Local LLM Hardware Guide

Everything you need to know about consumer, prosumer, and pro hardware for running LLMs and AI models locally — February 2026

---

1. The #1 Principle — Memory Is King

2. How to Size Your Hardware

3. Quantization

4. GPU Options — NVIDIA, AMD & Apple Silicon

5. Supporting Hardware (CPU, RAM, Storage, PSU, Cooling)

6. Software Stack

7. Recommended Models by Tier

8. Key Gotchas & Pitfalls

---

1. The #1 Principle: Memory Is King, Not Compute

This is the most counterintuitive and most important concept. LLM inference has **low arithmetic intensity** — it performs relatively few math operations per byte of data fetched from memory. The bottleneck is **data movement** (getting billions of model weights from memory to compute units), not raw computation.

A 2025 NVIDIA Research paper (*Efficient LLM Inference*) confirmed that inference throughput scales primarily with memory bandwidth. IBM's *Mind the Memory Gap* study found that even GPUs with vast unused computational headroom become bandwidth-saturated during inference, leaving compute units idle.

Concrete example: an older M3 Max with 48 GB unified memory (400 GB/s bandwidth) will actually run LLMs faster during token generation than a newer M4 Pro — because the M3 Max's substantially higher memory bandwidth overwhelms any generational compute improvements.

> *When evaluating hardware, prioritise memory bandwidth and VRAM capacity first. Clock speed and TFLOPS are secondary.*

2. How to Size Your Hardware

Before buying anything, answer four questions:

- **What model size?** (7B, 14B, 32B, 70B parameters)
- **What quantization level?** (FP16 full precision, 8-bit, 4-bit)
- **What context window?** (8K, 32K, 128K — longer contexts eat more RAM)
- **How many concurrent sessions?**

Your available RAM must exceed: **model file size + ~2 GB OS overhead + context window buffer**. The KV cache for a 128K context window can consume an extra 2–8 GB. If the model doesn't fit, the system starts swapping to SSD and performance drops from ~35 tokens/sec to ~1.5 tokens/sec — essentially unusable.

3. Quantization: Running Bigger Models on Less Memory

Quantization compresses model weights from 16-bit floats to 4-bit or 8-bit integers. For example, a 13B model that needs ~26 GB at FP16 can run in 8–10 GB when quantized to 4-bit. The standard sweet spot is **Q4_K_M** quantization — it preserves most quality while dramatically shrinking memory requirements. This is what makes 70B models runnable on consumer hardware.

For image and video generation models, quantization is less helpful and can degrade output quality. VRAM capacity matters more there.

4. GPU Options

4a. NVIDIA (Path of Least Resistance)

NVIDIA's CUDA platform is the industry standard for AI workloads — every major framework is built with CUDA in mind, making NVIDIA GPUs the easiest path for local inference.

Tier	GPU	VRAM	Bandwidth	Runs	Price (approx.)
Entry	RTX 3060 12 GB	12 GB	360 GB/s	7B–13B quantized	$250–300 (used)
Mid	RTX 3090 / 4070 Ti S	24 / 16 GB	936 / 672 GB/s	Up to 30B quantized	$700–900
High	RTX 4090	24 GB	1,008 GB/s	30B comfortably; 70B w/ quant	$1,600–2,000

Tier	GPU	VRAM	Bandwidth	Runs	Price (approx.)
Top	RTX 5090	32 GB	~1,792 GB/s	70B quantized, single card	$2,000+
Multi-GPU	2× RTX 3090	48 GB total	—	70B with less quantization	~$1,400 total
Pro/Server	RTX PRO 6000, L40S, H200	48–96 GB	Varies	Full-precision large models	$5K–$30K+

Choose the RTX 5090 for speed and simplicity — one card, normal case. Choose dual RTX 3090s if on a budget or if you need maximum VRAM (48 GB total), but plan for a larger motherboard, hefty PSU, and model-splitting configuration.

> *Don't force a 70B model into 24 GB — it will run at ~2 tok/s. A well-quantized 32B model (e.g. Qwen 2.5 Coder 32B at Q4_K_M) fits comfortably in 24 GB and can outperform a heavily quantized 70B.*

4b. AMD GPUs

The Radeon RX 7900 XTX ($1,200–1,500) offers 24 GB VRAM and 960 GB/s bandwidth, roughly matching RTX 3090 performance. The major issue is software maturity: ROCm (AMD's CUDA equivalent) lags considerably, especially on Windows where support is essentially nonexistent. Even on Linux, you may face manual configuration. ROCm has improved, but always check your exact software stack before committing to AMD.

