Why Enterprises Pay the Frontier Premium

REV 2.1

September 2026 update · five mechanisms behind the closed-vs-open procurement decision · modelled economics

STRUCTURED ABSTRACT

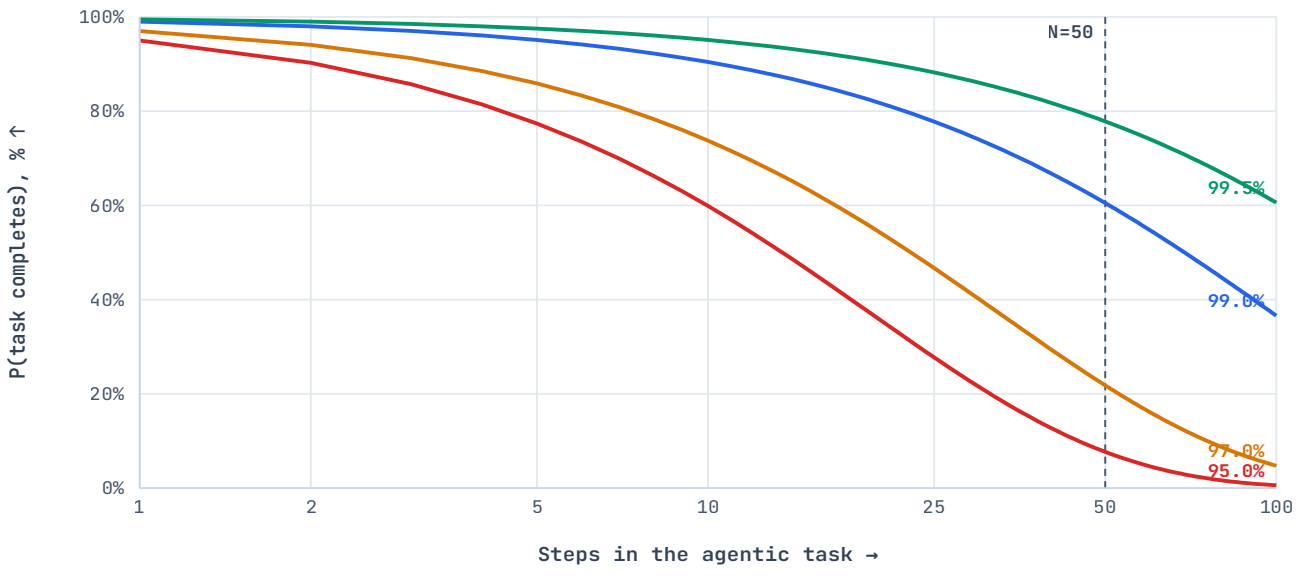
QUESTION	Why do large enterprises pay a substantial premium for frontier proprietary models when open weights reach roughly 87% of composite benchmark capability at a fraction of the token price?
FINDING	Token price is the wrong denominator. Cost per unit of output is (1 + f) / m: model spend enters additively while throughput divides. At realistic spend the frontier needs only a ~4% throughput edge to break even. A two-point per-step reliability edge (97%→99%) clears that bar at any horizon N ≥ 2, and by a widening margin as N grows.
METHOD	Derived arithmetic (p^N compounding; the unit-cost identity and its isoclines) combined with parameterised models for self-hosting TCO and threshold-gated value. Benchmark figures are third-party reported; per-step reliability values are illustrative parameters, not measurements (Note 2).
SCOPE	Applies to knowledge-work and agentic software deployment at enterprise scale. Does not address consumer products, on-device inference, or research settings.

Token price is the wrong denominator. Cost per unit of output is `(1 + f) / m` — model spend is additive, throughput is a divisor. At a revised **2.4–5.8% of loaded labour**, the frontier needs only a **4.0% throughput edge** over open weights to break even — a bar that p^N compounding clears at any horizon N ≥ 2. The genuine overpayment isn't choosing proprietary; it's **routing everything to a flagship max-effort tier**.

