How Much VRAM Do You Need for Local LLMs?

January 27, 2025 · 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. 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.

📚 **More on this topic:** GPU Buying Guide · Quantization Explained · Context Length Explained · What Can You Run on 24GB VRAM

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—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.

VRAM Requirements by Model Size

7B-8B Models (Entry Point)

The 7-8B parameter range is where most local LLM journeys begin. Models like Llama 3.1 8B, Mistral 7B, and Qwen 2.5 7B hit a remarkable quality-to-size ratio.

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

At Q4_K_M quantization, these models fit comfortably on 8GB cards with room for context. This is the gold standard for accessible local deployment.

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.

30B-34B Models (Serious Performance)

Models like DeepSeek-R1-Distill-Qwen-32B and CodeLlama 34B deliver noticeably better reasoning and coding ability. This tier requires real hardware commitment.

Precision	VRAM Required
FP16	~60-68 GB
Q8_0	~32-34 GB
Q4_K_M	~18-20 GB

32B models require approximately 19-20GB at Q4 quantization. An RTX 4090 or 3090 handles these comfortably. The RTX 5090 with 32GB is the first single consumer card that can run these 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

Precision	VRAM Required
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.

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

Quality vs VRAM Tradeoffs

Q4_K_M is the sweet spot for most users. It offers the best balance of quality, speed, and memory efficiency. Going lower (Q3, Q2) causes significant quality degradation and

unpredictable behavior. Going higher (Q6, Q8) requires substantially more VRAM with diminishing quality improvements.

For coding tasks specifically, the quality difference between Q4 and Q8 becomes more apparent. If you have the VRAM headroom, Q5_K_M or Q6_K provides a meaningful improvement 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 7-8B model at Q4 quantization performs remarkably well. Llama 3.1 8B, Mistral 7B Instruct, or Qwen 2.5 7B Instruct are all excellent choices that 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 significantly from larger models. While 7B models can handle simple code completion, 32B models like DeepSeek-Coder-V2 or CodeQwen show dramatically better understanding of 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
7-8B models at Q4	13B+ at any quality
SD 1.5, SDXL	FLUX full precision
Short-medium context	Long context (32K+)

At 8GB, you're running the smallest serious models at aggressive quantization. It works, but you'll hit the ceiling quickly.

12GB VRAM (Practical Minimum)

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

What Works	What Doesn't
7-8B models at Q6-Q8	32B+ models
13B models at Q4	70B at any setting
FLUX at FP8	Multi-model workflows
Longer context windows	

12GB is where local LLMs become genuinely useful. You get quality headroom on smaller 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
13B models at Q6-Q8	70B without offloading
32B models at Q3-Q4	Full precision anything large
FLUX at FP8 comfortably	
Good context windows	

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.

24GB VRAM (The Sweet Spot)

Cards: RTX 4090, RTX 3090, RTX 5070 Ti

What Works	What Doesn't
32B models at Q5-Q8	70B without some offloading
70B models at Q2-Q3 (degraded)	Full precision 70B
FLUX full precision	
LoRA training	
Large context windows	

This is the serious enthusiast tier. 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)
Multiple models loaded	Full 671B DeepSeek R1
Massive context windows	
Professional workflows	

The RTX 5090 at 32GB ($1,999) is a game-changer—it runs 32B models at Q8 and can handle 70B at aggressive quantization on a single card. It achieves 213 tok/s on 8B models and outperforms the A100 in many benchmarks.

Master VRAM Reference Table

VRAM	Best GPU Options	Max Model (Q4)	Max Model (Q8)	Best For
6GB	RTX 4060, 3060	7B	3B	Testing only

VRAM	Best GPU Options	Max Model (Q4)	Max Model (Q8)	Best For
8GB	RTX 4060, 3070	8B	7B	Casual chat, SD/SDXL
12GB	RTX 4070, 3060 12GB	13B	8B	Daily driver, entry coding
16GB	4070 Ti Super, 4060 Ti 16GB	32B (tight)	13B	Coding, FLUX, serious use
24GB	RTX 4090, 3090	32B (comfortable)	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

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

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 genuinely useful local LLM setup is 12GB. At 8GB, you're constantly compromising. At 24GB, almost nothing is off-limits except the largest models. And at 32GB with the RTX 5090, you're running what required datacenter hardware two years 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

Source: https://insiderllm.com/guides/vram-requirements-local-llms/