Mac vs PC for Local AI: Which Should You Choose?

January 30, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Both work. PC wins on raw speed for models that fit in GPU VRAM — an RTX 3090 generates 2-3x faster than an M4 Pro on 7B-14B models. Mac wins when you need to load models that exceed GPU VRAM — a 128GB M4 Max runs 70B models at 8-12 tok/s natively, while a single RTX 4090 crawls at 2-5 tok/s with CPU offloading. Pick PC if you run 7B-32B models and want maximum speed. Pick Mac if you need 70B+ models, quiet operation, or a laptop that runs large models.

📚 **More on this topic:** Laptop vs Desktop for Local AI · GPU Buying Guide · LM Studio Tips & Tricks · VRAM Requirements

The Mac vs PC debate for local AI isn't about brand loyalty. It's about two fundamentally different memory architectures, and which one matches what you're actually trying to do.

A PC with an RTX 4090 has 24GB of extremely fast VRAM (1,008 GB/s) that runs 7B-32B models faster than anything Apple makes. A Mac Studio with an M4 Max has 128GB of unified memory (546 GB/s) that loads 70B models a PC can't touch without multi-GPU setups. Neither is "better." They solve different problems.

This guide covers where each platform actually wins, with real benchmarks, real prices, and honest recommendations.

The Real Question: VRAM vs. Unified Memory

Everything in this debate comes down to one technical difference:

PC (NVIDIA GPU): Your model lives in dedicated VRAM — fast but small. Consumer GPUs max out at 24GB (RTX 3090/4090). When a model doesn't fit, it spills to system RAM over the PCIe bus, and speeds collapse.

Mac (Apple Silicon): Your model lives in unified memory — slower per-byte but large. The CPU and GPU share the same memory pool (up to 128GB on M4 Max, 512GB on M3 Ultra). There's no spilling — if it fits in memory, it runs at full speed.

	PC (NVIDIA GPU)	Mac (Apple Silicon)
Memory type	Dedicated VRAM	Unified memory
Capacity	16-24GB (consumer)	24-512GB
Bandwidth	288-1,008 GB/s	120-800 GB/s
When model doesn't fit	Crashes or crawls (PCIe offloading)	Doesn't happen (more memory available)
Upgradeable	Yes (swap GPU)	No (soldered)

This single difference determines everything that follows.

Mac Advantages

Unified Memory: Load What GPUs Can't

This is the Mac's killer feature for local AI. A 70B model at Q4 needs ~35-41GB — more than any consumer GPU holds. On a PC, you're stuck with CPU offloading at 2-5 tok/s. On a Mac with 64GB+ unified memory, the same model runs entirely in memory at 8-12 tok/s.

Mac Config	Unified Memory	Largest Model (Q4)
Mac Mini M4 Pro	24-64 GB	32B (64GB config)
Mac Studio M4 Max	36-128 GB	70B-104B (128GB config)
Mac Studio M3 Ultra	96-512 GB	405B+ (512GB config)
MacBook Pro M4 Max	36-128 GB	70B-104B (128GB config)

A Mac Studio M4 Max with 128GB can run Qwen3 235B at Q3 quantization (~88GB). That would require four RTX 3090s on PC — assuming you could even get tensor parallelism working. On Mac, you just load it.

No Driver Headaches

Metal works. You install Ollama, LM Studio, or any llama.cpp-based tool, and GPU acceleration is automatic. No CUDA version conflicts, no driver updates breaking things, no `nvidia-smi` troubleshooting. Apple's MLX framework is even faster — it's built specifically for unified memory and achieves 20-87% higher throughput than llama.cpp on the same hardware.

Power Efficiency and Silence

A Mac Studio M4 Max draws 40-80W during heavy LLM inference. A PC with an RTX 3090 draws 350W for the GPU alone — 500W+ total system.

