The Denominator Problem: Compute Commitments and Delivered Work

A Preprint

Sidney Scott
Mind Meets Matter
sidney@mindmeetsmatter.com

Abstract

A frontier lab commits capital in megawatts, chip-counts and fixed dollars, and earns revenue in tokens whose count the model itself chooses. Those two ledgers are denominated in units that do not convert into one another, and no contract, instrument or standard performs the conversion. We state the consequence precisely. Under take-or-pay capacity commitments, a reduction in cost per token produces no cash saving at all, because committed spend is fixed. On flat-rate products, contribution margin falls whenever tokens per task and tasks per seat grow faster than cost per token falls, and every efficiency instrument in production use reports improvement throughout, because each measures one term of a three-term product and none measures the product. A sensitivity sweep over the parameter plane shows the crossover binds across most of the region real serving organizations occupy, and two ablations isolate the mechanism: on a metered contract the crossover does not exist at any parameter value, and the efficiency available from cache composition is bounded and is exhausted above a hit rate of 82.8 percent, below the rates now reported. Across the four laboratories with the largest announced commitments we find every agreement indexed to capacity, power, chip-count or a fixed dollar schedule and none to delivered work, and every named megasite placed by power availability with no declared service-level class, specify a denominator that fuses cost, service level and quality with the conditions under which a value of it is defined, and demonstrate on fourteen production endpoints serving one fixed artifact that the cheapest providers by price and the fastest by delivered speed are disjoint sets.

Keywords compute procurement · unit of account · service-level objectives · goodput · take-or-pay contracts · inference economics

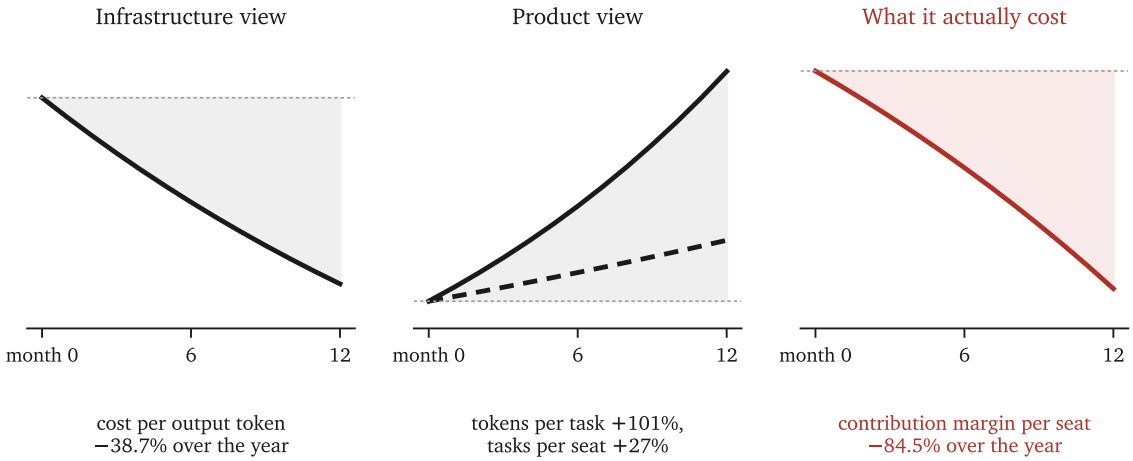


Figure 1: The same twelve months, read three ways. Every series is generated by the five parameters of Appendix B. 1. The infrastructure and product views are what an organization measures and both move the approved direction. The third panel is what the product earned. No instrument in standard production use plots them together. Illustrative parameters [EST].

1 Introduction

The four laboratories with the largest announced compute commitments have between them contracted for capacity on a scale without precedent in corporate history. One reports commitments of up to $517 billion across at least 14.8 gigawatts (The Information, 2026); another has stated that its cloud and buildout commitments total approximately $1.4 trillion (Data Center Dynamics, 2026). Every one of these agreements is written in units of capacity, power, chip-count or fixed dollars. Revenue, for those of them that sell inference, arrives per output token or per seat, and the number of output tokens a task consumes is chosen by the model rather than by the customer who wanted the task done.

This is a missing conversion rate between two ledgers, and its absence has three consequences that compound at the scale the commitments now reach. Committed capacity is paid for whether or not it is used, so efficiency gains on the committed portion of a fleet produce no cash saving. Flat-rate products break the identity that makes per-token efficiency and margin move together. And the instruments that would reveal either effect are distributed across three organizations that each hold one term of a three-term product.

The industry has assembled every component of the missing unit separately. Goodput counts requests meeting declared latency objectives but carries no price and no acceptance test (Zhong et al., 2024). MLPerf Inference enforces percentile latency bounds at accuracy floors but has no cost denominator in any scenario (MLCommons, 2025). Software Carbon Intensity standardizes a rate per functional unit, for carbon (ISO/IEC, 2024). Token Arena fuses price, speed and live-endpoint quality into dollars per correct answer, as a benchmark rather than a settlement basis (Token Arena, 2026). Each is real, useful, and missing exactly one leg of the unit a contract would need.

Our contributions are as follows. **1) The crossover condition, measured.** We state the condition under which every efficiency instrument improves while contribution margin compresses, sweep it over the parameter plane, and locate where it binds. **2) Two ablations that isolate the mechanism.** The crossover disappears entirely on a metered contract at every parameter value, and the efficiency available from cache composition is bounded and already exhausted at reported hit rates, so neither the growth rates nor the caching term is what carries the result. **3) The unit, its conditioning tuple, and the contract form it implies.** We specify a denominator fusing cost, service level and quality, demonstrate the inversion it exposes on fourteen production endpoints, and derive where utilization and obsolescence risk move under each indexation form.

2 The two ledgers

One side of a frontier lab's ledger is written in physical units. The other is written in a unit the seller controls.

The commitment ledger. The Amazon agreement secures up to 5 gigawatts of capacity, with nearly 1GW of Trainium2 and Trainium3 online by the end of 2026 (Anthropic, 2025a). The Google and Broadcom agreement covers multiple gigawatts of next-generation TPU capacity from 2027 (Anthropic, 2026). The Microsoft agreement is $30 billion of Azure compute capacity plus up to one gigawatt contracted additionally (Microsoft, 2025). The Fluidstack agreement is a $50 billion buildout in Texas and New York (Anthropic, 2025b). Not one is denominated in output.

The revenue ledger. Revenue arrives per output token on metered products and per seat per month on subscriptions. The token count for a given piece of work is set by reasoning length, tool-call depth, and how many turns an agent takes. It is not a property of the task.

3 The central result

> **Claim 1.** On flat-rate products, contribution margin per seat falls whenever the combined growth rate of tokens per task and tasks per seat exceeds the rate of decline in cost per token, and no instrument in standard use reports this composite.

