Best VRAM Cheat Sheet for Local LLMs: Every Model, Every Quant

January 27, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** For most users, 12GB is the practical minimum, 24GB is the sweet spot, and anything less than 8GB will severely limit what you can run. Qwen 3.5 9B now fits in 6.6GB at Q4, making it the best 8GB VRAM option available. A used RTX 3090 (24GB for ~$700) or RTX 4060 Ti 16GB (~$400) offers the best value depending on whether you prioritize VRAM capacity or budget. Apple's M5 Max with 128GB unified memory at 614GB/s opens up 70B+ models on a laptop.

📚 **More on this topic:** GPU Buying Guide · Quantization Explained · Context Length Explained · What Can You Run on 24GB VRAM · Qwen 3.5 9B Setup Guide · Qwen 3.5 Small Models: 9B Beats 30B · Llama 4 Guide

If you're looking to run large language models locally, you've probably noticed that every guide eventually lands on the same question: how much VRAM do you actually need? The answer isn't as simple as "more is better"—though that's technically true. What matters is understanding the relationship between model size, quantization, and your specific use case.

This guide cuts through the confusion with concrete numbers based on real-world testing. You'll learn exactly what fits in 8GB, 12GB, 16GB, and 24GB of VRAM, and which GPU makes the most sense for your budget.

Why VRAM Is the Bottleneck

The Memory Wall Problem

When you run an LLM, the entire model needs to be accessible for inference. Unlike gaming where textures can be streamed in and out, language models perform calculations across billions of parameters simultaneously. If those parameters don't fit in VRAM, you're stuck.

The math is straightforward: a model's parameter count directly determines its base memory footprint. A 7 billion parameter model in FP16 (16-bit) precision uses approximately 14GB of

VRAM. Double the parameters, double the VRAM. This is the memory wall, and no amount of clever optimization can eliminate it entirely.

Why System RAM and CPU Don't Save You

You can technically offload model layers to system RAM, but the performance penalty is brutal. GPU memory bandwidth on an RTX 4090 hits 1,008 GB/s. Your DDR5 system RAM? Maybe 50-80 GB/s. That's a 12-15x difference.

In practice, offloading a 70B model's excess layers to RAM drops inference speed from 25+ tokens/second to 3-5 tokens/second. It works for testing, but it's not a real solution for daily use.

CPU-only inference is even worse. Running a 7B model on a modern CPU might give you 2-5 tokens/second compared to 40+ on a mid-range GPU. The CPU path exists for compatibility, not performance.

What Actually Lives in VRAM

VRAM doesn't just hold model weights. You're also paying for:

- **Model weights**: The actual parameters (the big one)
- **KV cache**: Stores attention state and grows linearly with context length
- **Activation memory**: Temporary tensors during forward pass
- **Framework overhead**: CUDA, your inference backend, typically 0.5-1GB

The KV cache is often overlooked. An 8B model at 32K context length needs approximately 4.5GB just for the KV cache with FP16 precision. At longer contexts, the KV cache can exceed the model weights themselves.

Qwen 3.5 partially sidesteps this problem. It uses Gated Deltanet (a form of linear attention) in 75% of its layers, which keeps KV cache growth much lower than standard transformer attention — especially at the 262K native context length these models support. This is one reason Qwen 3.5 9B fits so well on 8GB cards even with reasonable context windows.

VRAM Requirements by Model Size

7B-9B Models (Entry Point)

The 7-9B parameter range is where most local LLM journeys begin. Models like Llama 3.1 8B, Mistral 7B, and Qwen 2.5 7B pack solid quality into a small footprint. As of March 2026, Qwen

3.5 9B is the one to beat here. It outscores models 3x its size on reasoning benchmarks while fitting in 6.6GB on Ollama.

Precision	VRAM Required	Speed (RTX 4090)
FP16	~16-19 GB	80+ tok/s
Q8_0	~8-13 GB	70+ tok/s
Q4_K_M	~5-6.6 GB	60+ tok/s

At Q4_K_M quantization, these models fit on 8GB cards with room for context. This is the entry point that actually works well.

