Rent the Machine or Buy the Tokens? We Spent $18 Finding Out.

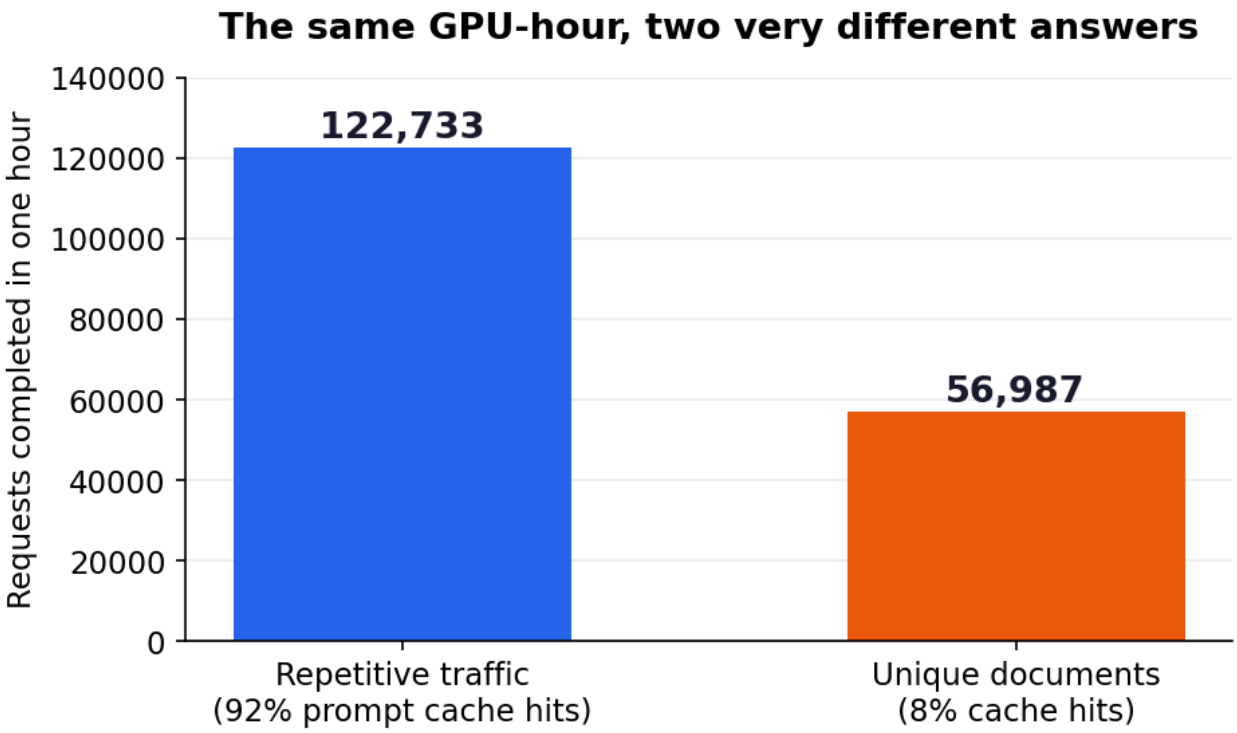
Every company building with AI eventually faces the same procurement fork: pay a provider for each unit of AI work (per-token pricing, like a taxi meter), or rent dedicated computing hardware by the hour and run an open model yourself (like leasing a car). The debate is usually settled by opinion. We settled it with invoices.

What we did

We took two freely available AI models — one small, one 50 times larger — and deployed each on rented GPU hardware in a single afternoon, using a platform that bills by the second and charges nothing when idle. We then pushed roughly 345,000 real requests through them under a controlled, repeatable test, and compared the resulting costs and speeds against the leading pay-per-token providers serving the exact same models. Total research budget: about $18.