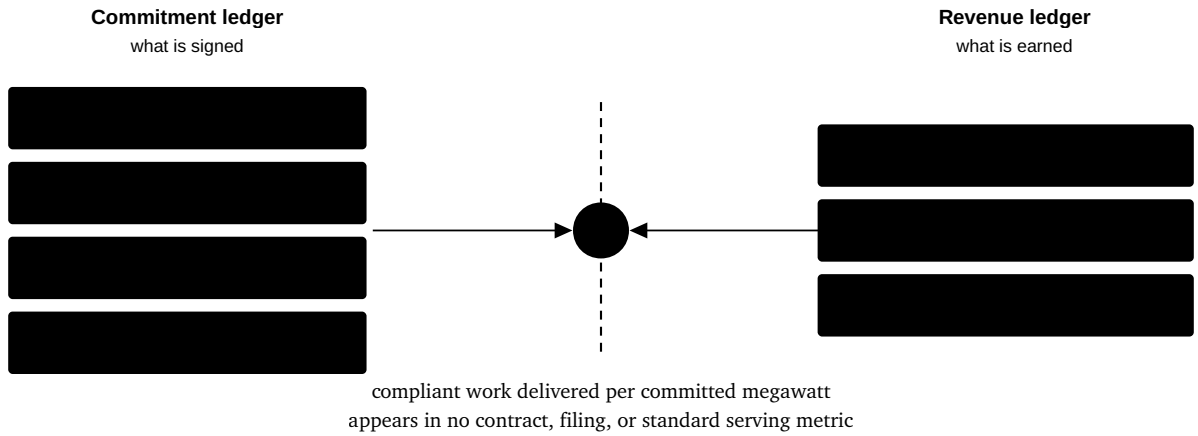


Figure 2: The two ledgers and the conversion between them. Converting one into the other requires knowing how much compliant work a committed megawatt delivers, where compliant means inside the latency contract under which the output has value and above the quality floor at which the customer would accept it. That quantity appears in no contract, no filing, and none of the standard serving metrics.

Let c be cost per output token served, n the tokens consumed to complete one task, and T the tasks a seat runs per month. Cost to serve one seat per month is

$$K = c \cdot n \cdot T \tag{1}$$

In growth rates, $\mathrm{d}(\ln K)/\mathrm{d}t = \mathrm{d}(\ln c)/\mathrm{d}t + \mathrm{d}(\ln n)/\mathrm{d}t + \mathrm{d}(\ln T)/\mathrm{d}t$. Revenue per seat R is fixed by the price list, so contribution margin $R - K$ compresses whenever

$$d(\ln n)/dt + d(\ln T)/dt > -d(\ln c)/dt \tag{2}$$

The right-hand side is the efficiency term, and it is the only term any serving instrument measures. Cost per million tokens, model FLOP utilization, tokens per second per megawatt, goodput and the InferenceMAX frontier (SemiAnalysis, 2026) all describe how c moves. They are silent on n and T by construction, while the product direction of 2025 and 2026 raises both deliberately.

Why this is not the usual efficiency story. On a per-token contract the identity is benign. Revenue per task is $p{\cdot}n$ and cost per task is $c{\cdot}n$, so gross margin ratio $(p - c)/p$ is independent of n, and margin per task rises with n. Token growth is accretive on metered products. The exposure is created by the price form, not the model, which Section 7.3 tests directly.

3.1 The efficiency term is not one thing

Equation (2) treats c as a scalar. In a production serving stack it is a blend. With cache hit rate h and a cached-read price ratio κ (0.1 or lower against base input on current published schedules), the effective rate is

$$c = c_0 \cdot [\ \kappa h + (1 - h)\ ] \tag{3}$$

where c_0 is the uncached rate. This matters because h and n are mechanically coupled rather than independent. An agentic loop that raises tokens per task raises cache hit rate at the same time, since each turn re-presents a growing shared prefix. Analyst reconstructions put hit rates above 90 percent on agentic workloads (SemiAnalysis, 2026), which is why a blended cost per token can fall steeply while nothing about the hardware improved. Section 7.4 shows this source of efficiency is bounded and, at reported hit rates, already spent.

3.2 A worked example

The parameters are illustrative and tagged [EST]. They favour the efficiency case: cost per token falls 4 percent a month, faster than most serving organizations achieve, while tokens per task grow 6 percent and tasks per seat 2 percent a month, slower than the agentic product direction implies. The exact monthly composite is $0.96 \times 1.06 \times 1.02 = 1.03795$. Take a seat priced at \$30 a month with a starting cost to serve of \$18, a 40 percent contribution margin.

Table 1: Twelve months of successful efficiency work under growing consumption. Every figure is generated by the five parameters above; the reproduction is Appendix B.1.

Month	Cost per token (index)	Cost to serve one seat	Contribution margin	Margin %
0	100.0	$18.00	$12.00	40.0%
3	88.5	$20.13	$9.87	32.9%
6	78.3	$22.51	$7.49	25.0%
9	69.3	$25.17	$4.83	16.1%
12	61.3	$28.14	$1.86	6.2%

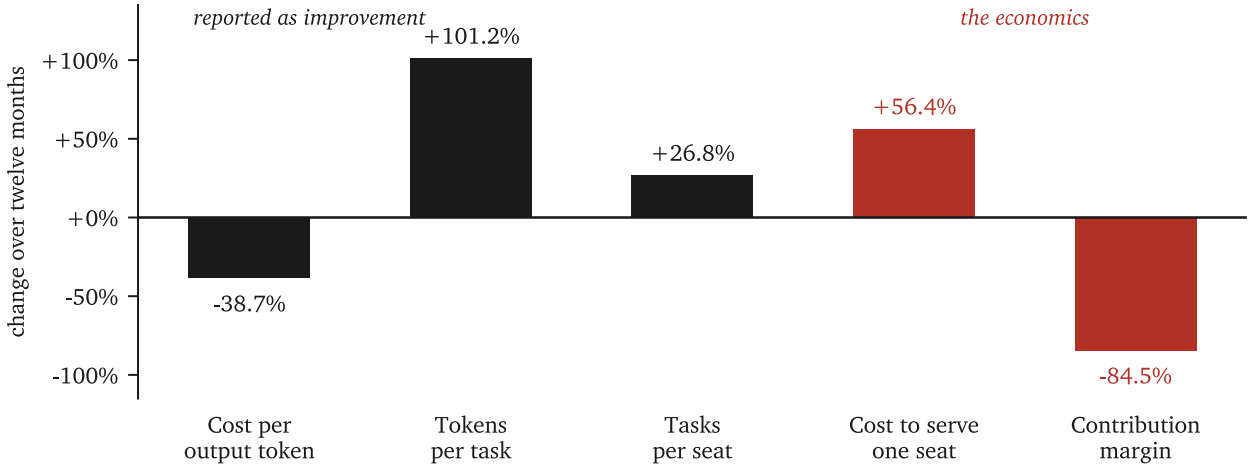


Figure 3: Twelve-month change by instrument. The three quantities an organization reports all move the approved direction: cost per token falls 38.7 percent, while tokens per task rise 101.2 percent and tasks per seat 26.8 percent, both of which read as engagement on an agentic product. The composite of the three rises 56.4 percent and contribution margin falls 84.5 percent. Absolute values appear in Table 1.

