Hugging Face got hacked -- Open local AI came to the rescue!

July 20, 2026 · by Mark Bartlett

Download this post as PDF

InsiderLLM Weekly issue 14 – July 20, 2026

Two stories this week look like opposites — one about models you can't run, one about a company getting broken into — and they land on the same point from opposite ends: the open-weight tier everyone writes off as second-rate is the one that actually works when it counts. Trillion-parameter "open" launches you can't download, and the open-weight model a Fortune-scale AI company reached for mid-crisis. Same lesson, two directions. Here's the week.

Two Trillion-Parameter "Open" Models in Ten Days. You Can Run Neither.

Kimi K3 landed July 16 — 2.8 trillion parameters, open weights promised for July 27. Three days later Alibaba announced Qwen 3.8 at 2.4T, open weights "soon," no date. Both got the "open-weight" headline. Neither is downloadable today, and even once they are, they're datacenter objects — nothing you load onto a 3090. The tier that actually tops the charts is the closed, paid Max API; the "open" part is either unshipped or too big to run.

What the horse-race coverage misses is the tier that never slowed down. Qwen 3.6 (27B dense, 35B-A3B MoE) and Gemma 4 are sitting right there — downloaded, fast, and, on the independent Artificial Analysis index, still the best open models that fit a single 24GB card. The headline chases the trillion-parameter number. The model on your drive doesn't need it to.

📖 **The full breakdown — both launches, and when runnable weights realistically land — is here.**

Hugging Face Got Breached — and Your Downloads Are Fine

First, the part that matters if you pull models off Hugging Face, which is all of us: your downloads are safe. HF disclosed on July 16 that an autonomous AI agent breached its internal infrastructure — and in the same post confirmed no evidence of tampering with public models, datasets, or Spaces, with its software supply chain verified clean. The GGUF you grab tonight is

not implicated. The breach hit internal systems and service credentials, not the weights you download.

Now the part that ties back to the spine. When HF's own responders went to analyze the attack with AI, the commercial models balked — feeding a model real exploit payloads and command-and-control artifacts tripped their safety guardrails. So the incident team ran the forensics on GLM 5.2, an open-weight model, on their own hardware. Not to fight off the attack live — to investigate it after, once the commercial tools locked them out. The tier that gets dismissed as second-rate is what the world's largest model hub reached for when its defenders needed an AI that would actually look at the evidence.

📖 **What was and wasn't hit, and why none of it changes your download habits, is here.**

So What Do You Actually Run?

Both stories point at the same answer, and it's boring in the best way. On a 24GB card, Qwen 3.6-27B dense is the pick — I clock it at 38 tok/s at Q4 on my own 3090, and it still tops the independent open-model charts in its class. Prefer the MoE? Our 35B-A3B guide has it running at 101 tok/s at UD-Q4_K_XL on the same card. Neither is new this week — that's exactly the point. While the trillion-parameter headlines play out, the model that fits your GPU is already downloaded, already fast, and already better than anything you'd have run a year ago.

That's the week. If the throughline landed for you — that the open tier isn't the consolation prize, it's the one that shows up when the stakes are real — forward this to someone who runs local AI. This newsletter is the one channel I've got that isn't at some algorithm's mercy, and the way a good issue travels is one real person handing it to another. That's the whole ask.

New here, reading this on the web? Subscribe and the next one lands in your inbox.

— Mark, InsiderLLM

Running frontier-ish open weights on a card that has no business running them? Reply, or hit me at hello@insiderllm.com. I read everything.

Source: https://insiderllm.com/blog/newsletter-2026-07-20/