Power week in local AI: Mythos, MiroThinker, real Qwen 3.6 builds

May 18, 2026 · by Mark Bartlett

Download this post as PDF

InsiderLLM Weekly issue 9 – May 18, 2026

Three threads converged this week and they tell the same story: local AI moved from “interesting” to “serious” on measurable terms. Two researchers using a local AI agent broke through Apple’s biggest defensive investment in five days. An open-source research agent landed that actually beats closed-source on real benchmarks. And r/LocalLLaMA stopped debating whether multi-GPU Qwen 3.6 setups would work and started posting their tok/s numbers.

Calif Cracked Apple’s M5 With Mythos in Five Days

On May 14, Calif disclosed that two researchers using their Mythos Preview agent found a data-only kernel local privilege escalation chain on macOS 26.4.1, running on Apple M5. The chain bypasses Memory Integrity Enforcement — the runtime defense Apple spent five years and billions of dollars engineering. The work took five days. The 55-page report was laser-printed and hand-delivered to Apple Park.

The same week, GTIG published its own report documenting the first AI-generated zero-day found in the wild. It also named OpenClaw and OneClaw as tools showing up in active attack workflows.

Mainstream tech press led with Google’s narrative — Wired, Bloomberg, the usual circuit. Most coverage framed this as Google’s threat intelligence arm scoring a win against the bad guys. That’s not the actual story. The actual story is what Calif did, and who the white hats are.

The white hats are Calif, Koi Security, 6mile, Snyk, and Zscaler. They are the ones disclosing vulnerabilities, patching skill registries, and shipping the audit tooling that catches the threats GTIG documents after the fact. The OpenClaw skill weaponization problem isn’t new either. InsiderLLM tracked it in February — three months before GTIG named it.

Two researchers. Five days. A defense that took 60 months and a budget you can’t write on the back of an envelope. That’s where local AI actually is now.

📖 **Full piece on the Mythos crack and the framing the mainstream missed is here.**

Three Models Worth Pulling This Week

MiroThinker-1.7 is the headline. Open-source deep research agent built on Qwen3 MoE — both the 30B-A3B mini and the 235B-A22B big variant. SOTA among open-source models on BrowseComp at 74.0% and BrowseComp-ZH at 75.3%. 256K context. Up to 300 tool calls per task. Apache 2.0. The mini fits consumer hardware. HuggingFace card: `miromind-ai/MiroThinker-1.7-mini`. This is the first open-source research agent I've seen beat closed-source on real research benchmarks rather than synthetic ones. Worth pulling and pointing at your actual workload.

DeepSeek V4 Pro GGUF dropped. Previously API-only, now local-runnable on serious hardware. `batiai/DeepSeek-V4-Pro-GGUF` on HuggingFace. Our Flash vs Pro guide covers when each version makes sense.

Qwen 3.5 122B-A10B MTP GGUF landed via Unsloth — `unsloth/Qwen3.5-122B-A10B-MTP-GGUF`. MTP-enabled, so the speed gains for that scale of MoE actually compound on serial workloads.

Gemma-4-Gembrain-31B showed up on Ollama — a community merge of multiple Gemma 4 31B finetunes, uncensored. Mixed reviews on quality, but available if you want to pull and form your own opinion.

Multi-GPU Qwen 3.6 Builds Got Real

Three r/LocalLLaMA threads this week show the multi-GPU story moving out of theoretical territory:

- A backend comparison on Qwen 3.6 27B at 24GB VRAM — llama.cpp, ik_llama.cpp, BeeLlama, and vLLM head to head, score 19 and growing. Real per-token numbers, not vendor charts.
- A 4× Nvidia RTX A4000 build (16GB each) running Qwen 3.6 27B Q8 with MTP. The kind of homelab spec that was theoretical six months ago.
- Luce DFlash + PFlash on an AMD 7900XTX hitting Qwen 3.6 27B at 2.24× decode and 3.05× prefill versus llama.cpp HIP.

Three communities, three price-performance points, same model. That's convergence.

📖 **Refreshed multi-GPU setup guide is here.**

What We Shipped This Week

Monday May 11 through Sunday May 17:

- **New:** Mythos cracked Apple M5 — May 15
- **Refreshed:** OpenClaw tool call failures, FLUX locally (now covering FLUX.2 + FLUX.1), multi-GPU local AI with the honest Q4_K_M trade-off note, managing multiple Ollama models
- Earlier in the week: Qwen 3.5 9B, structured output, PI Agent, function calling, llama.cpp build errors

Eleven pieces total. The refresh cadence is what's compounding.

Next issue lands May 25. Phase 1 of the cross-vendor GPU guide is in progress. More back-catalog refreshes continuing. The 32 use cases piece is still percolating — that one's taking longer because the use cases keep getting better, and I want to ship the version that holds up six months from now.

— Mark, InsiderLLM

Running local AI on weird hardware? Built something novel with it? Drop us a line at hello@insiderllm.com.