Nothing here is a failure of execution. The efficiency work succeeded. A 4 percent monthly gain was set against 8.1 percent monthly growth in consumption, and the composite compounds at 3.795 percent a month in the wrong direction. Margin crosses zero in month 13.7.

3.3 The take-or-pay corollary

Under a fixed capacity commitment the cash version of the identity is harsher. Spend S over the window is contractually fixed, so cost per delivered task is

$$C = S / W \tag{4}$$

where W is compliant work delivered. Improving c does not reduce S. It raises the ceiling on W.

> **Claim 2.** Because S is contractually fixed and W is unmeasured, the reported cost-per-token series for a committed fleet is not a cost series. It is an allocation of a fixed number across a varying divisor, and it improves automatically as volume grows regardless of whether anything got cheaper.

There are exactly two ways freed capacity becomes cash. It is filled with additional compliant work, which requires demand clearing the relevant service-level class, or it is resold. The second route now has venues: cash-settled rental index futures are listed (CME Group and Silicon Data, 2026) and secondary capacity marketplaces exist. Both are denominated in chip-hours, so a buyer can hedge the price of capacity but not the rate at which capacity converts into delivered work. That residual is the exposure this paper is about, and it is unhedgeable because it is unmeasured.

4 Method: the unit

A service-level objective is the latency contract under which output has value, stated as percentile bounds on time-to-first-token and time-per-output-token or their task-level equivalents, over a declared window. Compliant work is

output passing the quality floor and meeting every term of the objective. Output failing either is not discounted work; it is zero work that was paid for. Retried requests count their full spend and only their final compliant output.

4.1 The unit. The cost of useful work over a window is $C = S/W$, equation (4), where S is total spend attributable to the workload in the window including failed and retried requests. W is denominated in compliant tasks: one request-with-defined-acceptable output that passed its quality floor and met every term of its objective. Units: dollars per compliant task. This is the primary form and the only form comparable across model artifacts, because the number of output tokens a model emits to complete a task is chosen by the seller.

The served token, SVT, is the sub-case. Where the artifact is held fixed, meaning same weights, version and quantization and therefore same tokenizer, output token counts become a stable measure of delivered quantity and W may be denominated in millions of compliant output tokens. One SVT is one output token delivered inside its service-level objective and above its quality floor.

4.2 The conditioning tuple. A value of C is non-conforming and undefined unless stated with all six of: the workload, meaning the trace or trace-class measured; the model artifact, meaning weights, version, quantization and tokenizer explicitly; the placement, meaning provider or self-host, silicon, region, batching policy and offered load; the service-level objective, with percentile bounds stated numerically; the quality floor, as an operational acceptance test with the evaluated fraction; and the window, with start, end and sustained load, reported at declared percentiles.

The window must be long relative to the system's own response time and stationary within itself. A conforming record reports window length divided by measured p99 request latency, and a split-half comparison of C over the first and second halves. Where the halves differ by more than ten percent, the window is non-stationary, C describes a transient rather than a placement, and the record is marked accordingly.

4.3 Commensurability. Token-denominated cost is defined within one artifact class and undefined across artifact classes. Any comparison spanning artifacts must be denominated in compliant tasks. A ranking of dollars per million served tokens across different models is not a conservative approximation; it is the error this specification exists to correct. The rule also closes an incentive defect: under a token denominator a model that pads its answer lowers its apparent cost, while under a task denominator padding raises S and leaves W unchanged.

	Cost	Work	Service level	Quality
Dollars per million tokens	●	●	⊘	⊘
Dollars per GPU-hour	●	⊘	⊘	⊘
Benchmark score	⊘	●	⊘	●
Goodput	⊘	●	●	⊘
MLPerf Inference, Server scenario	⊘	●	●	●
Served token	●	●	●	●

A filled mark carries the leg. Every existing metric is the served token with at least one leg removed.

Figure 4: What each candidate metric measures. Raw dollars per million tokens is C with the objective and quality floor deleted. Dollars per GPU-hour is C with W deleted entirely. Benchmark scores are W with S deleted. Goodput is C with S and the quality floor deleted. MLPerf's Server scenario carries three legs and is the nearest live falsifier: adding a cost denominator to it would take the unit most of the way to existing.

5 What is actually contracted

> **Claim 3.** Across the four largest announced commitment programmes, and within the agreement-level record of the most fully documented of them, the indexation form is capacity, power, chip-count or a fixed dollar schedule in every case. We find no publicly documented agreement indexed to delivered compliant work, and no announcement that declares a service-level class.

Provenance is tiered: P primary (company announcement, regulatory filing, named-executive quote), S credible secondary reporting, E estimate.

Table 2: The four largest announced commitment programmes, their indexation form, and the basis on which their flagship capacity is sited. No programme is indexed to delivered work and no announcement declares a service-level class.

Programme	Announced scale	Denomination	Flagship site and basis	Tier
OpenAI, Stargate and cloud	about $1.4T stated total; Oracle $300B over five years from 2027; Broadcom 10GW custom accelerators; AMD 6GW	Fixed dollars, gigawatts, chip-count	Abilene, Texas: land, gas generation and grid access	S
Anthropic, multi-vendor	Up to $517B across at least 14.8GW; AWS up to 5GW; Google and Broadcom multi-GW TPU; Azure $30B	Gigawatts, chip-count, fixed dollars	Project Rainier, Indiana: power and land	P and S
Meta, owned buildout	$64–72B AI data centre capex for the year; Hyperion 5GW compute at full build; Prometheus about 1GW	Gigawatts and capex dollars	Richland Parish, Louisiana: ten new gas plants plus solar	P and S
xAI, owned buildout	Colossus at about 2GW; about 555,000 GPUs for approximately $18B	Gigawatts and chip-count	Memphis, Tennessee: on-site gas turbines	S

The scale figures in Table 2 are announcements rather than cash spent, and several are company statements or journalist-assembled aggregates spanning horizons past 2030. They are reported here to establish the indexation form, which is consistent across all four, rather than to compare programme sizes. For the most fully documented of them, agreement-level detail is available and appears in Table 3. Its aggregate was estimated at up to $517 billion covering at least 14.8GW, roughly 2.9 times what the company had told investors to expect in server rentals through 2029 (The Information, 2026), against a revenue run-rate near $65 billion annualized at the end of July 2026 (TechCrunch, 2026). The aggregate is an order of magnitude, not a liability figure [EST].

Table 3: Agreement-level detail for the most fully documented programme, October 2025 to September 2026. The indexation column is the point: it is capacity, power, chip-count or a fixed schedule in every row.