1 · Agentic error compounding

Task completion probability vs horizon · P(success) ≈ p^N

The largest single driver, and the one invisible on single-turn benchmarks. A two-point per-step reliability delta rounds away in a composite index and is **decisive** by step 50.



PER-STEP RELIABILITY	10 STEPS	25 STEPS	50 STEPS	100 STEPS
95.0%	59.9%	27.7%	7.7%	0.6%
97.0%	73.7%	46.7%	21.8%	4.8%
99.0%	90.4%	77.8%	60.5%	36.6%
99.5%	95.1%	88.2%	77.8%	60.6%

WHY THIS DOMINATES

Moving per-step reliability from **97% to 99%** is a **2.8× swing in completion at N=50** and **7.7× at N=100** — the boundary between "the agent ships the PR" and "an engineer babysits and re-drives." Read against driver 2: the completion-ratio gain from that edge is **2.1% at N=1, 4.2% at N=2, 22.6% at N=10** — it clears the 4.0% break-even bar from N=2 onward, by a widening margin. The bridge from completion probability to m is **accepted output** (limit 4 in driver 2): a task that fails to complete ships no feature and adds nothing to m.

2 · The labour denominator

REVISED REFRAMED

Unit cost = (1 + f) / m · f = spend as a fraction of loaded labour, m = throughput multiplier

f is additive; m is a divisor. A small numerator penalty buys a large denominator gain, which is why the frontier wins on unit cost while spending several times the tokens.

Reading the formula

1

One seat, one year. Start with what a person costs you.

$250,000 loaded salary

+ f

Add what their AI costs, expressed as a fraction of that salary. f = 0.05 means you are now paying 1.05 salaries for that seat.

f = model spend ÷ salary

÷ m

Divide by how much more work the seat now produces. m = 1.45 means 1.45× the output.

m = throughput multiplier

=

What one unit of work now costs you, relative to before. Below 1.00 means cheaper.

(1 + f) / m

Worked example · one engineer, one year

Illustrative units of completed work — not a measurement

LINE	BASELINE	OPEN WEIGHTS	FRONTIER
Loaded salary	$250,000	$250,000	$250,000
Model spend	$0	$2,500/yr	$12,500/yr
as fraction of salary (f)	0%	1%	5%
Total cost of the seat	$250,000	$252,500	$262,500
Features shipped in the year	40	48	58
throughput multiplier (m)	1.00×	1.20×	1.45×
Cost per feature	**$6,250**	**$5,260**	**$4,526**
Unit cost index · (1 + f) / m	**1.000**	**0.842**	**0.724**

WHAT THE EXAMPLE SHOWS

The frontier seat costs 5% more and produces 45% more, so each feature lands at **72.4% of baseline cost** — and **14.0% below the open-weights seat**. Break-even against open weights needed only m ≥ 1.25 (about 50 features). It delivered 58. Feature counts are illustrative units of completed work, not a measurement.