Qwen 3.5 note: All Qwen 3.5 models are natively multimodal (text + images + video from the same weights). Processing images adds VRAM overhead on top of the base model — roughly 0.5-1.5GB depending on image resolution and count. If you're using vision features on an 8GB card, stick to single images at default resolution to stay within budget.

13B Models (The Sweet Spot That Was)

The 13B class (Llama 2 13B, CodeLlama 13B) was the previous sweet spot before 7B models got smarter. They're still relevant for specific fine-tuned variants.

Precision	VRAM Required
FP16	~26 GB
Q8_0	~14 GB
Q4_K_M	~8 GB

A 13B at Q4 fits on 12GB cards. If you have 16GB, you can run Q6 or Q8 for better quality.

27B-34B Models (Serious Performance)

Models like DeepSeek-R1-Distill-Qwen-32B, Qwen 3.5 27B, and CodeLlama 34B are where you start seeing real jumps in reasoning and coding ability. You need real hardware for these.

Precision	VRAM Required
FP16	~54-68 GB
Q8_0	~30-34 GB
Q4_K_M	~17-20 GB

Qwen 3.5 27B lands at ~17GB at Q4, which means it fits on a 24GB card with room for a reasonable context window. On Apple Silicon with 32GB+ unified memory, you can run it at Q6 or Q8 for better quality. An RTX 4090 or 3090 handles this whole range without issue. The RTX 5090 with 32GB is the first single consumer card that can run 32B models at Q8.

70B+ Models (The Big Leagues)

Llama 3.1 70B, Qwen 2.5 72B, and DeepSeek-V2.5 represent the upper limit of what's remotely practical on consumer hardware.

Precision	VRAM Required
FP16	~140-168 GB
Q8_0	~70-75 GB
Q4_K_M	~35-40 GB

Running 70B models requires either dual 24GB GPUs (48GB total), a single 48GB workstation card, or aggressive RAM offloading with significant speed penalties. Dual RTX 5090s (64GB total) can run 70B at higher quantization with approximately 27 tokens/second.

MoE Models: Big Brains, Smaller Footprint

Mixture-of-Experts (MoE) models change the VRAM math. They have a large total parameter count but only activate a fraction of those parameters per token. The catch: you still need the full model in VRAM even though only a subset fires during inference.

The Qwen 3.5 family leans heavily on this architecture, and Llama 4 Scout joined the MoE party in April 2025:

Model	Total Params	Active Params	VRAM at Q4	Speed Benefit
Qwen 3.5 35B-A3B	35B	3B	~22-24 GB	Fast — only 3B fires per token
Llama 4 Scout	109B	17B	~55 GB	109B smarts, 17B inference cost
Qwen 3.5 122B-A10B	122B	10B	~70-81 GB	122B smarts at 10B inference cost
Qwen 3.5 397B-A17B	397B	17B	~214 GB	Cloud/API only for most users

The 35B-A3B is the interesting one for consumer hardware. It needs 24GB to load the full model at Q4, but inference speed is closer to a 3B model because that's how many parameters actually compute per token. On an RTX 3090 or 4090, it runs fast while giving you 35B-level quality.

Llama 4 Scout sits in between — at ~55GB for Q4, it needs dual 24GB GPUs or an M5 Max with 128GB. It activates 17B parameters per token across 16 experts, so inference speed is closer to a 17B model. Aggressive quantization (Q2-Q3) can squeeze it onto a single 24GB card, but quality suffers.

The 122B-A10B needs an M4 Max or M5 Max with 128GB unified memory, or multi-GPU setups. The 397B is realistically cloud-only unless you have an M5 Max 128GB and are willing to run heavy quantization (Q3 or lower).