Counterparty	Announced scale	Denomination	Tier
Amazon / AWS	Up to 5GW; >$100B; about 1GW Trainium2/3 by end 2026	Gigawatts and chip-count	P
Google / Broadcom	Multi-gigawatt TPU from 2027; earlier tranche up to 1M TPUs	Gigawatts and chip-count	P
Microsoft Azure	$30B purchase commitment plus up to 1GW contracted	Fixed dollars and gigawatts	P
Fluidstack	$50B US buildout, Texas and New York	Fixed dollars and sites	P
SpaceX / Colossus	about $1.25B per month to May 2029; about 300MW; about 220k GPUs	Fixed monthly dollars	P via S
Nscale	about $45B over 6 years; 460MW	Megawatts	S
Lambda	about $35B, reported take-or-pay	Take-or-pay dollars	S
Volta Infrastructure	about $10B over 6 years; $1.3B standby letter of credit	Fixed dollars	S

Why suppliers write contracts this way. Take-or-pay revenue is what makes neocloud debt financeable. The progression is visible in one supplier's cost of capital: an initial 2023 GPU-backed facility priced in the mid-teens all-in, and by March 2026 an $8.5 billion delayed-draw term loan rated A3 and A(low), priced at SOFR plus 225 basis points on the floating tranche (CoreWeave, 2026). What earned the investment grade was neither the neocloud's own credit nor GPU residual value. It was contracted revenue from a creditworthy offtaker. This locates the buyer's leverage: the buyer supplies the credit that makes the financing work and currently receives no measurement concession for it.

5.1 Siting is disclosed by power, not by service level

Every flagship site in Table 2 is placed where power could be assembled quickly: land and gas generation in Abilene, ten new gas plants and solar in Richland Parish, on-site turbines in Memphis. That is the correct basis for training

capacity, because a gradient step has no latency contract and a training cluster can sit anywhere power is available. It is an undetermined basis for serving capacity, and the announcements do not resolve the question because none of them declares a service-level class.

This matters to the denominator in a specific way. Compliance under Section 4 includes the latency contract, and latency has a geographic component that capacity cannot overcome. A megawatt outside the latency radius of its demand produces zero compliant work for an interactive class however large the site or however cheap the power, and under take-or-pay it is paid for in full regardless. The same arithmetic as Claim 2 applies, with distance rather than utilization as the cause.

We make the narrow claim only. We do not assert that any programme has mis-sited its capacity, which would require knowing the intended workload mix and the declared classes, and neither is public. We assert that from the public record it is not possible to determine which service-level classes any announced gigawatt can serve, because the class is never stated. That is Claim 3 on a second axis: the commitments disclose the input and not the delivered work, and here the missing disclosure is geographic rather than financial. A buyer or a lender reading these announcements can price the power and cannot price the service.

The off-balance-sheet layer. A private-credit package of approximately $35 billion, arranged to purchase TPUs through a vehicle that leases them to the lab, is the denomination problem crystallized into credit (Bloomberg, 2026) [S]. Repayment depends on lease payments, a residual-value guarantee on the senior stack, and a backstop from the chip vendor. None of it is contingent on tokens delivered.

6 The outside view

> **Claim 5.** For one fixed model artifact served by many providers, the set of cheapest providers by price and the set of fastest providers by delivered speed are disjoint, so the two rankings do not merely differ in order; they share no members at the top.

This section contains the only information a lab cannot generate internally. A lab measures its own fleet with better instrumentation than any outsider will have. It cannot measure the serving economics of providers it competes with, and could not publish the result if it did.

Table 4: The inversion on one fixed artifact. Leaders by price, output speed and time to first token across fourteen benchmarked endpoints serving Llama 3.3 Instruct 70B (Artificial Analysis, 2026) [MEASURED, THIRD PARTY]. No provider appears in more than one column.

Rank	Cheapest by blended price	Fastest by output speed	Lowest time to first token
1	DeepInfra Turbo FP8, $0.12 / 1M	Groq, 292.4 t/s	Google Vertex, 0.65s
2	Nebius Base, $0.16 / 1M	SambaNova, 290.7 t/s	Groq, 0.82s
3	Novita, $0.16 / 1M	Google Vertex, 153.3 t/s	CoreWeave, 0.89s
Spread	Up to 8.6x, to $1.05 / 1M	45.8x fastest to slowest	Not reported across the set

Three things follow, none visible in a price comparison. First, under an interactive bound the cheapest endpoints deliver no compliant work at all: output speed across the set varies by a factor of 45.8, a class requiring sustained interactive generation excludes the slow tail entirely, and for an excluded endpoint W is zero, so C is undefined rather than large. Price rankings treat those endpoints as the winners. Second, the cheapest endpoint in the set is FP8-quantized, so two buyers comparing prices are not entitled to assume the same quality floor until it is declared and tested, and under current market convention they are never asked to declare it. Third, cross-artifact token comparisons are invalid rather than imprecise, per Section 4.3.

The figures in Table 4 are a third-party measurement at one date and will move. The structural finding does not depend on which providers occupy the rows: it depends on the leaderboards being disjoint, which follows from price and latency being purchased from different parts of the cost structure.

7 Experiments

7.1 Setup

The crossover condition of Section 3 is a statement about a region of parameter space, not about one point, so we sweep it. Cost per token declines at monthly rate a over 0 to 10 percent; consumption per seat grows at combined monthly rate b over 0 to 15 percent, where b aggregates tokens per task and tasks per seat. Seat price and starting

cost to serve are held at the Table 1 values, $30 and $18. The reported quantity is the month at which contribution margin reaches zero, $\ln(R/K_0) / \ln[(1 - a)(1 + b)]$, which is undefined where the composite does not exceed 1. Both arms of every comparison use the same grid, the same seat economics, and the same expression, so the only structural difference is the one under test.

7.2 Sensitivity of the crossover

Figure 5, left panel, contours the crossover month over the plane. The binding region is large. At the 4 percent monthly efficiency gain of the worked example, consumption growth above 4.2 percent a month brings margin to zero eventually, above 5.1 percent within five years, and above 8.0 percent within fourteen months. The safe region is the wedge below the diagonal where efficiency outpaces consumption, and it narrows as consumption growth rises regardless of how good the serving organization is: even a 10 percent monthly decline in cost per token, which no published series approaches, is overtaken by 12 percent monthly consumption growth.

