MiniMax M3's asterisk, the Windows shift, and World's Fair plans

June 1, 2026 · by Mark Bartlett

Download this post as PDF

InsiderLLM Weekly issue 11 – June 1, 2026

Big model launch this week with an asterisk, a real shift in what "local AI hardware" is about to mean on the Windows side, and a llama.cpp fix that quietly matters if you run Qwen across multiple GPUs. Plus a personal note at the bottom: I'm planning to be at a conference in SF at the end of the month, and I'd like to know if you'll be there too.

MiniMax M3: "Open-Weight Frontier" — Hold the Champagne

MiniMax dropped M3 today, and the pitch is loud: the first open-weight model to combine frontier coding, a one-million-token context window, and native multimodality. The benchmarks are real and they're good — 59.0% on SWE-Bench Pro (edging out GPT-5.5), 66.0% on Terminal-Bench 2.1, and 83.5 on BrowseComp, which actually beats Opus 4.7's 79.3. The architecture story is the genuinely interesting part: a new sparse-attention design (MSA) that MiniMax says delivers 9.7x faster prefill and 15.6x faster decode than M2 at a million tokens. If that holds up, it's the real advance here — long context that doesn't make inference cost explode.

Here's the catch, and it's the whole story for anyone reading a local AI newsletter: the weights aren't out. M3 is API-only right now (live on OpenRouter at a promo ~$0.30/$1.20 per million tokens). MiniMax says weights and a technical report land on Hugging Face and GitHub "within ten days." So "open-weight" is a promise, not a fact you can act on today.

And the second catch: even when the weights drop, this is a frontier-size model. Nobody runs it on a single 3090. A 1M-context frontier model is data-center / multi-GPU territory — the open-weight release matters for the ecosystem, but "can I run it at home" is a different question with a much less exciting answer for most of us.

The honest take: watch this one. If the weights actually ship in ten days, it's a real milestone for open frontier models. But the headline you're seeing everywhere — "open-weight model challenges GPT-5.5" — is running ahead of what's actually downloadable, and "open-weight" doesn't mean "local" when the model's this big. We'll cover it properly once the weights are real and someone's benched what it takes to serve.

The Windows Local-AI Hardware Shift Is Coming

For a couple of years the answer to “what’s the best unified-memory box for running models at home” has basically been Apple Silicon. That’s about to get contested. NVIDIA is pitching the RTX Spark — a Blackwell GPU paired with an Arm-based Grace CPU and up to 128GB of shared memory — as the chip that finally makes local AI agents practical on Windows. And reporting says Microsoft and NVIDIA are teaming on AI PCs built to run actual agents rather than just Copilot, with the first machines from Dell and Microsoft’s Surface line expected to be shown next week at Computex and Build.

Why it matters for you: the Mac-vs-PC question for local AI has been lopsided toward Mac on the unified-memory front. Big shared-memory Windows boxes change that math. I’m not buying the marketing sight unseen — “practical local agents” is a claim that needs real tok/s and real thermals behind it before I believe it — but the hardware direction is worth tracking if you’re planning a build this year. I’ll have firsthand takes once there’s actual hardware to test, not slideware.

Three Things Worth Pulling This Week

llama.cpp b9434 — multi-GPU Qwen fix. If you run Qwen 3.5 or 3.6 split across three GPUs, this release fixes tensor-parallel granularity for exactly that case (plus an afmoe TP fix). The kind of unglamorous correctness fix that’s easy to miss in the changelog but saves you a confusing afternoon. Update if multi-GPU Qwen is your setup.

Liquid AI LFM2.5 8B-A1B. Last week’s release — 8B total, ~1B active — trained on 38T tokens. Liquid keeps aiming squarely at local hardware (their whole thing is models built to actually run on your machine, not scaled-down cloud leftovers). Worth a look if you want MoE efficiency without the VRAM bill of the big mixture models. Skip if you’ve already got a 9B dense pick you’re happy with.

ATOM00blue/machine-learning-library. A trending GitHub repo (Karpathy gave it a nod) — a curated ML-education corpus: papers, course lectures, and explainers normalized into clean Markdown for RAG or fine-tuning a domain model. I added it to the awesome-local-ai list this morning. Useful if you’re building a local ML tutor or a fine-tuning dataset. One honest caveat: it’s redistributing full-text papers and lecture transcripts, so the licensing is dicey and the repo may not be long-lived — grab what’s useful while it’s there.

Industry Watch

Groq raising $650M. After NVIDIA's reported $20B not-an-acquihire, inference-chip startup Groq is said to be raising $650M as it leans harder into inference (the serving side, not training). Worth noting because the inference-platform layer — the companies whose whole job is running models fast — is consolidating and getting capitalized. That's the layer a lot of local-AI tooling eventually plugs into.

Heading to AI Engineer World's Fair — Are You?

Personal note to close. I'm planning to be at the AI Engineer World's Fair at Moscone in San Francisco, June 30 – July 2. It's the biggest gathering of the people actually building this stuff — the inference platforms, the GPU-cloud folks, the agent-tooling crews — and I want to walk the floor, see what's real versus slideware, and bring back firsthand takes for you.

If you're going too, reply and say hi — genuinely. Part of why I run InsiderLLM is to stay plugged into what's actually happening in local AI, and meeting a few of you in person would be the best version of that. And if you work at (or know) one of the inference / GPU-cloud companies exhibiting, I'd love an excuse to stop by your booth.

Next issue lands in about a week. If MiniMax ships those M3 weights, I'll tell you what it actually takes to run the thing. The Qwen 2.5 -> 3.6 catalog refresh continues, and I'll have World's Fair logistics sorted by then.

— Mark, InsiderLLM

Running local AI on weird hardware? Built something novel with it? Going to be at the World's Fair? Drop us a line at hello@insiderllm.com.