Qwen 3.5 Full Family VRAM Reference

The Qwen 3.5 family shipped in three waves (Feb 16 flagship, Feb 24 mid-range, Mar 2 small models). Every model uses a Gated DeltaNet hybrid architecture with 262K native context, and every one is natively multimodal (text + images + video from the same weights, early fusion, not bolted on).

Model	Ollama Size (Q4)	Min VRAM	Architecture	Notes
0.8B	~500 MB	2 GB	Dense	Phone, Raspberry Pi, edge devices
2B	~1.5 GB	4 GB	Dense	Laptop integrated graphics
4B	~2.5 GB	6 GB	Dense	Laptop dGPU, multimodal agent base
9B	~5 GB	8 GB	Dense	The new 8GB default. Beats GPT-OSS-120B on GPQA Diamond (81.7 vs 71.5)
27B	~16 GB	24 GB	Dense	Ties GPT-5 mini on SWE-bench Verified (72.4)
35B-A3B	~20 GB	24 GB	MoE (3B active)	112 tok/s on RTX 3090. Beats previous-gen 235B-A22B on benchmarks
122B-A10B	~70 GB	80 GB+	MoE (10B active)	Best tool use (BFCL-V4: 72.2). M4/M5 Max 128GB or multi-GPU
397B-A17B	~214 GB	256 GB	MoE (17B active)	Flagship. M3 Ultra or datacenter multi-GPU

The MoE trap: The 35B-A3B only activates 3B parameters per token, so inference is fast. But you still need ~20GB to load all 35B parameters into memory. Don't assume "3B active" means "3B VRAM." The same applies to the 122B-A10B and 397B-A17B at their respective sizes.

KV-cache advantage: Qwen 3.5's Gated DeltaNet architecture uses linear attention in 75% of its layers (3:1 ratio of DeltaNet to full attention). This keeps KV-cache growth significantly lower than standard transformers at long context lengths. A Qwen 3.5 9B running at 32K context uses roughly 40% less KV-cache memory than a standard 9B transformer would. This is part of why the 9B fits so comfortably on 8GB cards even with usable context windows.

How Quantization Changes Everything

What Quantization Actually Does

Quantization reduces the precision of model weights from 16-bit floating point to smaller representations. Instead of storing each parameter as a 16-bit number, you store it as 8-bit, 4-bit, or even 2-bit. For a deeper dive into how this works and which format to choose, see our quantization explainer.

The basic formula: **VRAM (GB) ≈ (Parameters in Billions × Bits) / 8**

A 7B model: FP16 = 14GB, Q8 = 7GB, Q4 = 3.5GB. Simple math, dramatic savings.

Common Quantization Levels

Format	Bits	VRAM Reduction	Quality Impact
FP16/BF16	16	Baseline	None (reference)
Q8_0	8	50%	Negligible
Q6_K	6	62%	Minimal
Q5_K_M	5	69%	Minor
Q4_K_M	4	75%	Noticeable on complex tasks
Q3_K_S	3	81%	Significant degradation
Q2_K	2	87%	Severe degradation

New Formats Worth Knowing (2026)

Two developments are changing the quantization landscape:

I-quants (IQ4_XS, IQ4_NL) use non-linear reconstruction with lookup tables instead of simple scaling. They retain ~95% quality at 4-bit — better than Q4_K_M — but decode slightly slower due

to table lookups. If your hardware can handle the speed hit, IQ4_XS is the new quality king at 4-bit.

Unsloth Dynamic 2.0 GGUFs analyze each layer individually and pick the quantization type that minimizes accuracy loss for that specific layer. The result outperforms both imatrix and QAT-based quants on MMLU and KL divergence benchmarks. If you see a "Dynamic" GGUF on HuggingFace, it's worth grabbing over the standard quant.

Quality vs VRAM Tradeoffs

Q4_K_M is the sweet spot for most users. It's the best tradeoff between quality and memory savings. Going lower (Q3, Q2) tanks quality and makes output unpredictable. Going higher (Q6, Q8) eats a lot more VRAM for diminishing returns.

