# Ollama's quiet Mac shift, the Qwen refresh, and the closed-weight drift

June 8, 2026 · by Mark Bartlett

Download this post as PDF

InsiderLLM Weekly issue 12 – June 8, 2026

A quiet but real change in how Ollama runs on Apple Silicon, the model picks worth updating in your setup right now, and one trend worth keeping half an eye on: Qwen's best models are starting to ship closed.

---

## Ollama 0.30 Changed the Apple Silicon Story

If you run Ollama on a Mac, 0.30.0 is worth understanding — not for a flashy feature, but for a change underneath that affects how your models actually run. Back in the 0.19 preview, Ollama added MLX as the engine for safetensors models. As of 0.30.0, it layers llama.cpp's Metal backend in alongside it — so GGUF models, which is most of what `ollama pull` lands, get first-class Metal support too. Ollama now auto-routes by file format: safetensors go to MLX, GGUF goes to llama.cpp Metal. You don't pick; it picks.

Why it matters: for a while the MLX-vs-llama.cpp question on Mac was an either/or you had to think about. Now the common case just works, and you get broader compatibility with community GGUF fine-tunes pulled straight from Hugging Face. One thing to expect on the upgrade — a first-launch storage migration to a content-addressable format, so the first run after updating takes a minute.

And the Mac-specific tuning that most setup guides skip is still the real win: confirm Metal is actually active ( `ollama ps` should say GPU, not CPU — a startling number of "Ollama is slow on Mac" complaints are just silent CPU fallback), set your env vars with `launchctl`, not `.zshrc`, and on Sonoma+ you can push past the default GPU memory cap with `iogpu.wired_limit_mb` if you need to load something big.

📖 **Our updated Ollama-on-Mac guide is here.**

---

## The Model Picks Worth Updating in Your Setup

If your local setup still defaults to last season's models, this is the month to refresh. The current open workhorses are Qwen 3.5 9B (the 16GB-Mac sweet spot, fits cleanly) and Qwen 3.6 27B (48GB+ unified memory, or a used 24GB GPU). On the smaller end, Qwen 3.5's dense lineup runs from 0.8B up through 27B, so there's a fit for almost any RAM tier.

Two practical notes from refreshing our guides this month. First, if you upgraded Ollama through 0.30.0 and you care about reproducible RAG results, that release corrected nomic-embed-text to lowercase its inputs — which means embeddings land slightly differently than before, and a re-embed realigns an older index. Most general corpora won't notice; case-sensitive content (code, proper nouns) will. Second, a heads-up that saves confusion: Qwen 3.5 dropped the typed `/think` slash-command toggle that Qwen3 had. Reasoning mode is now a chat-template parameter (`enable_thinking`), not something you type in the prompt. If you had `/think` in muscle memory, that's why it stopped doing anything.

📖 **Our Qwen guide is here.**

---

## One to Watch: Qwen's Best Models Are Going Closed

Worth keeping half an eye on, even though it changes nothing today. Qwen's flagship tier has now shipped closed twice — 3.7-Max in May, the 3.7-Plus multimodal agent in early June — both API-only on Alibaba Cloud, neither with open weights. For now that's a non-event for local users: 3.5 and 3.6 are still open, still excellent, still the models you should be running.

But it's a pattern worth noting. If Qwen keeps its frontier line proprietary and you specifically need open weights, it's good to know your fallback has gotten stronger: Google's Gemma 4 moved to a full Apache 2.0 license this generation (more open than Gemma 3 was), shipped a new 12B on June 3, and trades benchmark wins with Qwen — Qwen still leads coding and MMLU, Gemma leads math and vision tasks. Not a reason to switch anything today. Just a reason to know the open field is healthy and you have options if the closed trend continues.

---

### Quick Hits

- **Check your `ollama ps` output.** Simplest local-AI tune-up there is: if it says CPU when you expected GPU, your model is running on the wrong hardware and everything's slow for no good reason. Thirty seconds, big payoff.

- **Gemma 4 12B (June 3).** New mid-sized member of the Gemma 4 family — slots between the edge E4B and the 26B MoE, runs on a 16GB machine, Apache 2.0. A reasonable "I want one capable all-rounder" pick.
- **eBay over Amazon for used GPUs.** Quiet operational note: if you're shopping a used 3090 or 4090, the sold-listings data on eBay is where the real street price lives, not retail aggregators. Prices on the 3090 have been holding firm — it's not getting cheaper the way you'd expect a two-gen-old card to.

---

## Heading to AI Engineer World's Fair

Personal note to close. I'll be at the AI Engineer World's Fair at Moscone, June 30 – July 2 — it's under three weeks out and I'm locking in plans. If you're going, reply and let me know. Meeting a few of you in person is one of the better reasons I do this.

I'm planning to work the whole floor — the inference platforms, the GPU-cloud companies, the agent-tooling crews — and bring back firsthand takes on what's real versus slideware. If you're exhibiting, I'll likely come find you.

That's the week. Next edition drops next Monday — I'll have World's Fair logistics and a "who's exhibiting that's worth your attention" rundown as it gets closer.

— 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.

---

Source: https://insiderllm.com/blog/newsletter-2026-06-08/