System	Power During AI Inference	Noise
Mac Studio M4 Max	40-80W	Near silent
MacBook Pro M4 Max	38-50W	Quiet (fan ramps under load)
PC + RTX 3090	~500W total	Loud
PC + RTX 4090	~385-600W total	Loud

If you're running models 24/7 — as a local API server, for example — electricity costs matter. A Mac Studio costs ~$50/year in electricity versus ~$340/year for a PC with an RTX 4090.

Portability

A MacBook Pro M4 Max with 128GB is a portable AI workstation that runs 70B models. Nothing on the PC side comes close. Gaming laptops with an RTX 4090 Mobile have only 16GB VRAM, throttle under sustained loads, and weigh twice as much.

Mac Disadvantages

Slower Per-Token Than Discrete GPUs

For models that fit in GPU VRAM, NVIDIA is faster. This is the fundamental tradeoff — GPU memory bandwidth wins the speed race.

Model	M4 Max 40c (546 GB/s)	RTX 3090 (936 GB/s)	RTX 4090 (1,008 GB/s)
7B-8B Q4	~83 tok/s	~100 tok/s	~130 tok/s
14B Q4	~38 tok/s	~55 tok/s	~70 tok/s
32B Q4	~20 tok/s	~40 tok/s	~30 tok/s

An RTX 4090 generates tokens roughly 1.5-2x faster than an M4 Max for 7B-14B models. That gap is noticeable in interactive chat. At 83 tok/s (M4 Max) you won't feel the difference from 130 tok/s (RTX 4090) — both are faster than you can read. But at larger models where speeds drop, every tok/s matters.

Prompt processing (prefill) is where NVIDIA dominates even harder — 5-8x faster due to massively higher compute TFLOPS. If you're doing batch processing or RAG with long documents, this gap stings.

No CUDA

CUDA is the default in AI. Some important tools don't support Metal:

Tool	Mac Support?
llama.cpp / Ollama / LM Studio	Yes (Metal)
MLX (Apple's framework)	Yes (native)
PyTorch (MPS backend)	Partial — no FlashAttention, limited torch.compile
ComfyUI / Stable Diffusion	Works via MPS, slower than CUDA
vLLM	Not natively (vllm-mlx exists, very new)
ExLlamaV2	No — CUDA only
TensorRT-LLM	No — NVIDIA only
bitsandbytes	No — CUDA only

For inference with Ollama, LM Studio, or MLX, Mac is fine. For training, advanced quantization, or research tools that assume CUDA, you'll hit walls.

Expensive Per GB of Memory

Apple's memory upgrades are not cheap:

Upgrade	Cost
Mac Studio: 48GB → 64GB	+$200
Mac Studio: 64GB → 128GB	+$800
MacBook Pro: 48GB → 128GB	+$1,000
Mac Studio M3 Ultra: 192GB → 512GB	+$2,400

A Mac Studio M4 Max with 128GB costs ~$3,499. A used RTX 3090 (24GB) costs ~$800. Per gigabyte, Apple charges roughly $10-17/GB for unified memory versus ~$33/GB for GPU VRAM — but the GPU memory is 2x faster.

The real comparison is total system cost, which we'll cover below.

Non-Upgradeable

This is the big one. Mac memory is soldered. The amount you buy is the amount you have forever. If you buy 64GB and realize you need 128GB, you're buying a new Mac.

A PC lets you swap GPUs, add a second card, or upgrade from an RTX 3060 to a 3090. That flexibility has real value — especially when models keep getting bigger.

PC Advantages

Faster Inference on Models That Fit

When a model fits entirely in GPU VRAM, nothing beats a dedicated GPU. The combination of high bandwidth and massive parallel compute delivers speeds Mac can't match.