For coding tasks, the quality gap between Q4 and Q8 is more noticeable. If you have the VRAM headroom, Q5_K_M or Q6_K makes a real difference for technical work.

Use Case	Recommended Minimum
Casual chat	Q4_K_M
Creative writing	Q4_K_M
Coding assistance	Q5_K_M or higher
Technical analysis	Q6_K or Q8_0
Research/accuracy-critical	Q8_0 or FP16

Practical Recommendations by Use Case

Casual Chat and General Assistant

Minimum: 8GB VRAM | **Recommended:** 12-16GB VRAM

For everyday questions, summarization, and general conversation, a 7-9B model at Q4 quantization works well. Qwen 3.5 9B is the current best pick. Llama 3.1 8B and Mistral 7B Instruct are also solid and fit on 8GB cards.

If you want longer conversations without context window issues, 12GB gives you comfortable headroom for larger KV caches.

Best value: RTX 4060 Ti 16GB (~~$400) or used RTX 3060 12GB~~ ($200)

Coding and Development

Minimum: 16GB VRAM | **Recommended:** 24GB VRAM

Coding tasks benefit from larger models. 7B models handle simple code completion fine, but 32B models like DeepSeek-Coder-V2 or CodeQwen are much better at understanding complex codebases.

At 24GB, you can run 32B coding models at Q4 with room for decent context windows. This is where the RTX 4090 and 3090 shine.

Best value: Used RTX 3090 (~$700) for 24GB at the best price-per-VRAM ratio — see our buying guide

Image Generation (Stable Diffusion, Flux)

Minimum: 8GB VRAM | **Recommended:** 12-16GB VRAM

Image generation has different VRAM characteristics than LLMs:

Model	Minimum VRAM	Recommended
Stable Diffusion 1.5	4GB	6GB
SDXL	6GB	8GB
FLUX (NF4 quantized)	6GB	8GB
FLUX (FP8)	12GB	16GB
FLUX (Full precision)	22GB	24GB

FLUX at full precision needs 22GB+, but NF4 quantized versions run on 6-8GB with minimal quality loss. For LoRA training, 24GB is strongly recommended.

Running Multiple Models / Hybrid Workflows

Minimum: 24GB VRAM | **Recommended:** 32GB+ VRAM

If you want to run an LLM and image generation simultaneously, or switch between multiple models without reloading, you need substantial headroom. The RTX 5090's 32GB makes this practical for the first time on a single consumer card.

What You Can Actually Run: VRAM Tier Guide

8GB VRAM (Budget Entry)

Cards: RTX 4060, RTX 3070, RTX 3060 Ti

What Works	What Doesn't
Qwen 3.5 9B at Q4 (6.6GB)	13B+ at any quality
7-8B models at Q4	FLUX full precision
SD 1.5, SDXL	Long context (32K+)
Short-medium context	

Qwen 3.5 9B is now the default recommendation for 8GB cards. At 6.6GB on Ollama, it leaves room for context and beats older 8B models on every benchmark. If you're on 8GB, this is your model. See our setup guide for the walkthrough.

12GB VRAM (Practical Minimum)

Cards: RTX 4070, RTX 3060 12GB, RTX 3080 10GB/12GB

What Works	What Doesn't
Qwen 3.5 9B at Q6_K or Q8_0	27B+ models
7-8B models at Q6-Q8	70B at any setting
13B models at Q4	Multi-model workflows
FLUX at FP8	
Longer context windows	

12GB is where local LLMs become actually useful. With Qwen 3.5 9B, you can run Q6_K (~9GB) or Q8_0 (~13GB, tight) for clearly better quality than Q4, especially on coding and reasoning tasks. You also get quality headroom on older 7-8B models and can dip into the 13B class.

