Mac Studio for Local AI: M4 Max vs M3 Ultra, Every Config Ranked

February 24, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** The Mac Studio M4 Max 128GB ($3,499) is the sweet spot for local AI. It runs 70B models at 10-15 tok/s, handles 100B+ MoE models, uses 60-100W under load, and stays near-silent. MLX is 30-50% faster than llama.cpp on Apple Silicon. For models under 24GB, an RTX 3090 PC is faster and cheaper. For 70B+ models, the Mac Studio wins by running them at all. The M3 Ultra 512GB ($9,499+) is the only consumer hardware that runs DeepSeek-V3 671B locally at ~20 tok/s. There is no M4 Ultra yet — the current lineup is M4 Max or M3 Ultra.

📚 **More on this topic:** Running LLMs on Mac M-Series · Best Local LLMs for Mac · Mac vs PC for Local AI · VRAM Requirements · Planning Tool

The Mac Mini M4 is the best always-on local AI box for 32B models and under. But if you want to run Llama 70B, Qwen 72B, or anything larger, you need more memory and more bandwidth than the Mini can provide.

That's where the Mac Studio comes in. The M4 Max with 128GB unified memory loads a 70B model entirely in memory — no offloading, no VRAM limits — and generates tokens at 10-15 tok/s while drawing less power than a gaming laptop. The M3 Ultra with 512GB is the only consumer hardware on the market that runs DeepSeek-V3's full 671B parameters locally.

One thing to clear up first: there is no M4 Ultra. The current Mac Studio ships with either the M4 Max or the M3 Ultra. An M4 Ultra (or M5 Ultra) is expected later in 2026, but it doesn't exist yet. If someone tells you to wait for it, that's reasonable — but you can't buy it today.

This guide covers every Mac Studio configuration, what models each one runs, how it compares to NVIDIA GPU builds on price and speed, and who should actually spend this kind of money.

The Lineup

M4 Max Configurations

Config	CPU / GPU Cores	Memory	Bandwidth	Price
Base	14-core / 32-core	36GB	410 GB/s	$1,999
Mid	16-core / 40-core	48GB	546 GB/s	$2,699
Mid-high	16-core / 40-core	64GB	546 GB/s	$2,899
Sweet spot	**16-core / 40-core**	**128GB**	**546 GB/s**	**$3,499**

The 36GB base model uses the lower-bin M4 Max chip (14-core CPU, 32-core GPU) with only 410 GB/s bandwidth. Every other config gets the full 16-core / 40-core chip at 546 GB/s. That bandwidth difference matters — it's a 33% speed gap for inference.

M3 Ultra Configurations

Config	CPU / GPU Cores	Memory	Bandwidth	Price
Base	28-core / 60-core	96GB	819 GB/s	$3,999
Mid	28-core / 60-core	256GB	819 GB/s	$6,599
High	32-core / 80-core	512GB	819 GB/s	$9,499+

The M3 Ultra is two M3 Max chips fused together via UltraFusion. It has 50% more memory bandwidth than the M4 Max (819 vs 546 GB/s), which means faster token generation for the same model. The tradeoff: it's the previous generation chip, so single-threaded CPU performance is slightly lower.

There's no 192GB option — the M3 Ultra jumps from 96GB directly to 256GB.

What Each Config Actually Runs

macOS reserves 5-8GB for itself. The numbers below reflect real usable memory.

36GB M4 Max ($1,999) — Up to 14B comfortably

Model	Quant	Memory Used	Speed	Verdict
Llama 3.1 8B	Q8_0	~9 GB	~55 tok/s	Fast
Qwen 2.5 14B	Q4_K_M	~8 GB	~30 tok/s	Good
Mixtral 8x7B	Q4_K_M	~26 GB	~15 tok/s	Tight fit, works
Llama 3.3 70B	Any	Won't fit	—	No

This is a Mac Studio that runs small and medium models well. It handles the same models as a Mac Mini M4 Pro 48GB but with more bandwidth (410 vs 273 GB/s), so everything runs about 50% faster. If you already have a Mac Mini M4 Pro, the 36GB Studio isn't worth the upgrade. If you're buying fresh and want faster 14B inference, it's a reasonable pick.