Why the asymmetry is the whole argument

Elasticity of unit cost with respect to each input

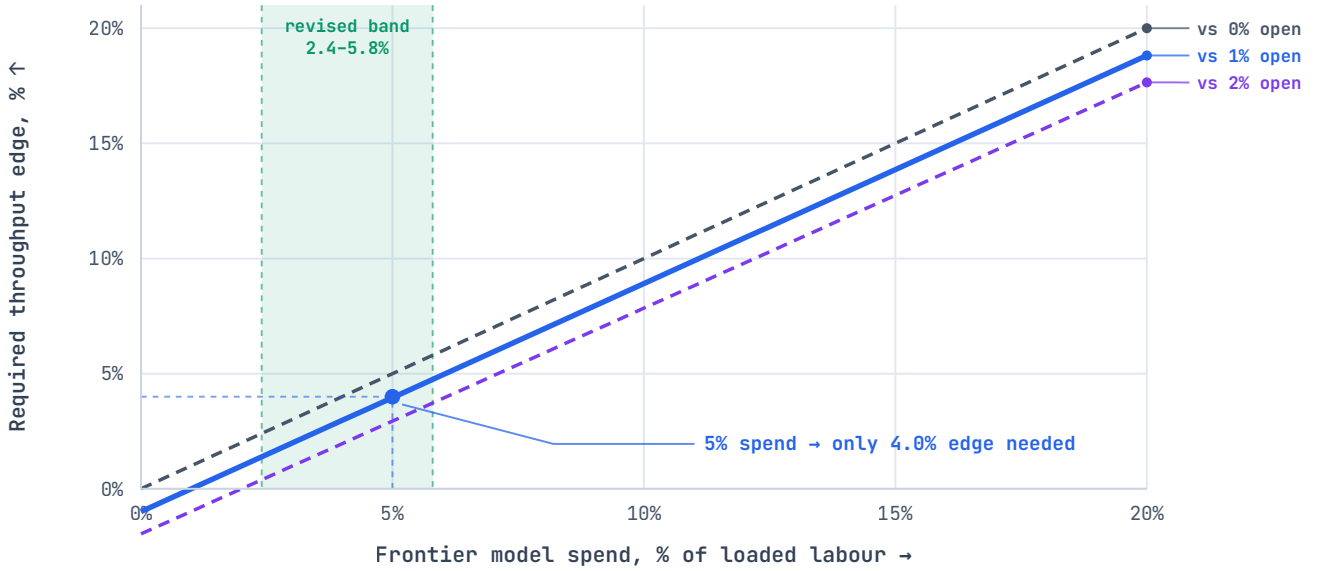
INPUT	ELASTICITY	AT F = 5%	READING
Model spend · f	f / (1 + f)	**+0.048**	Nearly inelastic. Doubling the AI bill raises unit cost ~4.8%.
Throughput · m	-1 (exactly, always)	**-1.000**	Unit-elastic. A 10% throughput gain cuts unit cost 9.1%.

LEVERAGE

Percent change in unit cost per percent change in the input, evaluated at f = 5%. Throughput carries roughly **21× the leverage of spend**. The elasticity in f is self-limiting: spend matters more the more you spend, but only linearly and from a very low base.

Scenario comparison

SCENARIO	F · SPEND	M · THROUGHPUT	UNIT COST	VS BASELINE
Baseline · no tooling	0%	1.00×	**1.000**	—
Open weights	1%	1.20×	**0.842**	-15.8%
Frontier	5%	1.45×	**0.724**	-27.6%
Frontier · heavy agentic	10%	1.55×	**0.710**	-29.0%



BREAK-EVEN AT 5% SPEND	BREAK-EVEN AT 10% SPEND	BREAK-EVEN AT 15% SPEND	FRONTIER UNIT COST
+4.0% edge	**+8.9%** edge	**+13.9%** edge	**0.724** vs 0.842 open

BREAK-EVEN IN CLOSED FORM

Frontier beats open weights when m_F / m_O ≥ (1 + f_F) / (1 + f_O). Rearranged, the required throughput edge is **E = (f_F - f_O) / (1 + f_O)**. Two consequences: **E does not depend on m at all**, so the multipliers used throughout this document set the size of the win but never whether there is one; and for small f_O, **E ≈ Δf** — the extra spend expressed as a percentage of salary. In plain terms: if the frontier model costs four more points of a salary, it has to make that person 4% more productive.

What the identity cannot see

1. **Output must be homogeneous**
   Q is "units of completed work." Where the frontier model does things the alternative cannot do at all, m is undefined — that is a capability discontinuity, and driver 4 governs, not this identity.

2. **m is a workload-weighted average**
   Real m is high on refactoring and boilerplate, low on novel design, and plausibly below 1 where review overhead exceeds generation savings. A composite m hides that spread — the same averaging trap criticised in driver 4.

3. **Fixed costs are omitted**
   Integration, eval harnesses, training, and change management are absent. The fuller form is (1 + f + k) / m, where k is amortised adoption cost as a fraction of salary. In year one k can exceed f outright, which is why pilots underperform the model.

4. **Quality is invisible**
   Two outputs count as one unit each regardless of defect rate. Strictly, m should be accepted output, not generated output. Driver 1 exists because this identity cannot see reliability.