16GB VRAM (Comfortable Middle Ground)

Cards: RTX 4060 Ti 16GB, RTX 4070 Ti Super, RTX 5060 Ti

What Works	What Doesn't
Qwen 3.5 9B at Q8_0 with room	70B without offloading
Qwen 3.5 27B at Q3 (tight)	Full precision anything large
13B models at Q6-Q8	
32B models at Q3-Q4	
FLUX at FP8 comfortably	

The RTX 4070 Ti Super at 16GB is the performance choice here, hitting 25-35 tok/s. The RTX 4060 Ti 16GB is the budget choice at 12-18 tok/s — slower due to its 128-bit bus, but the VRAM capacity is the same. Qwen 3.5 27B at Q3 is technically possible (~14GB) but you're leaving very little room for context. If you're on 16GB, the 9B at Q8_0 is the smarter play.

24GB VRAM (The Sweet Spot)

Cards: RTX 4090, RTX 3090, RTX 5070 Ti

What Works	What Doesn't
Qwen 3.5 35B-A3B MoE at Q4	70B without some offloading
Qwen 3.5 27B at Q4 (17GB)	Full precision 70B
32B models at Q5-Q8	Qwen 3.5 122B+
70B models at Q2-Q3 (degraded)	
FLUX full precision	
LoRA training	

This is the serious enthusiast tier. Qwen 3.5 opens up two strong options here: the 27B dense model at Q4 (17GB, leaves 7GB for context) and the 35B-A3B MoE at Q4 (~22-24GB). The MoE variant delivers 35B-level quality but infers at roughly 3B speed since only 3B parameters fire per token.

The RTX 4090 ($1,599 new) delivers approximately 52 tok/s, while the used RTX 3090 ($650-750) hits around 42 tok/s. For pure VRAM-per-dollar, the 3090 is unbeatable.

48GB+ VRAM (No Compromises)

Cards: RTX 5090 (32GB), Dual 4090/3090 (48GB), RTX 6000 Ada (48GB)

What Works	What Doesn't
70B models at Q4-Q8	70B full precision (single card)
Llama 4 Scout at Q4 (~55GB, dual GPU)	Full 671B DeepSeek R1
Qwen 3.5 122B-A10B at Q4 (70-81GB, multi-GPU)	Qwen 3.5 397B-A17B
Multiple models loaded	
Massive context windows	
Professional workflows	

The RTX 5090 at 32GB ($1,999) changed the math here. It runs 32B models at Q8 and can handle 70B at aggressive quantization on a single card. It hits 213 tok/s on 8B models and outperforms the A100 in many benchmarks.

For the Qwen 3.5 122B-A10B, you need 70-81GB at Q4. That means dual RTX 3090s (48GB, tight with Q3), an RTX 6000 Ada (48GB, same story), or Apple Silicon with 128GB unified memory. The M4 Max and M5 Max both support 128GB and are the most practical way to run this model locally. See the Apple Silicon section below.

Master VRAM Reference Table

VRAM	Best GPU Options	Max Model (Q4)	Max Model (Q8)	Best For
6GB	RTX 4060, 3060	Qwen 3.5 4B, 7B	3B	Testing only
8GB	RTX 4060, 3070	Qwen 3.5 9B (6.6GB)	7B	Casual chat, SD/SDXL
12GB	RTX 4070, 3060 12GB	13B	Qwen 3.5 9B	Daily driver, entry coding
16GB	4070 Ti Super, 4060 Ti 16GB	32B (tight)	13B	Coding, FLUX, serious use
24GB	RTX 4090, 3090	Qwen 3.5 35B-A3B, 32B	32B (tight)	Power user, training
32GB	RTX 5090	70B (tight)	32B	Enthusiast, production
48GB	2×24GB, RTX 6000 Ada	70B	70B (tight)	Professional, no compromises
64GB	M5 Pro, 2×RTX 5090	Llama 4 Scout, 122B-A10B (tight)	70B	Apple pro workflows

VRAM	Best GPU Options	Max Model (Q4)	Max Model (Q8)	Best For
128GB	M5 Max, M4 Max	Qwen 3.5 122B-A10B	70B+	Everything local

→ Use our Planning Tool to check exact VRAM for your setup.