64GB M4 Max ($2,899) — The 70B entry point

Model	Quant	Memory Used	Speed	Verdict
Qwen 2.5 72B	Q4_K_M	~47 GB	~10-12 tok/s	Usable
Llama 3.3 70B	Q4_K_M	~40 GB	~11-12 tok/s	Usable
Mixtral 8x22B	Q4_K_M	~50 GB	~8 tok/s	Fits, slow
Llama 3.3 70B	Q8_0	Won't fit	—	Needs 128GB

64GB is where 70B models become possible. At Q4 quantization, Llama 70B and Qwen 72B both fit with room for a decent context window. The speed — 10-12 tok/s — is fast enough for interactive use. Not fast, but not painful.

The problem: Q4 is the ceiling. If you want Q6 or Q8 quantization for better quality, the model won't fit. And context length is constrained — a 70B Q4 model uses ~40GB, leaving ~16GB for KV cache. That limits you to roughly 8K-16K context depending on the model.

128GB M4 Max ($3,499) — The sweet spot

Model	Quant	Memory Used	Speed	Verdict
Llama 3.3 70B	Q4_K_M	~40 GB	~11-12 tok/s	Comfortable
Llama 3.3 70B	Q8_0	~70 GB	~6.5 tok/s	Quality upgrade
Qwen 2.5 72B	Q6_K	~55 GB	~9 tok/s	High quality

Model	Quant	Memory Used	Speed	Verdict
Command R+ 104B	Q4_K_M	~62 GB	~7 tok/s	Fits well
DBRX 132B	Q4_K_M	~64 GB	~6 tok/s	Works
Qwen3-235B-A22B (MoE)	Q4	~90 GB	~10-15 tok/s	Sparse MoE advantage

This is where the Mac Studio makes sense as an AI machine. 128GB gives you 70B models at high quantization with long context windows, 100B+ dense models at Q4, and large MoE models that would need multiple GPUs on a PC. The ~97-120GB of usable memory means a 70B Q4 model has 50-70GB left over for KV cache — enough for 32K+ context.

At $3,499 for a complete silent system that runs 70B models, this is the config most local AI users should consider.

96GB M3 Ultra ($3,999) — More bandwidth, same model tier

Model	Quant	Memory Used	Speed	Verdict
Llama 3.3 70B	Q4_K_M	~40 GB	~14-17 tok/s	Faster than M4 Max
Qwen 2.5 72B	Q4_K_M	~47 GB	~13-15 tok/s	Faster
Llama 3.3 70B	Q8_0	~70 GB	~8-9 tok/s	Tight but faster

The M3 Ultra 96GB at $3,999 runs the same 70B models as the M4 Max 128GB but 30-40% faster thanks to 819 GB/s bandwidth (vs 546). The tradeoff: you get less total memory (96GB vs 128GB), so Q8 70B models are a tight squeeze and 100B+ models don't fit.

If you only care about 70B Q4 speed and don't need the extra headroom, the 96GB M3 Ultra is actually the better buy for $500 more. If you want flexibility with larger models, the 128GB M4 Max wins.

256GB M3 Ultra ($6,599) — The 405B tier

Model	Quant	Memory Used	Speed	Verdict
Llama 3.1 405B	Q4_K_M	~230 GB	~5-7 tok/s	Tight but works
Multiple 70B models	Q4	~80 GB each	~15+ tok/s	Room for 2-3
Qwen3-235B-A22B	Q4	~90 GB	~15-20 tok/s	Comfortable

256GB opens up 200B+ dense models and lets you run multiple large models simultaneously. Llama 405B at Q4 barely fits (~230GB) and runs at 5-7 tok/s — slow, but it runs at all. On any NVIDIA single-GPU system, this model simply doesn't load.