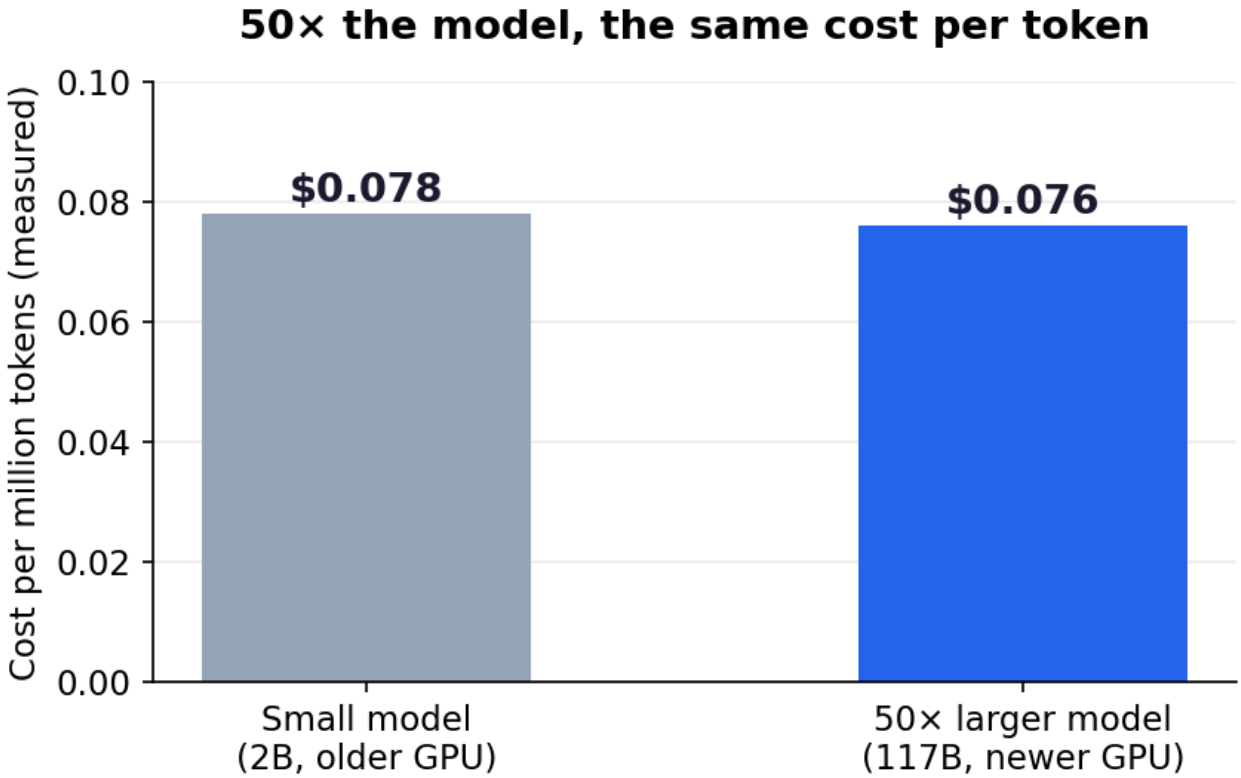
The three findings that matter

1. Your traffic pattern — not the model, not the vendor — decides which option is cheaper. The same hour of rented hardware processed 122,000 requests one day and 57,000 the next. The only difference: the first day's requests were repetitive, and modern AI infrastructure recognizes and reuses repeated work. Here is the commercial catch: per-token providers charge you full price for reused work. When you rent the machine, that efficiency is your savings; when you pay per token, it is your supplier's margin. Companies with repetitive AI workloads — templated document processing, standardized extraction, assistants with fixed instructions — are systematically overpaying under per-token pricing, and most have never measured by how much. In our test, the difference was 2.2x.