5. **Loaded labour is assumed invariant**
   If AI-augmented engineers command a wage premium, L rises and f falls mechanically — flattering the ratio for the wrong reason.

REVISED SPEND BAND — THE EARLIER 1–2% WAS ANCHORED TOO LOW

A saturated agentic engineer running parallel sessions lands nearer **$500–$1,200/seat/month**, or **2.4–5.8% of $250k loaded labour**, with a fat right tail (MED — inferred from agentic usage patterns, not a surveyed figure). Raising f does not weaken the case: at 5% frontier spend against 1% open weights, break-even is a **4.0% throughput edge**. Even at 15% spend it is only 13.9%. The token-price argument was never load-bearing.

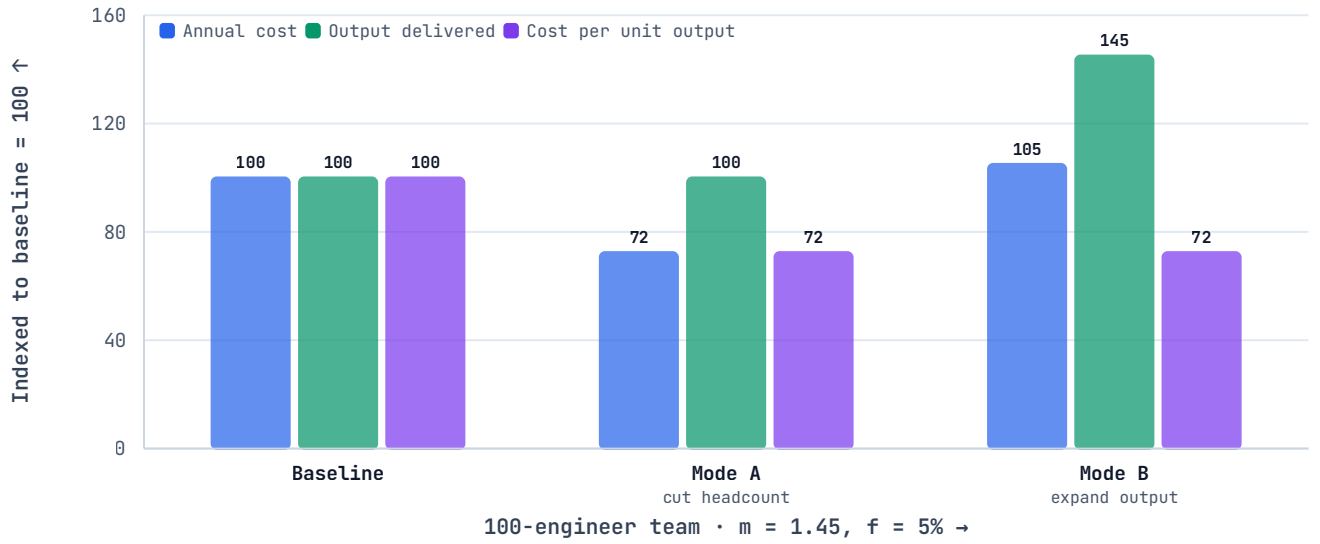
JEVONS DRIFT — AND WHERE THE FRAME BREAKS

As inference gets cheaper per unit of work, consumption rises faster than unit cost falls, so **total spend goes up**. The 2.4–5.8% figure will drift upward, and that drift is a health signal rather than a cost problem. Past some point "tooling as a percentage of labour" stops being the right accounting frame entirely: compute becomes a **factor of production alongside labour** rather than an overhead line charged against it. That is when the procurement conversation changes shape.

3 · Substitution modes NEW

Two ways to bank the same multiplier · identical on unit cost, different on total value

A throughput multiplier can be taken as fewer engineers for the same output, or the same engineers for more output. Illustration: a 100-engineer team at m = 1.45, f = 5%.