512GB M3 Ultra ($9,499+) — The 671B machine

Model	Quant	Memory Used	Speed	Verdict
DeepSeek-V3 671B	4-bit	~405 GB	~20 tok/s	Works
DeepSeek-R1 671B	quantized	~400 GB	~17-18 tok/s	Works
Llama 3.1 405B	Q8_0	~400 GB	~8-10 tok/s	High quality
Multiple 100B+ models	Various	—	—	Yes

The 512GB M3 Ultra is the only consumer hardware that can run full DeepSeek-V3 671B locally. At ~20 tok/s via MLX, it's not fast — but it's local and private, running on a box that fits on a shelf. This is a niche machine for researchers, teams running private inference on frontier-scale models, and people who want to say they run 671B at home.

MLX vs llama.cpp on Apple Silicon

Both work. MLX is faster.

A November 2025 academic benchmark (arxiv) comparing inference frameworks on Apple Silicon found:

Framework	Token Generation	Notes
MLX	~230 tok/s	>90% GPU utilization
MLC-LLM	~190 tok/s	Second place
llama.cpp	~150 tok/s	Metal backend

MLX is 30-50% faster than llama.cpp for token generation on Apple Silicon. The reason: MLX uses zero-copy unified memory access, avoiding the memory transfer overhead that llama.cpp's Metal backend incurs. MLX also uses lazy evaluation to fuse operations, reducing kernel launch overhead.

Where llama.cpp still wins: prompt processing on quantized models can be ~15% faster, and model loading is faster. If you're doing batch processing with short prompts, llama.cpp might edge ahead. For interactive chat with long conversations, MLX is the better choice.

How to use MLX

```
# Via Ollama (easiest, uses llama.cpp by default, MLX runner expanding)
ollama run llama3.3:70b

# Via LM Studio (GUI, has MLX backend option)
# Download from lmstudio.ai, select MLX backend in settings

# Via mlx-lm directly (fastest, most control)
pip install mlx-lm
mlx_lm.generate --model mlx-community/Llama-3.3-70B-Instruct-4bit \
  --prompt "Your prompt here"
```

Ollama is gradually adding MLX runner support (Gemma 3, Llama, Qwen 3 as of 0.16-0.17), but it's not yet the default for all models. For maximum speed today, use mlx-lm directly or LM Studio with the MLX backend enabled.

Thermal and Sustained Performance

The Mac Studio doesn't throttle during normal AI inference. The dual-fan cooling system keeps the M4 Max at comfortable temperatures under sustained load. Reviewers consistently describe it as near-silent — fan RPM stays around 1,700 under load, which is barely audible from a few feet away.

This matters because AI inference is a sustained workload. You're not doing a quick render and stopping — you might run inference for hours. MacBook Pros with the same M4 Max chip start throttling after several minutes because the thin chassis can't dissipate 80W+ continuously. The Mac Studio sustains 145W without issues.

Metric	Mac Studio M4 Max	MacBook Pro M4 Max	PC + RTX 4090
Sustained power	145W	50-80W (throttles)	600-800W (system)
Idle power	6W	5W	80-120W
Fan noise under AI load	Near silent	Audible after minutes	Loud

Metric	Mac Studio M4 Max	MacBook Pro M4 Max	PC + RTX 4090
Throttling	No (standard inference)	Yes (sustained load)	No (with good cooling)

The power difference is striking. A Mac Studio under full AI load draws 60-100W from the wall. A PC with an RTX 4090 under load draws 600-800W. Running inference 24/7, that's roughly $50-80/year for the Mac Studio vs $400-600/year for the PC at average US electricity rates.

Price/Performance vs NVIDIA Builds

The honest comparison: NVIDIA GPUs are faster per-token for models that fit in VRAM. Mac Studio wins when you need to run models that don't fit in 24-32GB.

Mac Studio M4 Max 128GB ($3,499) vs PC + RTX 3090 (~$1,800 total build)

	Mac Studio M4 Max 128GB	PC + Used RTX 3090
Price	$3,499	~$1,800
AI memory	~120GB usable	24GB VRAM
7B Q4 speed	~70 tok/s	~121 tok/s
70B Q4 speed	~11 tok/s (in memory)	Can't load (24GB limit)
Power under load	60-100W	350-450W
Noise	Near silent	Loud
Largest model	100B+ Q4	~33B Q4

The RTX 3090 is 1.7x faster for models that fit in its 24GB. But 70B models don't fit. If you're running 7B-14B models, the PC is better value. If you need 70B+, the Mac Studio is the only option that doesn't involve multiple GPUs.