GPU	7B Q4 tok/s	14B Q4 tok/s	Bandwidth
RTX 4090	~130	~70	1,008 GB/s
RTX 3090	~100	~55	936 GB/s
RTX 5060 Ti 16GB	~51	~33	448 GB/s
M4 Max (40c)	~83	~38	546 GB/s
M4 Pro	~50	~23	273 GB/s

The RTX 3090 — a five-year-old card available for ~$800 used — outperforms the M4 Max on 7B-14B models. For a used RTX 3090 system at ~$1,600-2,500, you get faster inference on everything up to 32B than a $3,499 Mac Studio.

CUDA Ecosystem

Every AI tool works with CUDA. Every tutorial assumes CUDA. Every optimization — FlashAttention, TensorRT, ExLlamaV2, vLLM, bitsandbytes — is CUDA-first. Some are CUDA-only.

If you plan to train models, fine-tune with LoRA, or use cutting-edge research tools, PC with NVIDIA is the path of least resistance. You'll spend zero time debugging Metal compatibility.

Upgradeable

A PC grows with you:

- Start with an RTX 3060 12GB (~$200 used). Run 13B models comfortably.
- Upgrade to an RTX 3090 24GB (~$800). Run 32B models, squeeze 70B.
- Add a second RTX 3090 with NVLink (~$1,600 total). Run 70B at Q4 properly.
- Swap system RAM from 32GB to 128GB for $150.

Each step reuses your existing CPU, motherboard, PSU, and case. On Mac, every step is a new machine.

Budget Flexibility

You can build a capable local AI PC for far less than the cheapest Mac option:

PC Build	Cost	What It Runs
Budget build + used RTX 3060 12GB	~$500-700	13B at Q4, 7B at Q8
Mid build + used RTX 3090	~$1,600-2,500	32B at Q4, 14B at Q8
High build + RTX 4090	~$3,100-4,400	Same models, 40-70% faster

The cheapest Mac that runs anything beyond 7B comfortably is the Mac Mini M4 Pro 48GB at ~$1,800. A used RTX 3090 PC system at similar cost runs larger models faster.

PC Disadvantages

VRAM Ceiling

Consumer NVIDIA GPUs max out at 24GB. That's enough for 32B models at Q4 (tight) but nothing larger without painful workarounds. A 70B model at Q4 on a single 24GB GPU means CPU offloading at 2-5 tok/s — essentially unusable.

Getting past 24GB on PC means:

- Dual RTX 3090s with NVLink (48GB, ~$1,600 in GPUs alone, 700W, requires careful cooling)
- Professional cards (RTX A6000 48GB, ~$3,000+ used)
- The RTX 5090 at 32GB (~$2,000+, still not enough for 70B at Q4)

None of these are simple or cheap. A Mac Studio with 128GB unified memory is a single, quiet box.

Driver and Compatibility Issues

NVIDIA CUDA mostly works, but "mostly" still means:

- CUDA version mismatches with PyTorch
- Driver updates that break things
- ROCm on AMD requiring workarounds for consumer cards

AMD's ROCm has improved significantly — vLLM now considers it a first-class platform with 93% test pass rate — but the NVIDIA CUDA experience is still smoother.

Power and Noise

An RTX 3090 runs at 350W and sounds like it. An RTX 4090 at 450W is worse. Multi-GPU setups can draw 900W+ from the wall, require 1,200W+ power supplies, and need aggressive cooling.

If your setup is in a living room, bedroom, or shared office, noise matters. Mac wins this comparison completely.

Head-to-Head Speed Comparison

Here's the direct comparison everyone wants, using Q4 quantization across platforms:

Model Size	M4 Pro (273 GB/s)	M4 Max 40c (546 GB/s)	M3 Ultra 80c (800 GB/s)	RTX 3090 (936 GB/s)	RTX 4090 (1,008 GB/s)
7B-8B	~50 tok/s	~83 tok/s	~92 tok/s	~100 tok/s	~130 tok/s
14B	~23 tok/s	~38 tok/s	~55 tok/s	~55 tok/s	~70 tok/s
32B	~11 tok/s	~20 tok/s	~41 tok/s	~40 tok/s	~30 tok/s
70B	Can't fit (64GB)	~8-12 tok/s	~14 tok/s	~2-5 tok/s (offload)	~2-5 tok/s (offload)
Memory	24-64 GB	36-128 GB	96-512 GB	24 GB	24 GB
System power	30-60W	40-80W	50-100W	~500W	~400-600W