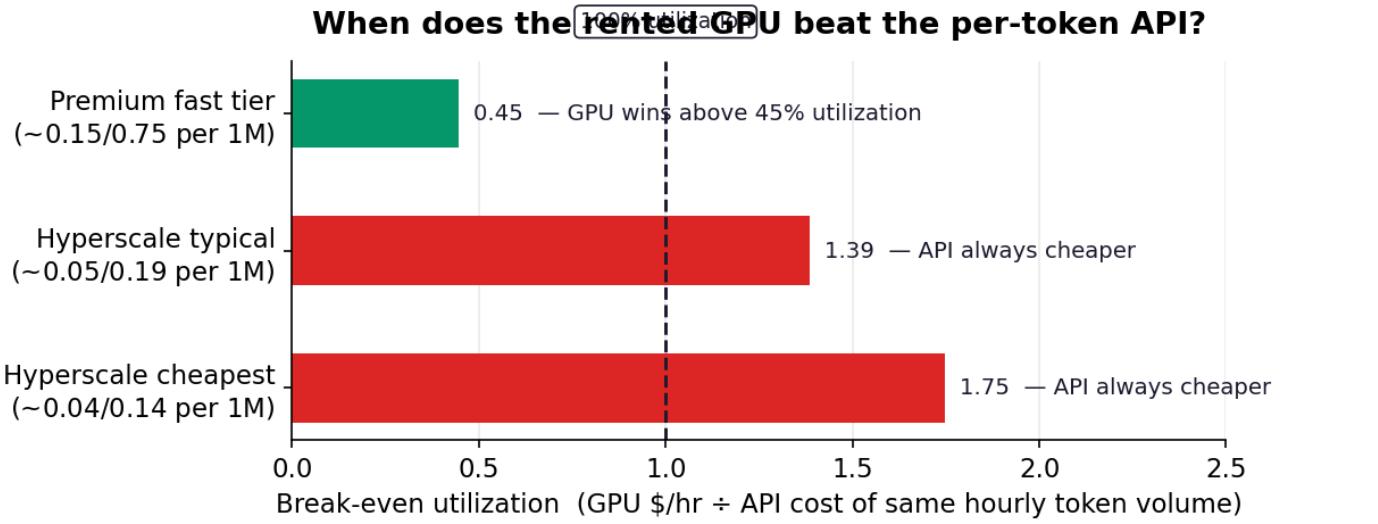
Gemma 4 E2B-it, 1×A100-40GB ($2.50/hr), identical configuration — only the traffic changed. 2.2× difference.

2. Bigger models are no longer proportionally more expensive to run. Our 50x-larger model, on newer hardware, cost the same per unit of work as the small one — and three times less per request. The instinct that "serious models need serious budgets" is increasingly outdated. What this means practically: capability upgrades that looked cost-prohibitive on last year's assumptions may already be affordable. The assumptions deserve a refresh, not a renewal.



Measured at maximum sustained load, unique-document traffic, July 2026. Hardware rented per hour.

3. Neither option wins outright — and the deciding number fits on a sticky note. The largest per-token providers pool demand from thousands of customers, achieving an efficiency no single company can match. Against their cheapest tiers, renting hardware never breaks even. But against the premium tiers — the ones sold on speed and reliability, at up to 8x the price — our rented machine broke even at 45% usage and delivered more consistent response times, with zero rate-limit interruptions and zero corrupted responses (the standard API route failed 7 times out of 300 in our comparison test). The deciding number is simple: *what your monthly AI volume would cost per token, divided into the hardware's rental price.* Your team can compute it in an hour. We published the full method, free, so they can.



GPT-OSS-120B on 1×B200 at $9/hr, measured saturation throughput (~119M tokens/hr, 98% input). July 2026 list prices.

What this means for your organization

Three questions for your next AI budget review:

1. Has anyone measured how repetitive our AI traffic actually is? (If the answer is no, you don't yet know what you should be paying.)
2. Are we buying premium per-token tiers for speed or reliability that dedicated hardware would deliver cheaper at our volumes?
3. What would our current monthly AI spend buy in rented hardware hours — and which side of break-even are we on?

There is also a strategic dimension money doesn't capture: running your own models means your data stays within infrastructure you control, your capacity can't be rate-limited during your busiest hour, and your supplier can't quietly change the model behind your product. Those risks rarely appear in cost comparisons. They appeared in ours.

The era when self-hosting AI required an infrastructure team is over — our entire evaluation was one command to deploy and $18 to run. The question is no longer whether your organization *can* compare the two options. It's whether it has.

Based on CognitX's July 2026 field benchmark: two open models, three GPU-hours, ~345,000 requests, full methodology and data published at [article].

Revision #4
Created 15 July 2026 09:45:09 by EMB
Updated 10 August 2026 10:19:34 by EMB