MODE A

Reduce headcount

Value is **capped at the labour line** — you recover at most the salaries eliminated, and the gain terminates. Note that **f itself does not move**: N/m engineers at the same per-seat spend scale numerator and denominator together. What shrinks is the absolute AI budget, which can make a high-leverage line item look small and invite under-investment — a perception risk, not arithmetic. **Forced when demand is the binding constraint.**

MODE B

Hold headcount, expand output

Value is bounded by **opportunity set, not by cost.** Unbounded upside where a project backlog exists — additional revenue-generating work rather than a one-time cost recovery. **Available when engineering capacity is the constraint.**

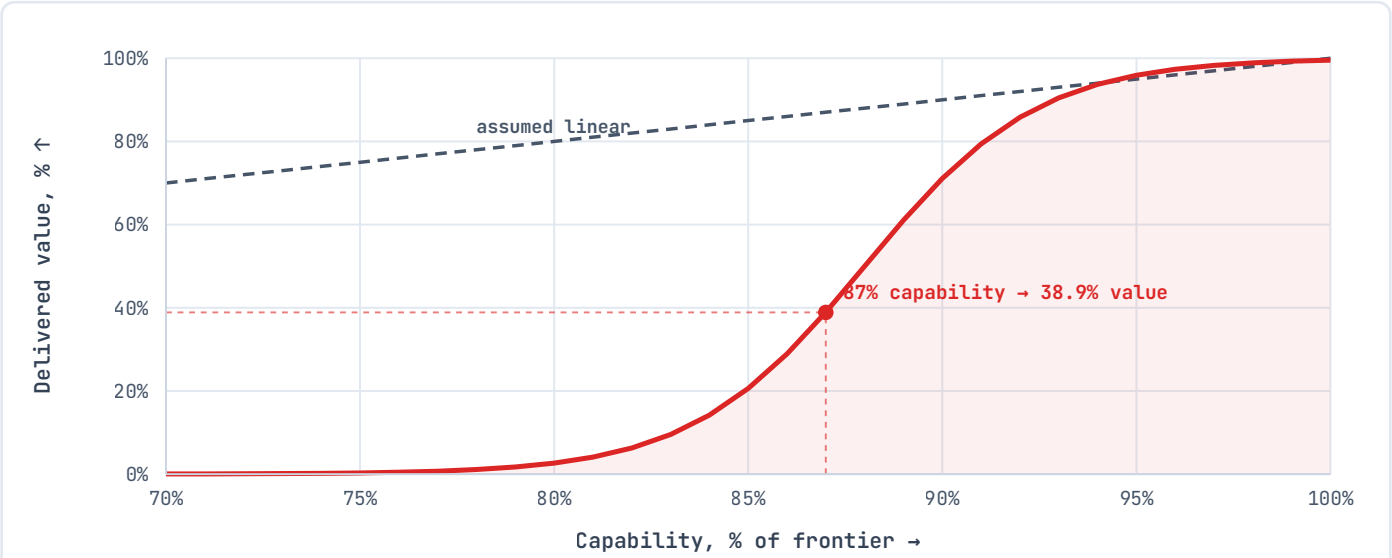
THE CHOICE IS A DIAGNOSTIC, NOT A PREFERENCE

Both modes land at **72.4 on cost per unit of output** — the arithmetic cannot distinguish them. What separates them is which constraint actually binds. Most software organisations claim engineering capacity is their limit, often accurately; firms that reflexively take Mode A are usually revealing that their real constraint was **never headcount**. That is a strategy disclosure, not a procurement decision.

4 · Threshold-gated value

Delivered value vs capability · the composite-average trap, priced

"Open weights reach ~87% of proprietary" assumes linear value transfer. On work gated by a pass threshold **delivered value is sigmoid**, and the last few index points sit on the knee.



READING IT

On this curve **87% of frontier capability delivers ~38.9% of the value**. The benchmark isn't wrong — a mean over ten evals is simply the wrong functional form for threshold-gated work. The easy 80% of tasks were automatable two model generations ago; the residue is exactly what sits on the knee. Illustrative — LOW confidence on curvature, HIGH on direction.

5 · Self-hosting total cost of ownership

Effective cost per 1M output tokens vs fleet utilisation · 4× 8×H200 + 3 ops FTE

Open weights are free; **serving them is not**. Fixed cost amortises over tokens actually produced, so unit cost is governed by utilisation — and enterprise diurnal load runs **5–10× peak-to-trough**.