Apple Silicon: Unified Memory Changes the Math

Apple Silicon Macs share one pool of memory between CPU and GPU. There's no separate VRAM — the entire unified memory pool is available for model weights. This means a MacBook Pro with 64GB of unified memory can load models that would require multiple discrete GPUs on a PC.

The tradeoff is bandwidth. Even the fastest Apple Silicon chips have lower memory bandwidth than a discrete GPU like the RTX 4090 (1,008 GB/s). You can fit bigger models, but they run somewhat slower per token.

Apple Silicon Memory & Bandwidth Comparison

Chip	Max Memory	Bandwidth	GPU Cores	Best Model Tier (Q4)
M4	32GB	~120 GB/s	10	9B-14B
M4 Pro	48GB	~200 GB/s	20	27B-32B
M4 Max	128GB	~400 GB/s	40	70B, 122B-A10B
M5	24GB	153 GB/s	—	9B
M5 Pro	64GB	307 GB/s	20	32B, Qwen 3.5 35B-A3B
M5 Max	128GB	614 GB/s	40	70B+, Qwen 3.5 122B-A10B

The M5 generation is a big step up. The M5 Pro pushes 307 GB/s (up from ~200 on the M4 Pro), and the M5 Max hits 614 GB/s, which puts it closer to discrete GPU territory. Both chips also have Neural Accelerators baked into every GPU core, so the effective AI throughput is higher than the bandwidth numbers alone suggest.

What This Means in Practice

M5 Pro (64GB): Runs Qwen 3.5 27B at Q8_0 (~30GB) with room left for large context windows. The 35B-A3B MoE fits at Q4 (~22-24GB) and infers fast. This is a 32B-class workstation in a laptop.

M5 Max (128GB): The only consumer hardware that can load Qwen 3.5 122B-A10B at Q4 (~70-81GB) on a single machine without multi-GPU hacks. 70B dense models at Q8 fit with room to spare. At 614 GB/s, you're looking at ~20-30 tok/s on a 70B model — not RTX 4090 fast, but fast enough for real work.

Upgrading from M4 to M5: The bandwidth increase matters more than the memory increase. If you already have an M4 Max 128GB, you can already fit the same models, but the M5 Max at 614 GB/s will spit out tokens noticeably faster. If you have an M4 Pro 48GB, the M5 Pro's jump to 64GB and 307 GB/s opens up models that flat out didn't fit before.

For detailed Mac setup instructions, see our running LLMs on Mac M-series guide.

The Bottom Line

If you're buying new today:

- **Budget (~$400):** RTX 4060 Ti 16GB—slow but capable
- **Mid-range (~$800):** RTX 4070 Ti Super 16GB—good balance
- **High-end (~$1,600):** RTX 4090 24GB—the proven workhorse
- **Flagship (~$2,000):** RTX 5090 32GB—if you can find one

If you're buying used:

- **Best value:** RTX 3090 at $650-750 on eBay or Amazon—24GB for the price of a new 12GB card

The minimum for a useful local LLM setup is still 12GB. But 8GB is more capable than it used to be, thanks to Qwen 3.5 9B fitting in 6.6GB. At 24GB, the Qwen 3.5 35B-A3B MoE gives you 35B-class quality at 3B inference speed. And on Apple Silicon, the M5 Max with 128GB at 614 GB/s runs models that needed a datacenter eighteen months ago.

VRAM is the one spec you can't fake or work around. Buy as much as you can reasonably afford. For specific card recommendations, check our GPU buying guide.

Related Guides

- GPU Buying Guide for Local AI
- What Quantization Actually Means (And Why It Matters)
- Used RTX 3090 Buying Guide for Local AI
- Qwen 3.5 9B Setup Guide
- Qwen 3.5 Small Models: 9B Beats Last-Gen 30B
- Llama 4 Guide: Running Scout and Maverick Locally
- Running LLMs on Mac M-Series

Get notified when we publish new guides.

Subscribe — free, no spam