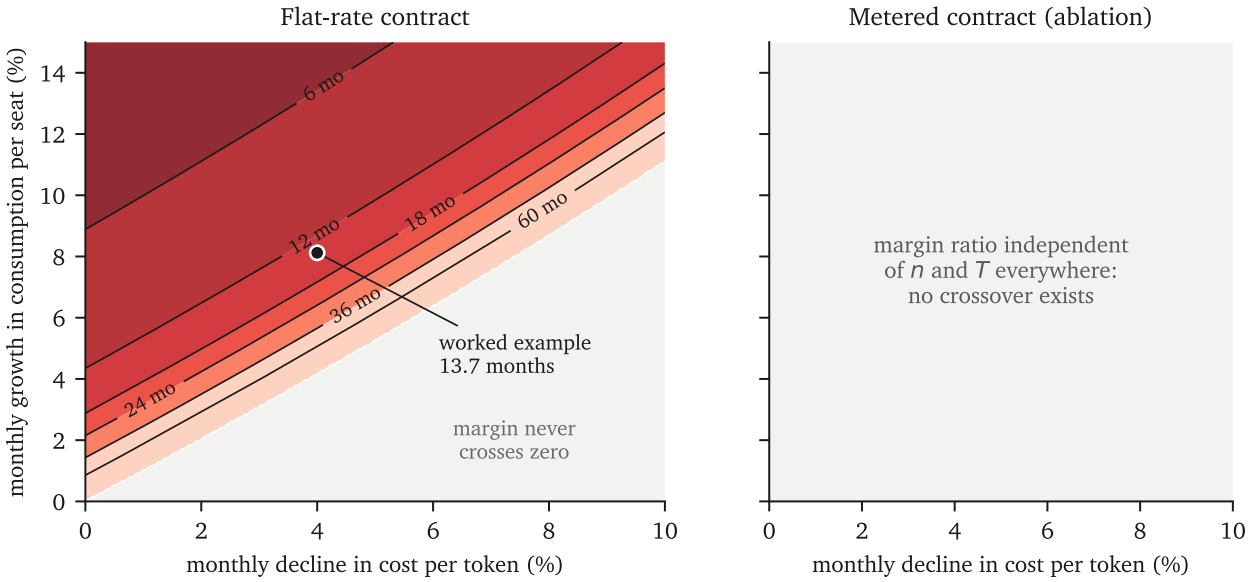


Figure 5: Sensitivity sweep and the metered-contract ablation. Left: month at which contribution margin reaches zero, over the plane of monthly efficiency gain against monthly consumption growth, with darker shading marking a faster crossover. The marked point is the worked example of Table 1. Grey marks the region where margin never crosses zero. Right: the identical sweep under a metered contract, where margin ratio is independent of n and T and no crossover exists at any parameter value.

7.3 Ablation: the price form, not the parameters

A natural objection to Figure 5 is that the result is an artifact of aggressive growth assumptions rather than of anything structural. To test this we re-run the identical sweep on a metered contract, holding every parameter, the grid, and the seat economics fixed and changing only the price form. Revenue per task becomes $p{\cdot}n$ rather than a fixed R, so margin ratio is $(p - c)/p$ and both n and T cancel. The crossover does not occur anywhere on the plane, at any growth rate, however extreme. The right panel of Figure 5 is therefore empty by construction, which is the result: the exposure is created by the price form and not by the growth rates, so an organization cannot manage it by forecasting consumption more carefully.

7.4 Ablation: the caching term saturates

The second objection is that caching supplies unlimited efficiency, so the right-hand side of condition (2) can always be made large enough. It cannot, and the reason is a bound rather than an observation. From equation (3), the further reduction in blended cost available from composition alone, starting at hit rate h, is $1 - \kappa/(\kappa h + 1 - h)$. It is bounded above by $1 - \kappa$ for every h, and falls steeply as h approaches 1. Offsetting one year of consumption growth at the worked example's 8.1 percent a month requires a 60.8 percent reduction, which composition can supply up to $h =$ 82.8 percent and cannot supply above it.

This holds at any hit rate and depends on no estimate of anyone's. What an estimate supplies is only the question of where current practice sits relative to the crossing, and the available figure, above 90 percent on agentic workloads

(SemiAnalysis, 2026), is tier E from a single analyst reconstruction. We therefore state the result as conditional: a serving organization operating above $h = 82.8$ percent has already exhausted composition as a source of offsetting efficiency, and one operating below it has not. The falsifier is a first-party disclosure placing agentic hit rates below that threshold, which would leave the bound intact and only move where practice sits on it.

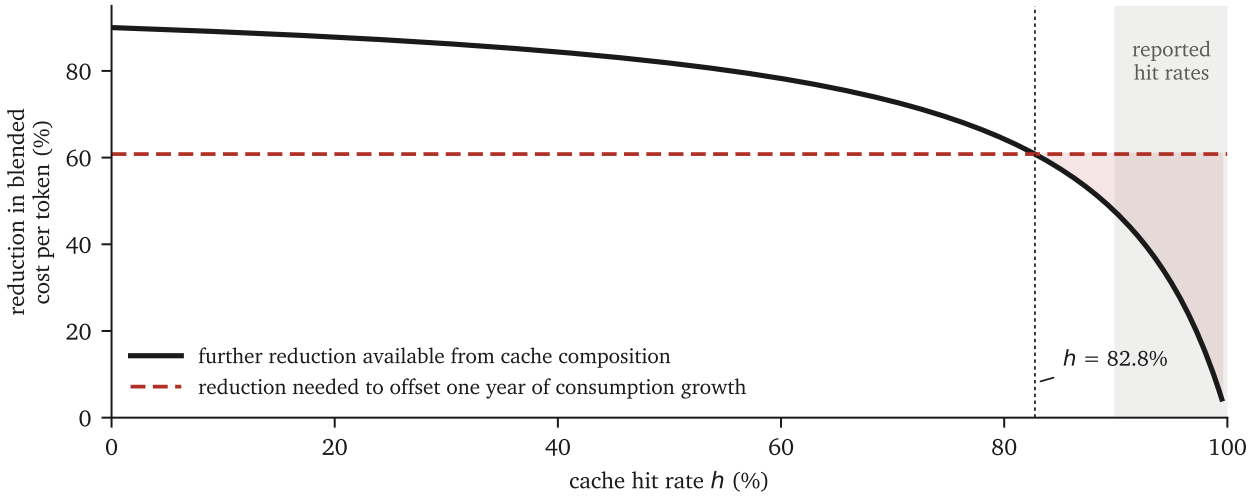


Figure 6: The caching ablation. Further reduction in blended cost per token available from cache composition, against cache hit rate, with the reduction needed to offset one year of consumption growth at the Table 1 parameters. The two cross at $h = 82.8$ percent. Above that hit rate, composition alone can no longer fund a year of growth, and the shaded band is the shortfall. The efficiency term is bounded; the consumption term is not.

7.5 Results

Table 5: Experiment summary. Arrows mark the direction that favours the serving organization. Every value derives from the parameters of Appendix B.1.

