Running local models on Macs gets faster with Ollama’s MLX support

Ollama, a runtime system for operating large language models on a local computer, has introduced support for Apple’s open source MLX framework for machine learning. Additionally, Ollama says it has improved caching performance and now supports Nvidia’s NVFP4 format for model compression, making for much more efficient memory usage in certain models.

Combined, these developments promise significantly improved performance on Macs with Apple Silicon chips (M1 or later)—and the timing couldn’t be better, as local models are starting to gain steam in ways they haven’t before outside researcher and hobbyist communities.

The recent runaway success of OpenClaw—which raced its way to over 300,000 stars on GitHub, made headlines with experiments like Moltbook and became an obsession in China in particular—has many people experimenting with running models on their machines.

As developers get frustrated with rate limits and the high cost of top-tier subscriptions to tools like Claude Code or ChatGPT Codex, experimentation with local coding models has heated up. (Ollama also expanded Visual Studio Code integration recently.)

The new support is available in preview (in Ollama 0.19) and currently supports only one model—the 35 billion-parameter variant of Alibaba’s Qwen3.5. Hardware requirements are intense by normal users’ standards. Users need an Apple

Silicon-equipped Mac, sure, but they also need at least 32GB of RAM, according to Ollama’s announcement.

https://arstechnica.com/apple/2026/03/running-local-models-on-macs-gets-faster-with-ollamas-mlx-support/