BREAK-EVEN VS OPEN-WEIGHTS 3RD-PARTY API	BREAK-EVEN VS MID-TIER PROPRIETARY	BREAK-EVEN VS FRONTIER FLAGSHIP	FIXED COST MODELLED
47.9% util	**19.1%** util	**3.8%** util	**$1.81M** /yr

MODEL ASSUMPTIONS — LOAD-BEARING, TAGGED

Infra **$911k/yr** (4 nodes × $26/hr rented — MED) plus ops **$900k/yr** (3 FTE × $300k loaded — MED) = **$1.81M/yr fixed**. Peak capacity 5,000 output tok/s per node for a ~100B-class MoE under heavy batching — LOW confidence; this is the swing variable. Halve throughput and every crossover doubles. The qualitative result survives that: self-hosting beats a flagship price almost immediately and a cheap open-weights API only at high sustained load.

Net read

Eight drivers, with basis and confidence

DRIVER	MECHANISM	BASIS	CONF
Agentic error compounding	97%→99% per-step = 2.8× completion at N=50	Derived (p^N)	HIGH
Labour denominator	f is additive, m is a divisor — break-even edge only ~4%	Derived	HIGH
Substitution symmetry	Cut headcount and expand output share identical unit cost	Derived	HIGH
Threshold-gated value	87% composite capability → ~39% delivered value	Modelled	MED
Self-host TCO	Break-even vs cheap API needs ≈48% sustained utilisation	Modelled	LOW
Indemnity & compliance	IP indemnity, SOC 2 / HIPAA / FedRAMP, residency	Contractual	HIGH
Provenance & sanctions	Training-data lineage, backdoor risk, Entity List exposure	Qualitative	MED
Harness, not checkpoint	Caching, batch, tool-call reliability, agentic scaffolds	Qualitative	HIGH

LANE 1 · OPEN WEIGHTS

Bounded, high-volume, private

Classification, extraction, routing, embedding. Error-tolerant, short-horizon, sustained enough to clear the utilisation bar. **The open-weights case holds cleanly here.**

LANE 2 · MID-TIER PROPRIETARY

General reasoning, good-enough

Where the **dominated-flagship trap** lives. Cost-tuned proprietary out-values open weights on both axes — same index, a fifth the cost.

LANE 3 · FRONTIER

Long-horizon, calibration-critical

Agentic chains, must-not-hallucinate, threshold-gated. p^N and the sigmoid knee are both load-bearing. **Token cost is the small line item.**

NET

Enterprises don't pay the premium because they're bad at arithmetic — they pay it because the arithmetic they're doing has a different denominator, and because f enters additively while m divides. Where open weights genuinely win is a real and sizeable quadrant, but it's **bounded work at sustained load**, not general reasoning. The overpayment that actually exists is **failure to segment traffic**, and it shows up inside the proprietary cohort as often as across the closed/open line.

Sources. Reliability compounding and the unit-cost identity are derived arithmetic. Index and cost figures from Artificial Analysis Intelligence Index v4.0, live June 9 2026. Every modelled parameter, limitation, and exclusion in this document — including throughput multipliers, the spend band, self-hosting assumptions, and omitted commercial terms — is itemised with confidence tags on the **Notes & clarifications** page.

Notes & clarifications

Everything that is a choice rather than a measurement, stated plainly

Change log

Post-review corrections · items 1–4 REV 2.1

Four corrections from an external arithmetic and consistency review, applied to both formats. (1) **Net read** — the withdrawn calibration-inversion driver had re-entered the summary table with a HIGH tag; row removed, driver count corrected to eight, and the hallucination-rate source line dropped since no figure in the body uses it. (2) **Driver 1** — the N=100 completion multiplier for 97%→99% is 7.7× (0.366 / 0.048), not 7.4×. (3) **Abstract, Method** — "reliability figures are third-party reported" contradicted Note 2; now states that per-step reliability values are illustrative parameters. (4) **Abstract, verdict, driver 1** — "clears the bar on any workload of length" replaced with the quantified claim: the two-point edge yields a completion-ratio gain of 2.1% at N=1, 4.2% at N=2, 22.6% at N=10, so it clears the 4.0% break-even from N=2 onward; the bridge from completion probability to m (accepted output, limit 4) is now stated explicitly.