Quantity	Flat-rate	Metered (ablation)
Crossover month at the worked example ↑	13.7	**never**
Share of swept plane where margin crosses inside 5 years ↓	58.3%	**0%**
Consumption growth tolerated at 4%/mo efficiency ↑	4.2%/mo	**unbounded**
Cost-per-token reduction over 12 months ↑	38.7%	38.7%
Contribution margin change over 12 months ↑	−84.5%	**0%**
Cache hit rate above which composition cannot fund one year ↑	82.8%	not binding

The headline is the fourth row read against the fifth. Under both price forms the serving organization delivers the same 38.7 percent efficiency gain and every instrument records it identically. Only the flat-rate arm loses 84.5 percent of its contribution margin while doing so.

At scale. Contribution per dollar of flat-rate revenue falls from 40.0 cents to 6.2 cents over the twelve months, a loss of 33.8 cents per revenue dollar. Against a $65 billion annualized run-rate (TechCrunch, 2026), every ten percentage points of revenue carried on flat-rate terms under these parameters is approximately $2.2 billion of annual contribution. The parameters are illustrative [EST] and no firm's flat-rate revenue share is public, so this is a statement of scale rather than an estimate of any firm's exposure. It is included because a margin percentage and a dollar figure are acted on by different people.

8 Contract design

What a unit of capacity is worth internally and what it should be indexed to externally are the same question asked from two sides. This section takes the internal side first.

8.1 The shadow price of SLO-constrained capacity

> **Claim 4.** In a heterogeneous fleet under take-or-pay, the internally correct price of a unit of capacity depends on whether that unit is latency-constrained, and the difference is a dual variable that no accounting system currently produces.

Allocation across training, reinforcement learning, internal use and customer serving is a daily decision, run across three silicon families used fungibly, inside what the chief financial officer has described as a cone of uncertainty (Rao, 2026) [S]. What the meeting lacks is a price.

Inference has two phases with opposite bottlenecks. Prefill is compute-bound and scales with prompt length. Decode is memory-bandwidth-bound and scales with generated tokens and batch size. A fleet sized for one is structurally mispriced for the other, which is why phase disaggregation onto distinct pools became the reference design (Zhong et al., 2024). Hardware validation on H100 PCIe shows a prefill fixed floor near 61 milliseconds that is independent of prompt length and does not appear on L40S. A floor of that size bounds which service-level classes that silicon can serve at all. Silicon that cannot meet a class is worth zero for that class and full price for another, which is a capability constraint rather than a price difference.

Write allocation as a constrained program: maximize the value of delivered work subject to fixed committed capacity by silicon type, a latency constraint per service-level class, and a quality floor per workload. The multiplier on the latency-constrained capacity constraint is the shadow price of SLO-bound capacity.

1. **The shadow price, not the invoice, should govern internal allocation.** Under take-or-pay the invoice is a sunk constant carrying no allocative information.
2. **Batch and interactive work consume different constraints.** Batch work consumes capacity but not the latency constraint, which is why a flat 50 percent batch discount and a comparable spot-versus-reserved spread can both be sustainable. A single blended cost per token prices them identically and therefore misprices both.
3. **Heterogeneity carries an option value no current metric prices.** Moving a workload across TPU, Trainium and GPU is a real option on relative capacity scarcity, and the fungibility took multiple years to build (Rao, 2026). Its value is the spread in shadow prices across the three fleets, computable only once shadow prices exist.

This claim is unquantified, deliberately. We assert that the spread exists and that no accounting system produces it. We do not state its size, because doing so requires measured cells on at least two silicon families under a common declared service-level class, and we hold measurements on one. The experiment that would settle it is stated in Section 10.4. A reader should treat Claim 4 as a structural claim with an open magnitude rather than as a result.

8.2 Indexation and where risk sits

> **Claim 6.** Indexation determines who holds utilization risk and who holds obsolescence risk, and the current form assigns both to the buyer at a price that is not separately quoted.

Table 6: Indexation forms and where each places risk. Only the last row is proposed here; the rest are observed.

Indexation form	Utilization risk	Obsolescence risk	Financeability	In use
Take-or-pay capacity	Buyer	Buyer	High, investment grade achievable	Dominant
Chip-hours, metered	Buyer partly	Buyer	Moderate	Spot and short term
Capacity with residual-value guarantee	Buyer	Shared or seller	High	SPV structures
Delivered compliant work	Seller	Seller	Low as sole form, untested	None documented
Take-or-pay floor plus delivered-work settlement	Shared	Shared	Retains the financeable floor	Proposed here

A supplier whose revenue varies with the buyer's realized compliant output cannot pledge that revenue against term debt on competitive terms, so removing the fixed floor removes the financing and the supplier will decline. This pa-

per does not propose it. The hybrid keeps the take-or-pay floor at the level supporting the supplier's debt and layers a settlement mechanism above it, denominated in delivered compliant work per unit of committed capacity, measured against a declared service-level class and quality floor.

8.3 Six clauses that would have to exist

Declared service-level classes, with percentile bounds stated numerically per class, the analogue of standard voltages and the clause most likely to require a multi-party body. **A declared quality floor**, as an operational acceptance test with the evaluated fraction stated, without which delivered work is undefined and the clause unenforceable. **An attribution basis**, reported on both reserved share and token share with the contract naming which governs. **Window stationarity**, per Section 4.2. **Commensurability**, per Section 4.3. And **measurement custody**: who runs the harness, whether it is open, and what happens on dispute, where the credible form is that the party being measured does not commission its own measurement.

9 Related work

Service-level-conditioned throughput, without price. Goodput, the count of completed requests per second adhering to declared objectives on p90 time-to-first-token and time-per-output-token, is the working vocabulary of serious serving systems and the denominator phase disaggregation optimizes against (Zhong et al., 2024). It is rigorous, load-swept, and carries no price and no acceptance test, so it can rank two schedulers and cannot rank two purchases.

Service level and quality, without cost. MLPerf Inference's Server scenario enforces percentile bounds on both quantities at declared accuracy floors, under consortium governance (MLCommons, 2025). It has no cost denominator in any scenario and runs vendor-submitted hardware against fixed datasets rather than production traffic.

A rate per functional unit, for carbon. Software Carbon Intensity supplies exactly the form this paper argues for, in a different quantity (ISO/IEC, 2024), which is why the energy axis is deferred rather than invented in Appendix A.

Efficiency figures of merit. Intelligence per Watt ranks systems on task accuracy per unit power (Saad-Falcon et al., 2025), and Token Arena fuses price, speed, context and live-endpoint quality into dollars and joules per correct answer, finding up to 6.2x variation in joules per correct answer for the same model across endpoints (Token Arena, 2026). Both rank and neither is additive, so neither can be invoiced or summed across a fleet. The distinction between a figure of merit and a unit of account is the distinction this paper turns on.

Price indices and standards bodies. Cash-settled rental index futures give the market a hedge on the price of capacity (CME Group and Silicon Data, 2026), denominated in chip-hours. The Tokenomics Foundation was chartered to write open standards for AI cost measurement and to extend billing schemas with token telemetry (Linux Foundation, 2026). These are the two places a delivered-work denominator would have to land, and both currently measure raw counts.