4c. Apple Silicon (The Sleeper Pick)

Apple's unified memory architecture is compelling for local LLMs: CPU and GPU share the same memory pool, so a 128 GB Mac means *all* of that is available to the model. NVIDIA still wins on absolute inference speed, but Apple offers competitive performance with excellent energy efficiency, silent operation, and macOS integration.

Key Apple Silicon considerations:

- Higher-end chips (Pro/Max/Ultra) have far more memory bandwidth — crucial for token generation speed.
- The M5 provides a 19–27% performance boost over M4 for token generation, thanks to higher bandwidth (153 vs 120 GB/s) and new GPU neural accelerators.
- **MLX** is Apple's open-source ML framework, integrated into LM Studio. Most popular models are available in MLX format.
- No external GPU expansion — you're locked into the config you buy.

Mac Config	Memory	Best For
Mac Mini M4	16–32 GB	7B–8B models only
MacBook Pro M4 Pro	36–48 GB	14B–32B quantized
Mac Studio M4 Max	64–128 GB	32B–70B quantized
Mac Studio / Mac Pro Ultra	192 GB+	70B+ full range

5. Supporting Hardware

CPU

For GPU-based inference the CPU is generally less important — it handles preprocessing and coordination. You need at least one core per GPU. Any modern Ryzen 7/9 or Core i7/i9 is fine. Threadripper or Xeon only for multi-GPU server builds.

System RAM

Even with a discrete GPU, 32 GB+ system RAM is recommended. System RAM matters more than VRAM for budget builds — when the model overflows VRAM it spills into system RAM, and swapping to SSD kills performance.

Storage

NVMe SSDs are essential. Models are large (7B ≈ 4 GB quantized, 70B ≈ 40 GB quantized). 2–8 TB NVMe drives are recommended if hosting multiple models. Fast SSD also helps with model loading times.

PSU & Cooling

A single RTX 4090 pulls ~450 W; an RTX 5090 pulls ~575 W. Dual-GPU builds need 1,000 W+ PSUs. Running LLMs pins your GPU at 100 % — thermal throttling is a real concern. Invest in good case airflow or aftermarket cooling.

6. Software Stack (2026)

Tool	Type	Notes
Ollama	CLI / API	Dead simple. 'ollama run qwen3:8b' and go. Local API compatible with OpenAI format.
LM Studio	GUI	Beginner-friendly; handles quantization & MLX. Great for Mac users.
llama.cpp	Library / CLI	The backbone most tools build on. Max control. CUDA, Metal, ROCm.
vLLM	Server	Optimised for multi-user serving with continuous batching.

Tool	Type	Notes
Jan / GPT4All	GUI	Other GUI options for non-technical users.

7. Recommended Models by Tier (Early 2026)

The biggest trend is the dominance of Qwen (Alibaba), DeepSeek, and GLM (Zhipu) — these models consistently outperform Western counterparts in coding and maths at smaller parameter counts.

Available RAM	Recommended Models
8 GB	Qwen 3 8B · Phi-3 · Gemma 3
16 GB	Qwen 3 14B · Qwen 2.5 Coder 14B · Mistral Small
32 GB	DeepSeek R1 32B · Qwen 2.5 Coder 32B
48 GB+	Llama 3.3 70B (Q3/Q4) · DeepSeek R1 70B

8. Key Gotchas & Pitfalls

- **Context window isn't free.** The KV cache for a long conversation can consume an extra 2–8 GB. If you're at the edge of capacity, keep context at 8K–16K tokens.
- **Reasoning models are hungrier.** Chain-of-thought models generate a hidden internal monologue (often thousands of tokens) before answering. They need more VRAM for context and take longer to reply.
- **Don't chase the biggest model.** A well-quantized 32B model that fits comfortably in VRAM will produce better real-world results than a 70B model crammed in with aggressive quantization and tiny context.
- **The NVIDIA vs AMD ecosystem gap is real.** If you're not on Linux and comfortable troubleshooting, stick with NVIDIA.
- **Used GPUs are a legitimate strategy.** Dual used RTX 3090s (~$700 each) give you 48 GB total VRAM for under $1,500 — excellent value if you're comfortable with multi-GPU setup.
- **Check power and cooling first.** High-end GPUs can turn a small office into a sauna and require serious PSU headroom.

Guide compiled February 2026. Hardware pricing, model rankings, and software compatibility can change rapidly — always cross-check before purchasing.