Model Size	M4 Pro (273 GB/s)	M4 Max 40c (546 GB/s)	M3 Ultra 80c (800 GB/s)	RTX 3090 (936 GB/s)	RTX 4090 (1,008 GB/s)
System price	$1,400-2,200	$2,000-4,700	$4,000-8,000	~$1,600-2,500	~$3,100-4,400

The crossover: At 32B, the M3 Ultra matches the RTX 3090. At 70B, Mac wins by a factor of 3-6x because unified memory eliminates the offloading penalty. Below 32B, NVIDIA GPUs are consistently faster.

Note: M4 Ultra does not exist — Apple confirmed the M4 Max lacks UltraFusion. The Mac Studio currently ships with M4 Max or M3 Ultra. An M5 Ultra is expected mid-2026.

→ Check what fits your hardware with our Planning Tool.

Price Comparison at Different Tiers

Budget ($700-1,500): PC Wins

Option	Cost	What It Runs	Speed
PC + used RTX 3060 12GB	~$700-1,000	13B at Q4, 7B at Q8	~30-45 tok/s (14B)
PC + used RTX 3090	~$1,600-2,500	32B at Q4, 14B at Q8	~55 tok/s (14B)
Mac Mini M4 Pro 24GB	~$1,400	8B at Q4 comfortable	~50 tok/s (8B)
Mac Mini M4 Pro 48GB	~$1,800	14B at Q4, 32B tight	~23 tok/s (14B)

At this tier, a PC with a used RTX 3090 is the clear winner: faster on sub-32B models and cheaper than comparable Mac configurations.

Mid ($2,000-3,500): Depends on Priorities

Option	Cost	What It Runs	Speed
PC + RTX 4090	~$3,100-4,400	32B at Q4 (max speed)	~70 tok/s (14B)
Mac Studio M4 Max 64GB	~$2,700	32B at Q4, 70B barely	~38 tok/s (14B)
Mac Studio M4 Max 128GB	~$3,500	70B at Q4, 104B at Q6	~10 tok/s (70B)
MacBook Pro M4 Max 48GB	~$4,000	32B at Q4, portable	~38 tok/s (14B)

Here's where the choice gets real. The RTX 4090 PC is faster for everything up to 32B. The Mac Studio M4 Max 128GB is the only single-device option that runs 70B models natively. If 70B matters to you, the Mac wins at this price.

High ($4,000-8,000+): Mac Becomes Compelling

Option	Cost	What It Runs	Speed
PC + dual RTX 3090 NVLink	~$2,500-3,500	70B at Q4 (48GB VRAM)	~40 tok/s (70B)
Mac Studio M3 Ultra 192GB	~$5,500	70B at Q4, 180B at Q3	~14 tok/s (70B)
PC + RTX A6000 48GB	~$5,000-6,000	70B at Q3 (tight)	~25-35 tok/s (70B)
Mac Studio M3 Ultra 512GB	~$8,000+	405B+	~5 tok/s (405B)

At this tier, the PC requires multi-GPU complexity (NVLink, massive power, cooling) or expensive professional cards. The Mac is a single quiet box. For 70B daily use, the dual RTX 3090 setup is faster and cheaper — but it's also loud, power-hungry, and requires significant technical setup. The Mac Studio M3 Ultra is the "it just works" option.

