Metric To Watch." Forbes, 10 July 2026. *Independent press* forbes.com

[21] Tom's Hardware, reporting a Morgan Stanley bill-of-materials estimate for VR200 NVL72, with GB200 and GB300 rack estimates from Loop Capital and others. Analyst estimates, not list prices. *Second-hand (unverified)* tomshardware.com

[22] TechTimes. Report that Anthropic confirmed a $16.6 million billing error, 12 July 2026. Seen only as a headline; the article could not be retrieved. Excluded from every finding in this report. *Second-hand (unverified)* techtimes.com

[23] Compiled LLM API pricing histories used only for the March 2023 GPT-4 launch price of $30 and $60 per million tokens, which is no longer published on OpenAI's own pages. *Second-hand (unverified)* deploybase.ai

Rights and permitted use

It is published for public reading. You are free to share the complete, unmodified file, to link to it, and to quote short passages in commentary, reporting, or discussion, provided the author and the site above are credited.

The following require prior written consent: republishing this work in whole or in substantial part under another name, masthead, or brand; excerpting or modifying it in a way that alters the analysis or removes the evidence classifications; incorporating any portion of it into a paid, commissioned, or client-facing deliverable; resale or distribution behind a paywall; and use of this document as training data for machine learning models.

Commissioned Market Assessment reports, commercial licensing, and briefings on this material are available.

This document is analysis, not investment, legal, or procurement advice. It was not commissioned, reviewed, or funded by any company named in it, and the author holds no position in any of them.

Market Assessment reports, commissioned research, and briefings: **davidsoden.com/market-assessment**