Mac Mini M4 for Local AI: Which Config to Buy and What It Actually Runs

February 17, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Get the Mac Mini M4 Pro 48GB ($1,799). It runs 32B models comfortably at 15-22 tok/s, fits Llama 3.3 70B at Q3 (slow but usable at 5-7 tok/s), draws ~40W under AI load, and costs about $25/year in electricity. The 24GB M4 Pro ($1,399) handles 14B models well but maxes out there. The 16GB base M4 ($599) is too constrained for serious local AI — stick to 3B-8B models. For raw speed, a PC with an RTX 3090 is 2-3x faster and cheaper. For silence, efficiency, and always-on use, the Mac Mini wins.

📚 **More on this topic:** Running LLMs on Mac M-Series · Best Local LLMs for Mac · Mac vs PC for Local AI · Ollama Troubleshooting

The Mac Mini M4 is the most efficient local AI box you can buy. Silent, palm-sized, idles at 5W, fits on a shelf behind your router. If you want a local AI server that runs 24/7 without sounding like a jet engine or costing $40/month in electricity, this is it.

But "efficient" and "fast" are different things. A $900 RTX 3090 in a used PC will generate tokens 2-3x faster for models that fit in 24GB VRAM. The Mac Mini trades speed for everything else: noise, power, unified memory, and model size ceiling.

Here's exactly which configuration to buy and what each one actually runs.

The Three Configs That Matter

Config	Chip	RAM	Bandwidth	Price	Best Model Tier
Mac Mini M4 16GB	M4	16GB	120 GB/s	$599	3B-8B
Mac Mini M4 Pro 24GB	M4 Pro	24GB	273 GB/s	$1,399	14B
Mac Mini M4 Pro 48GB	M4 Pro	48GB	273 GB/s	**$1,799**	**32B**

Apple also sells a 24GB M4 (non-Pro) for $999 and a 64GB M4 Pro for $2,199+. The 24GB M4 has only 120 GB/s bandwidth — half the Pro — which cuts inference speed significantly. The 64GB Pro is overkill unless you need 70B models at reasonable quality.

Memory bandwidth is what determines your speed on Apple Silicon. The M4 Pro's 273 GB/s is 2.3x the base M4's 120 GB/s. Same unified memory, very different throughput. This is why the M4 Pro 24GB is dramatically faster than the M4 24GB despite having the same RAM amount.

What Each Config Actually Runs

16GB — $599 (M4)

Model	Quant	VRAM Used	Speed	Verdict
Llama 3.2 3B	Q4_K_M	~2.5 GB	~50+ tok/s	Great
Qwen3-8B	Q4_K_M	~5.5 GB	~20-25 tok/s	Good, limited context
Qwen3-14B	Q3_K_M	~9 GB	~8-12 tok/s	Barely fits, short context only
Anything 32B+	—	—	—	Won't fit

The 16GB config works for casual experimentation. 8B models run well but context length is constrained — after the model loads, you have ~10GB left for KV cache. 14B barely squeezes in with heavy compromises.

Verdict: Not recommended for serious local AI. Fine as a $599 Mac Mini that can also run small models. Not a local AI machine.

24GB M4 Pro — $1,399

Model	Quant	VRAM Used	Speed	Verdict
Qwen3-8B	Q4_K_M	~5.5 GB	~35-45 tok/s	Fast, plenty of headroom
Qwen3-14B	Q4_K_M	~9 GB	~20-28 tok/s	**Sweet spot for this config**
DeepSeek-R1-14B	Q4_K_M	~9 GB	~18-25 tok/s	Solid reasoning
Qwen3-32B	Q3_K_M	~18 GB	~8-12 tok/s	Fits, but tight context

The M4 Pro 24GB is the minimum for serious local AI on Mac. 14B models at Q4 are comfortable with good context windows. 32B technically fits at Q3, but you're trading quality and context headroom.

Verdict: Good entry point. If you know you'll want bigger models eventually, spend the extra $400 for 48GB.

48GB M4 Pro — $1,799 (Recommended)

Model	Quant	VRAM Used	Speed	Verdict
Qwen3-14B	Q6_K	~12 GB	~25-32 tok/s	Overkill — runs beautifully
Qwen3-32B	Q4_K_M	~20 GB	~15-22 tok/s	**Best model for this config**
DeepSeek-R1-32B	Q4_K_M	~20 GB	~14-20 tok/s	Best reasoning at this tier
Qwen 2.5 Coder 32B	Q4_K_M	~20 GB	~15-22 tok/s	Coding powerhouse
Llama 3.3 70B	Q3_K_M	~40 GB	~5-7 tok/s	Fits, but slow

This is the sweet spot. 32B models at Q4 are comfortable with 16K+ context and room to spare. You can run two smaller models simultaneously or one large model with deep context.

The 70B option is real but slow. At 5-7 tok/s, it's below comfortable reading speed. Useful for batch processing or tasks where you'll wait for the answer anyway. For interactive chat, 32B at 3-4x the speed is the better experience.

Verdict: Best value per GB for local AI. The extra $400 over 24GB buys you the entire 32B model class. That's the jump from "good enough" to "genuinely impressive."