Mac Studio M3 Ultra 256GB ($6,599) vs PC + RTX 4090 (~$3,800 total build)

	M3 Ultra 256GB	PC + RTX 4090
Price	$6,599	~$3,800
AI memory	~240GB usable	24GB VRAM
70B Q4 speed	~17-20 tok/s	Can't load

	M3 Ultra 256GB	PC + RTX 4090
Power under load	100-150W	600-800W
Largest model	200B+ Q4	~33B Q4

Same story at a higher tier. The RTX 4090 is faster for small models but can't run anything over ~33B. The M3 Ultra runs 405B.

Cost per GB of AI Memory

System	Price	AI Memory	Cost/GB
RTX 4090	~$1,800	24GB	$75/GB
RTX 3090 (used)	~$800	24GB	$33/GB
Mac Studio M4 Max 36GB	$1,999	~28GB	$71/GB
Mac Studio M4 Max 128GB	$3,499	~120GB	$29/GB
Mac Studio M3 Ultra 96GB	$3,999	~88GB	$45/GB
Mac Studio M3 Ultra 256GB	$6,599	~240GB	$27/GB
Mac Studio M3 Ultra 512GB	$9,499	~480GB	$20/GB

At the high end, the M3 Ultra 512GB is $20/GB for a complete, silent system. The RTX 4090 is $75/GB for just the GPU card. Per-gigabyte, Apple Silicon is cheaper once you're past the 64GB tier. The catch: you're buying all that memory upfront whether you use it or not.

Who Should Buy This

The Mac Studio is premium hardware. It's not the budget option. Here's who it actually makes sense for.

The M4 Max 128GB ($3,499) makes sense if you want to run 70B models locally without building a multi-GPU PC. You value silence and low power. You want one box that handles daily Mac use and AI inference. You're okay with 10-15 tok/s instead of the 20+ an RTX 4090 would give on smaller models.

The M3 Ultra 256GB+ ($6,599+) is for 200B+ models or loading multiple large models simultaneously. Research that requires frontier-scale inference locally. This is a niche machine — you'll know if you need it.

Skip the Mac Studio entirely if your models fit in 24GB VRAM. A used RTX 3090 at $700-800 will be faster and much cheaper. If you're running 32B and under, a Mac Mini M4 Pro 48GB at $1,799 handles that tier well.

Consider waiting if you want an M4 Ultra. It doesn't exist yet, but it's expected in 2026. It should bring M4-generation performance to the Ultra's bandwidth tier — likely 192GB-512GB at 800+ GB/s. If you can wait 6-12 months, the M4 Ultra Mac Studio will probably be the best single-box local AI machine available.

The Bottom Line

The Mac Studio M4 Max 128GB at $3,499 is the best single-box solution for running 70B models locally. It's not the fastest — an RTX 3090 generates tokens faster for models that fit in 24GB. But the RTX 3090 can't load a 70B model. The Mac Studio can, and it does it silently at 80W.

For most local AI users who've outgrown 14B-32B models, the 128GB M4 Max is where you land. For researchers and teams who need 400B+ models, the M3 Ultra 512GB is the only consumer hardware that runs DeepSeek-V3 671B.

None of these are budget machines. If you're looking for the best value in local AI, a used RTX 3090 build or a budget AI PC will get you further per dollar. The Mac Studio is what you buy when you've decided that silence and unified memory are worth paying for.

📚 **Mac Guides:** Running LLMs on Mac M-Series · Best Local LLMs for Mac · Mac Mini M4 for Local AI · Mac vs PC for Local AI

📚 **Compare:** VRAM Requirements · RTX 4090 vs Used RTX 3090 · RTX 5090 for Local AI · Planning Tool

---

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

Free guides for running AI locally