The Crossover Point

When PC Wins

- **7B-32B models at maximum speed.** If your models fit in 24GB, an NVIDIA GPU is 1.5-3x faster. A used RTX 3090 at ~$800 is the best value in local AI.
- **Budget builds.** You can build a useful AI PC for under $1,000. The cheapest productive Mac is $1,400+.
- **CUDA-dependent workflows.** Training, fine-tuning, ExLlamaV2, TensorRT, vLLM — CUDA is the ecosystem. Mac support for these tools ranges from "not yet" to "never."
- **Upgradeability.** Start small, grow over time. Swap a 3060 for a 3090 for $600, not $2,000.
- **Image generation.** CUDA-accelerated Stable Diffusion and Flux are significantly faster than Metal/MPS equivalents. Flux on Mac can take 20 minutes per image versus under a minute on an RTX 4090.

When Mac Wins

- **70B+ models on a single device.** No consumer GPU holds 70B at Q4. Mac's unified memory does. This is Mac's strongest argument — not speed, but capability.

- **Quiet, always-on AI server.** A Mac Studio at 40-80W can serve a local LLM 24/7 without noise or significant electricity cost.
- **Laptop AI.** A MacBook Pro M4 Max with 128GB runs models that no PC laptop can touch. If you need large model inference on the go, there's no alternative.
- **Simplicity.** No driver debugging, no VRAM management, no power supply calculations. Plug in and run.
- **Power efficiency.** Mac delivers roughly 6x more tokens per watt than an NVIDIA GPU. If you're paying for electricity or care about energy consumption, this matters.

Recommendations by Use Case

Use Case	Recommendation	Why
7B-14B models, max speed	PC + used RTX 3090 (~$1,600-2,500)	2x faster than comparable Mac, half the price
32B models daily	PC + RTX 4090 (~~$3,100-4,400) or Mac Studio M4 Max 128GB~~ ($3,500)	PC faster, Mac loads 70B too
70B+ models, single device	Mac Studio M4 Max 128GB (~$3,500)	Only single-device option that works
Largest models (100B+)	Mac Studio M3 Ultra (~$5,500-8,000+)	Nothing else loads these models without multi-GPU
Training / fine-tuning	PC + NVIDIA GPU	CUDA ecosystem required
Image generation	PC + NVIDIA GPU	Significantly faster for SD/Flux
Portable AI workstation	MacBook Pro M4 Max 128GB (~$5,400)	No PC laptop comes close
Silent home server	Mac Mini M4 Pro 48GB (~$1,800) or Mac Studio	Near-silent, low power
Tight budget (<$1,000)	PC + used GPU	Mac doesn't play here
Coding assistant (14B-32B)	Either — PC for speed, Mac for quiet	Both run coding models well

---

The Bottom Line

This isn't a close call once you know your priorities:

Buy a PC if you primarily run models that fit in 24GB VRAM (7B-32B), care about speed per dollar, need CUDA for training or specialized tools, or want to start cheap and upgrade later. A used RTX 3090 system at ~$1,600-2,500 is the best value in local AI today.

Buy a Mac if you need to run 70B+ models on a single device, want a quiet and power-efficient workstation, need portable large-model inference, or value simplicity over raw speed. A Mac Studio M4 Max with 128GB at ~$3,500 is the single best device for running large models without complexity.

The uncomfortable truth: neither platform does everything well. The ideal setup for serious local AI might be both — a Mac for loading huge models and a PC for fast inference on smaller ones. But if you're choosing one, the decision framework is simple: count the parameters of the models you'll run most, check if they fit in 24GB, and let that answer guide you.

Related Guides

- GPU Buying Guide for Local AI
- How Much VRAM Do You Need for Local LLMs?
- Used RTX 3090 Buying Guide
- What Can You Actually Run on 24GB VRAM?
- AMD vs NVIDIA for Local AI

Sources: Hardware Corner GPU Benchmarks for LLMs, Hardware Corner Mac for LLMs, llama.cpp Apple Silicon Benchmarks, Apple MLX vs NVIDIA Inference, Apple Silicon vs NVIDIA CUDA 2025, XDA Used RTX 3090 Value King, GPU Benchmarks on LLM Inference, ArXiv: vllm-mlx on Apple Silicon

---