Substitution modes · Mode A second-order claim CORRECTED

The previous edition stated that reducing headcount raises model spend as a percentage of loaded labour "mechanically." That is false. Mode A retains the same per-seat spend and the same per-seat throughput, so headcount N/m scales numerator and denominator together and f is unchanged at S/L. Replaced with the weaker and correct claim: the absolute AI budget shrinks with headcount, which can make a high-leverage line item look small and invite under-investment — a perception risk, not an arithmetic one.

Labour denominator · plain-language walkthrough NEW

Added a term-by-term reading of (1 + f) / m, a worked single-seat example in dollars and features, an elasticity table, the closed-form break-even E = (f_F - f_O) / (1 + f_O), and an explicit statement of the identity's limits.

Labour denominator · spend band REVISED

Model spend raised from 1–2% to 2.4–5.8% of loaded labour ($500–1,200/seat/month). The earlier band was anchored on light interactive use and understated saturated agentic sessions. The revision strengthens rather than weakens the conclusion — break-even rises only to 4.0%.

Labour denominator · presentation REFRAMED

Replaced the break-even bar chart with the (1 + f) / m unit-cost identity and its isoclines. The bar chart answered "what lift repays the spend" against an absolute baseline; the isoclines answer the decision-relevant question, which is the lift required relative to the open-weights alternative.

Substitution modes NEW

Added as a distinct driver. Headcount reduction and output expansion are algebraically identical on cost per unit of output and differ only in where value is booked and whether it is bounded.

Jevons drift NEW

Noted that total spend rises as unit cost falls, and that percentage-of-labour eventually stops being the correct accounting frame.

Prior-revision corrections CARRIED

Corrections from the previous edition remain in force: the provenance-laundered calibration attribution and the invalid hallucination-rate denominator. The calibration-inversion driver remains withdrawn rather than reversed.

Notes

1 **Throughput multipliers are parameters, not measurements** LOW

m = 1.20 / 1.45 / 1.55 are chosen to illustrate the shape of the unit-cost relation. They set the size of the gain, not its direction. Break-even is independent of m entirely — it is (1 + f_frontier) / (1 + f_open) — so the 4.0% figure survives any revision to them. First-party measured throughput would rank above these and should replace them.

2 **p^N is an upper bound on decay, not a forecast** HIGH

The independence assumption overstates compounding: real agentic failures correlate, and recovery, retry, and human checkpoints truncate the chain. The curves describe the shape of the effect and the direction of the reliability premium. Specific completion rates are illustrative.

3 **Spend band is inferred, not surveyed** MED

$500–1,200/seat/month is reasoned from agentic usage patterns rather than measured across a population. The right tail is fat: parallel-session users can exceed it materially.

4 **Self-host TCO is dominated by one assumption** LOW

5,000 sustained output tok/s per node for a ~100B-class MoE governs every crossover in driver 5. Halve it and each crossover doubles. Node rental and loaded FTE cost are market estimates. The qualitative ordering survives; the specific percentages do not.

5 **Threshold curve is illustrative on curvature** MED

The sigmoid's steepness and midpoint are chosen, not fitted. The direction — that value transfer is convex rather than linear on gated work — is the load-bearing claim.

6 **No composite averaging across heterogeneous rows** HIGH

The 87% figure is quoted only as the thing being critiqued. It is a mean over ten evals and is not used here as an input to any calculation.

7 **Reference prices are output-only** MED

API prices exclude input tokens, cache reads, and the verbosity tax on output-heavy models. Folding verbosity in shifts effective frontier cost upward and moves the self-hosting crossovers left.

8 **Commercial terms are excluded** HIGH

Project-level discounts, committed-use pricing, and enterprise volume agreements materially compress the frontier premium and appear in no figure here. Every stated premium is therefore an upper bound on what a negotiating enterprise actually pays.