→ Check what fits your hardware with our Planning Tool.

Mac Mini M4 vs PC with RTX 3090

This is the real question most buyers are weighing.

Factor	Mac Mini M4 Pro 48GB	Used PC + RTX 3090
Price	$1,799	~$800-1,000
VRAM / Memory	48GB unified	24GB VRAM + system RAM
Bandwidth	273 GB/s	936 GB/s (GDDR6X)
14B Q4 speed	~20-28 tok/s	~45-55 tok/s
32B Q4 speed	~15-22 tok/s	Offload needed (~8 tok/s)
70B Q3 speed	~5-7 tok/s	~3-5 tok/s (heavy offload)
Idle power	~5W	~60-80W

Factor	Mac Mini M4 Pro 48GB	Used PC + RTX 3090
AI load power	~40W	~350W
Noise	Silent	GPU fans audible
Size	5" x 5"	Full tower

If the model fits in 24GB VRAM, the PC is faster and cheaper. The RTX 3090's 936 GB/s bandwidth crushes the M4 Pro's 273 GB/s. For 7B-14B models, the PC generates tokens 2-3x faster.

If you need 32B+ models, the Mac Mini wins. 48GB of unified memory means Qwen3-32B fits entirely in fast memory at Q4. On the PC, a 32B model at Q4 overflows 24GB VRAM and partially offloads to system RAM, cutting speed dramatically.

If you want always-on, the Mac Mini wins by a mile. A PC with a 3090 draws 350W under AI load. At $0.15/kWh running 8 hours daily, that's ~$150/year. The Mac Mini at 40W costs ~$18/year. If you want a local AI server that's always ready, the Mac Mini's power efficiency is the deciding factor.

Setup: Getting Started

Ollama (Recommended for Beginners)

```
# Install Ollama
brew install ollama

# Or download from ollama.com
curl -fsSL https://ollama.com/install.sh | sh

# Pull and run a model
ollama run qwen3:14b      # 24GB config
ollama run qwen3:32b      # 48GB config
```

Ollama uses llama.cpp as its backend. It works well on Mac, handles memory management automatically, and has the largest model library.

LM Studio (Recommended for Speed)

LM Studio supports the MLX backend, which is purpose-built for Apple Silicon. MLX is typically 20-50% faster than llama.cpp on Mac because it uses Apple's native GPU compute framework instead of the generic Metal backend.

Install LM Studio, download a model, and select the MLX backend in settings. For 48GB configs, grab the MLX-quantized version of Qwen3-32B.

Recommended First Models by Config

Config	Start With	Then Try
16GB	`ollama run qwen3:8b`	`ollama run llama3.2:3b` for speed
24GB	`ollama run qwen3:14b`	`ollama run deepseek-r1:14b` for reasoning
48GB	`ollama run qwen3:32b`	`ollama run qwen2.5-coder:32b` for coding

What NOT to Buy

Mac Mini M4 16GB for AI. Too constrained. You'll hit the wall immediately with anything above 8B. At $599 it's a great computer, just not a great AI computer.

Mac Mini M4 (non-Pro) 24GB for $999. The 120 GB/s bandwidth (vs 273 GB/s on the Pro) cuts inference speed nearly in half. For $400 more, the M4 Pro 24GB is dramatically faster. This is the worst value in the lineup for AI use.

Any Mac for image generation. Apple Silicon can't compete with NVIDIA for Stable Diffusion or Flux. CUDA acceleration on even a mid-range NVIDIA card outperforms the Mac's GPU compute for diffusion models. If image gen is your priority, build a PC.

The 36GB config. Apple doesn't sell a 36GB Mac Mini, but some older M-series configs existed at that tier. If you're looking at a used Mac with 36GB, know that the 48GB is only $200 more new — always get the 48.

Power and Cost: The Always-On Argument

Metric	Mac Mini M4 Pro	PC + RTX 3090
Idle power	~5W	~60-80W
AI inference power	~40W	~350W
Annual cost (8h/day AI, 16h idle, $0.15/kWh)	~$18	~$180
Annual cost (24/7 AI load)	~$53	~$460
Noise at idle	Silent	Fan hum
Noise under load	Silent	Audible
Physical size	5" x 5" x 2"	Full tower

The Mac Mini M4 draws less power running AI inference than most PCs draw while idle. If you want a local AI server that sits on a shelf, runs 24/7, and doesn't add noticeably to your electricity bill, this is the play.

Bottom Line

The Mac Mini M4 Pro 48GB at $1,799 is the best local AI server for people who value silence, efficiency, and model size over raw speed. It runs 32B models all day at 15-22 tok/s, fits behind a monitor, and costs $25/year in electricity.

If speed is your priority and you can live with a louder, larger box, a PC with an RTX 3090 is faster for models under 24GB and costs half as much.

The right answer depends on what you're optimizing for. Most people who buy the Mac Mini for AI end up loving the always-on convenience more than they expected. Most people who build the PC end up loving the speed. Neither choice is wrong.

Source: https://insiderllm.com/guides/mac-mini-m4-local-ai/