10 Discussion

10.1 The governance seam

> **Claim 7.** Every binding capability threshold in force is denominated in training compute, while an increasing share of frontier capability is produced at inference time, and no accepted denominator exists for the latter.

The thresholds are training-side without exception. The EU AI Act uses 10^{25} floating-point operations for systemic-risk classification; the United States executive order used 10^{26}; state statutes have adopted the same figure. Each measures the compute used to create an artifact. None measures the compute used to operate it. That was defensible when capability was fixed at training time. Reasoning models convert serving compute directly into task performance, continuously, on the same weights, so a model below a training threshold can be operated above the capability level the threshold was written to catch (Beyond Training, 2026).

Both the financial and the governance failure have one cause: the accepted unit measures a committed input rather than delivered work. This is a seam, not a proposal, and treating it as more would overstate what the paper establishes.

10.2 Objections and falsifiers

The metrics already exist. Largely true, leg by leg, as Figure 4 and Section 9 set out. What none of them does is settle a contract.

Labs already have this internally. Possible and not publicly determinable. The public record shows analyst reconstructions, not first-party disclosures. The contract-level claim does not depend on the answer: whatever exists internally, the supply agreements remain capacity-indexed.

Suppliers will refuse delivered-work terms. Correct as to pure delivered-work indexation, which is why Section 8.2 does not propose it. Whether suppliers accept a settlement layer above a preserved floor is an empirical matter to be tested rather than argued.

Dollars per million tokens is good enough. A quantity that moves when the vendor changes its tokenizer, with no change in the underlying work, is a price list rather than a unit. A kilowatt-hour is defined by physics and does not change when the utility upgrades its meters.

What would falsify this paper, in descending order of how fast it would resolve. The crossover condition is computed against real telemetry and does not bind in any period, which is the only falsifier that cannot be checked from outside. MLCommons adds a cost denominator to the Server scenario, or the Tokenomics Foundation ships token telemetry carrying a service-level condition and a quality floor rather than raw counts. A supplier signs and discloses an offtake agreement with a delivered-work settlement layer. A frontier lab discloses an internal unit of account fusing cost, service level and quality. Or the leaderboards of Table 4 converge, so that the cheapest providers are also the fastest.

10.3 Predictions

Prediction 3 is the public shadow of Claim 1: a vendor that introduces caps on a flat-rate product is a vendor whose crossover already bound. It resolves against a named watch list, so that it can be checked rather than argued: per-seat coding assistants and chat subscriptions sold with unlimited or high-limit usage. The observable is a pricing-page or changelog diff introducing caps, rate limits or consumption tiers, together with a stated cost-to-serve rationale. Both artifacts are public and dated, so the prediction resolves without access to any firm's internals. Prediction 4 is included against the paper's own direction, since if a delivered-work index displaces rental indexation quickly, the paper understated how fast the market moves.

Table 7: Dated predictions, each resolvable from public information. Prediction 4 is stated against the paper's own direction.

#	Prediction	Resolves by
1	A standards body publishes a cost denominator conditioned on a service-level objective	End 2027
2	A large compute offtake is disclosed with a settlement term indexed to delivered work above a take-or-pay floor	End 2028
3	A major vendor introduces usage caps, rate limits or consumption tiering on a previously unlimited flat-rate AI product, citing cost to serve	End 2027
4	Cash-settled compute futures trade against a rental index rather than a delivered-work index for their first full year	End 2027
5	A rating agency publishes methodology treating contracted compute utilization, not GPU residual value, as the primary credit driver	End 2027
6	A binding capability threshold denominated in inference-time compute is proposed in draft legislation or consultation	End 2028
7	The price and output-speed leaderboards for a single widely served open-weight artifact remain disjoint at the top three	End 2027

10.4 Limits and what would settle them

The experiments narrow what the paper claims. Section 7.3 shows the exposure is a property of the price form rather than of any growth forecast, so it cannot be managed by forecasting consumption more carefully, only by changing the contract or the instrument. Section 7.4 shows the largest recent source of per-token efficiency is bounded and, at reported hit rates, close to exhausted, so the right-hand side of condition (2) cannot be relied on to keep pace. Together they move the central result from a possibility to a structural property of flat-rate serving at current parameters.

Three limits are worth stating plainly. The sweep is parametric rather than empirical: it shows where the crossover binds given a growth pair, not what any particular organization's growth pair is, and only internal telemetry can sup-

ply that. The unit is specified for serving and not for training, which on the agreements of Table 2 is the larger portion of committed capacity; Appendix A states what a training denominator would need instead. And Table 4 is a third-party measurement at one date, so the providers in its rows will change even though the disjointness result is structural.

A fourth limit is predictive, and it is the one that most constrains what the unit is worth at signature. This paper measures delivered work after the fact. The question a buyer faces before committing capacity is what compliant work a given placement will deliver, conditional on workload shape, silicon, software stack and service-level class. Answering it requires a model of performance under those conditions, validated against measured cells on more than one silicon family, which is also what Claim 4 needs and what we do not hold. We state it as an open problem rather than sketch a solution, and note only that a denominator defined after the fact and a predictor available before it are different artifacts with different evidentiary standards.

The same measurement that settles a contract also serves the governance seam of Section 10.1, because a unit that is countable, additive and conditioned on a stated tuple is auditable in the way a threshold requires. That is a reason to expect the denominator to arrive from a body with no position in the answer rather than from a participant, which is what happened in every prior infrastructure that eventually settled on one.

11 Conclusion

A firm that commits hundreds of billions of dollars in units of capacity, and earns revenue in units the model chooses, is running two books that cannot be reconciled, and the reconciliation requires a measurement that does not yet exist anywhere. The crossover of Section 3 is computable this quarter from series that already exist in separate systems; the join is the work, and it either finds the exposure or rules it out. The settlement layer of Section 8.2 requires one supplier to accept a measurement term above a take-or-pay floor it already has, which is a negotiation rather than a research programme. Reference service-level classes require a multi-party body, and standards ossify around whatever the first implementations measured, which today is raw token counts.

A. Scope and attribution

A.1 Multi-tenant attribution. Two bases are defensible and they bracket the honest answer. Under reserved share, spend is attributed in proportion to reserved capacity, charging idle reserved capacity to the tenant that reserved it. Under token share, spend is attributed in proportion to realized compliant output, ignoring idle capacity, which is a strict lower bound. A conforming record reports C on both bases and names which it uses. The gap between them is the tenant's own load factor restated in dollars.

