# The 'expensive' GPU came out cheaper — we rented both to find out

July 6, 2026 · by Mark Bartlett

Download this post as PDF

One firsthand cost guide I've wanted to write for a while, a floor report from the AI Engineer World's Fair, and a quick status check on the Qwen 3.7 open weights everyone's still waiting on. Let's go.

---

## The Cheaper GPU Costs More to Train — We Rented Both to Prove It

We rented an A100 and an H100 back to back to train a 1.5B proto, and the bill flipped my intuition. The A100 was cheaper per hour. The H100 was cheaper per run — because it finished in roughly half the wall-clock. That's the whole trap: on a fixed-FLOP training job, the hourly rate lies.

Line all six cards up and it gets sharper. A $0.34/hr RTX 4090 has the cheapest invoice on the board and is the worst deal on it — a 1.5B/15B-token run takes that card about 200 days on a preemptible instance that gets reclaimed mid-run, and the model doesn't even fit 24GB cleanly. Plot total cost against speed and it's a U: cheap-slow cards (4090, A6000, L40S) pile up months of runtime, the H200 charges a premium for compute a small model can't use, and the H100 sits at the bottom.

The guide has the current rental rates, the throughput I actually measured, and the full total-cost table so you can find the bottom of the U for your own job. There's also an own-vs-rent section with real PG&E math — training a 4090 once here in Berkeley costs more in electricity than most people guess, though a cheap-power state flips that answer completely.

📖 **A100 vs H100 vs L40S vs 4090: Why the Cheaper GPU Costs More to Train On.**

---

## Notes From the AI Engineer World's Fair Floor

I spent June 29 – July 2 at the AI Engineer World's Fair at Moscone. Quick honest read, not the hype reel: the floor was heavy on agent tooling and GPU-cloud platforms, and local/on-device AI

was a smaller but real presence — more of it than last year, still a side conversation next to the hosted-inference crowd.

Two vendors there tie directly to what we do here, so I'll call those out specifically:

- **RunPod.** The H100 and A100 numbers in the cost guide above came from real RunPod rentals, so it was useful to talk to the people behind the platform I'd actually billed hours on rather than just a dashboard. Good to see them working the floor for the training crowd, not just serving.
- **Z.ai (Zhipu AI).** The GLM team had a genuine local-AI presence, which matters — GLM-5.2 is one of the better open models out right now, and our local guide for it is one of our top pages. Nice to see the open-weights side represented in person and not just as an API booth.

If you were there and I missed you, reply and tell me what stood out to you — I'd like to compare floor notes.

---

## Qwen 3.7 Watch: Still No Open Weights

Status check, because the question keeps landing in my inbox. Qwen 3.7 went Max-first back in May — closed API on Alibaba Cloud, no weights. The 27B and 35B open variants, the ones that actually run on a 3090, were announced but still haven't shipped. So as of today, nothing's changed for local users: you're still waiting.

That means the open backbone to actually deploy is still Qwen 3.6 — 27B dense and 35B-A3B MoE. Don't hold a project hostage to a release window that keeps sliding. We've got the cadence math on how overdue this is, plus a live monitor that'll catch the drop, and a firsthand RTX 3090 benchmark queued to publish within a day of it landing.

📖 **Our Qwen 3.7 open-weights watch page is here.**

---

### Quick Hits

- **Ollama's July 1 update makes Gemma 4 up to ~90% faster on Apple Silicon** — multi-token prediction, auto-tuned and on by default, with no change to outputs. If you run Gemma 4 on a Mac, update. 📖 **Ollama-on-Mac guide.**
- **Kimi K2.7 Code shipped mid-June** — Moonshot's open coding variant, out the same week as GLM-5.2. Worth a look if local coding is your main use and you've got the VRAM for it.

- **Frontier baseline moved:** Claude Sonnet 5 landed June 30. Not local, but if you run a tiered local-plus-API setup, that's your fallback tier getting stronger.
- **Stack versions:** llama.cpp at b9882, vLLM at 0.24.0.

---

That's the week. If the cost guide saves you a bad rental decision, forward it to whoever's about to book GPU hours — that's the best thanks I can get. New here? Subscribe at insiderllm.com and it lands in your inbox every Monday.

— Mark, InsiderLLM

---

Renting GPUs for a real training run? Running local AI on weird hardware? Reply, or hit me at hello@insiderllm.com. I read everything.

---