A.2 Training capacity is out of scope, and this is a real limitation. Training work has no service-level objective in the sense of Section 4: there is no latency contract under which a gradient step has value, and no per-request acceptance test. W is therefore undefined over the training portion of a committed fleet, which on the programmes in Table 2 is the larger portion. Three consequences follow. $C = S/W$ applies to the serving slice only, so any fleet-wide figure requires an attribution of S between slices, which A.1 governs but does not determine. The shadow-price program of Section 8.1 still forms, because training enters it as a competing claim on capacity valued by a different objective rather than as a term in this unit. And a delivered-work denominator for training is a separate specification, likely denominated in progress toward a declared checkpoint target; nothing here supplies it.

A.3 What the unit excludes, knowingly. Quality floors richer than binary acceptance, since graded usefulness resists standardization and this specification chooses auditability over resolution. The price basis, list, committed or spot, which the tuple discloses but does not normalize. An energy term, deferred to keep the unit's first version one-dimensional and because ISO/IEC 21031 already supplies its form. And reference service-level classes, which only a multi-party body can legitimately set.

B. Reproduction

B.1 Table 1 and Figures 1, 3, 5 and 6. All are generated by five parameters: starting cost to serve 18, seat price 30, and monthly growth rates of −0.04 for c, +0.06 for n, +0.02 for T, with cached-read price ratio $\kappa = 0.1$ for Figure 6. The composite is the product of the three growth factors, 1.03795. Cost to serve at month t is 18×1.03795^t; the cost-per-token index is 100×0.96^t; margin is 30 minus cost to serve. Zero-margin crossing is $\ln(30/18) / \ln(1.03795) = 13.71$. No figure in this paper's own results comes from anywhere else, and the table and the figures are generated from the same expression rather than transcribed separately.

B.2 Recomputing the crossover against real telemetry. Three monthly series over at least four quarters, for one flat-rate product line. c, cost per output token served, already produced by serving infrastructure; decompose it per

equation (3) into the uncached rate and the hit rate, or the composition effect will be mistaken for a hardware gain. n, output tokens per completed task, which requires a task boundary definition; for an agentic product the natural boundary is a session or completed job, not a request. T, completed tasks per active seat per month, already produced by product analytics. Compute $K = c{\cdot}n{\cdot}T$ against revenue per seat and test condition (2) in each month. The join across three sources is the work; each series already exists. The expected finding, stated in advance so it is not fitted after the fact, is that the composite moved against the firm in at least one quarter in which every individual instrument improved.

B.3 Recomputing the committed-fleet cost series. For the take-or-pay portion, hold S at the contracted schedule rather than an imputed usage-based cost, and compute $C = S/W$ with W per Section 4 and the training slice excluded per A.2. Compare against the reported cost-per-token series for the same period. The divergence measures how much of the reported improvement is volume growth against a fixed numerator.

B.4 Reproducing Table 4. One fixed open-weight artifact, provider pricing pages and a published throughput series. No model access, no API keys, no proprietary data. Rank by blended price, rank by output speed, and compare membership of the top three. The disjointness result reproduces in a spreadsheet and should be re-run at the date of reading.

B.5 Provenance discipline. Every figure carried into a derived document is tagged at its origin: MEASURED, VALIDATED, SPEC, CONFIG, SIM, EST, HYPOTHESIS, THIRD PARTY, or UNKNOWN. A number that cannot carry a tag does not go in. Where this paper cites hardware figures from the author's own validation, the regime to which each applies is declared with it and was pre-committed before the runs.

References

Anthropic (2025a). Anthropic and Amazon expand compute collaboration.

Anthropic (2025b). Anthropic invests $50 billion in American AI infrastructure.

Anthropic (2026). Anthropic expands Google and Broadcom compute deal.

Artificial Analysis (2026). Llama 3.3 70B: API provider performance benchmarking and price analysis. Accessed September 17, 2026.

Beyond Training (2026). A feasibility taxonomy for inference-time AI governance. arXiv:2609.10105.

Bloomberg (2026). Private-credit package of approximately $35bn arranged by Apollo and Blackstone to purchase TPUs through a leasing vehicle, reported June 5, 2026, with corroborating reporting by Axios, June 10, 2026. No open-access URL held; the reporting is paywalled and is cited on that basis. [S]

CME Group and Silicon Data (2026). GPU rental index futures, listed for October 5, 2026.

Data Center Dynamics (2026). Building Stargate: talking to OpenAI about its trillion-dollar data center vision, July 2026. Reports the approximately $1.4 trillion figure as a statement by OpenAI's chief executive. [S]

Introl (2026). xAI Colossus expansion to 2GW, 555,000 GPUs, January 2026, following Bloomberg reporting of December 30, 2025. [S]

Meta, via TechCrunch (2025) and IEEE Spectrum (2026). Meta is building a 5GW AI data centre; Hyperion siting, phased generation and cooling load per IEEE Spectrum, March 2026. [P and S]

CoreWeave (2026). CoreWeave closes landmark $8.5 billion financing facility, achieving first investment-grade rated GPU-backed financing. Investor release, March 31, 2026. DDTL 4.0: A3 (Moody's), A (low) (DBRS); floating tranche at SOFR + 2.25%, fixed tranche at approximately 5.9%; matures March 2032; anchored by Blackstone Credit & Insurance. [P]

ISO/IEC (2024). ISO/IEC 21031:2024, Software Carbon Intensity specification.

Linux Foundation (2026). Linux Foundation launches the Tokenomics Foundation, August 4, 2026. Vendor-neutral body for AI cost measurement standards, operating with the FinOps Foundation. [P]

Microsoft (2025). Microsoft, NVIDIA and Anthropic announce strategic partnerships, November 18, 2025.

MLCommons (2025). MLPerf Inference v5.1 results, September 2025.

Rao, K. (2026). Interview, *Invest Like the Best*, EP. 471, May 2026. Quotations taken from a published summary rather than a primary transcript. [S]

Saad-Falcon, J., Narayan, A., et al. (2025). Intelligence per Watt: measuring intelligence efficiency of local AI. arXiv:2511.07885.

SemiAnalysis (2026). InferenceMAX throughput-interactivity frontier, and serving-economics reconstructions including agentic cache hit rates above 90 percent. Analyst estimates, not first-party disclosures; Section 7.4 states the result so that it does not depend on them. [EST]

TechCrunch (2026). Anthropic's annualized revenue surges to $65B, August 17, 2026.

The Information (2026). Estimate of cumulative announced compute commitments, up to $517bn across at least 14.8GW over the eleven months to August 2026, September 6, 2026. Journalist-assembled aggregate spanning horizons past 2030; no open-access URL held. [EST]

Token Arena (2026). A continuous benchmark unifying energy and cognition in AI inference. arXiv:2605.00300.

Zhong, Y., Liu, S., Chen, J., et al. (2024). DistServe: disaggregating prefill and decoding for goodput-optimized large language model serving. arXiv:2401.09670.

Corresponding author: sidney@mindmeetsmatter.com
© 2026 Mind Meets Matter. Every figure is tagged at its origin. Predictions 1 through 7 and the falsifiers of Section 10.2 are the falsification surface of this paper; if they fail, so does it.