Mixtral VRAM Requirements: 8x7B and 8x22B at Every Quantization Level

February 17, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Mixtral 8x7B at Q4_K_M needs ~28GB VRAM — it fits on a 24GB card with some CPU offloading, or cleanly at Q3_K_M (~22GB). Mixtral 8x22B at Q4_K_M needs ~88GB, requiring multi-GPU or a 96GB Mac. The catch: despite only activating 12.9B parameters per token, all 46.7B must be loaded into VRAM. In 2026, dense models like Qwen3-14B (MMLU 81.1) outperform Mixtral 8x7B (MMLU 70.6) while using less VRAM. Mixtral still has a role — but it's narrower than it was.

📚 **More on this topic:** Mistral & Mixtral Guide · VRAM Requirements · What Can You Run on 24GB · Quantization Explained

Mixtral is confusing. The model has 46.7 billion parameters, but only 12.9 billion activate per token. That sounds like it should use 12.9B worth of VRAM. It doesn't. You need VRAM for all 46.7 billion.

If you've been searching for exactly how much VRAM Mixtral needs at each quantization level — and whether it's still worth running in 2026 — this is the guide.

Why MoE VRAM Is Confusing

Mixtral uses a Mixture of Experts (MoE) architecture. Each token routes through 2 of 8 expert networks, so only 12.9B parameters do work per token. This makes it fast for its quality — you get near-70B performance at 13B inference cost.

But all 8 experts must live in memory. The routing layer decides which experts to use after the input arrives. If any expert could be needed, all experts must be loaded. There's no way to predict which 2 of 8 experts a given token will need.

The result: Mixtral 8x7B needs VRAM for a 46.7B model, despite behaving like a 12.9B model during inference. This is the single most common point of confusion in MoE VRAM discussions.

Mixtral 8x7B VRAM Requirements

46.7B total parameters. 12.9B active per token. 32K context window.

Quantization	File Size	VRAM Needed	Fits On	Notes
FP16	~93 GB	~95 GB	Multi-GPU only	Not practical for consumer hardware
Q8_0	49.6 GB	~52 GB	Dual 24GB or 64GB Mac	Overkill — quality gain over Q6 is tiny
Q6_K	38.4 GB	~41 GB	48GB Mac or dual 24GB	Excellent quality, large
Q5_K_M	32.2 GB	~35 GB	48GB Mac or dual 16GB	Good quality/size balance
Q4_K_M	26.4 GB	~29 GB	24GB + offload	**Most popular choice**
Q3_K_M	20.4 GB	~23 GB	**Single 24GB card**	Best fit for RTX 3090/4090
Q2_K	15.6 GB	~18 GB	16GB GPU (tight)	Noticeable quality loss

VRAM estimates include ~2-3GB overhead for KV cache at moderate context lengths (4K-8K tokens). Longer context windows increase VRAM usage.

The 24GB sweet spot: Q3_K_M at ~23GB is the practical choice for a single RTX 3090 or 4090. Q4_K_M is better quality but needs partial CPU offloading on 24GB cards, which cuts speed.

Mixtral 8x22B VRAM Requirements

141B total parameters. 39B active per token. 64K context window.

Quantization	File Size	VRAM Needed	Fits On	Notes
FP16	~263 GB	~270 GB	Not practical	Theoretical only
Q8_0	149.5 GB	~155 GB	Specialized hardware	4x 48GB or 2x 80GB
Q6_K		~120 GB	128GB Mac Studio	Excellent quality

Quantization	File Size	VRAM Needed	Fits On	Notes
	115.6 GB			
Q5_K_M	100.1 GB	~105 GB	128GB Mac or 3x 48GB	
Q4_K_M	85.7 GB	~90 GB	96GB Mac or 4x 24GB	**Recommended if you have the hardware**
Q3_K_M	67.9 GB	~72 GB	3x 24GB or 64GB Mac	Quality holds up reasonably
Q2_K	52.2 GB	~56 GB	2x 24GB + offload	Significant quality loss

The 8x22B is serious hardware territory. Even at Q3_K_M, you need 72GB of VRAM. That's three RTX 3090s, a 96GB Mac, or multiple datacenter GPUs. Most people searching for Mixtral 8x22B VRAM requirements are hoping it'll fit somewhere reasonable. At Q4_K_M, it won't fit on anything less than ~90GB of combined VRAM.

What Actually Fits on Your GPU

Your GPU	Mixtral 8x7B	Mixtral 8x22B	Better Alternative
8GB	No	No	Mistral 7B
12GB	No	No	Qwen3-14B at Q3
16GB	Q2_K (quality loss)	No	Qwen3-14B at Q4 — better quality, less VRAM
24GB	**Q3_K_M fits**	No	Qwen3-32B at Q4 (same VRAM, better benchmarks)
2x 24GB	Q5_K_M+ fits	Q2_K (barely)	Depends on workload
48GB Mac	Q5_K_M fits well	No	Qwen3-32B at Q6
96GB+ Mac	Q8_0 comfortably	Q4_K_M	The 8x22B finally makes sense

→ Not sure what fits? Try our Planning Tool.

The Honest Take: Is Mixtral Still Worth Running?

Mixtral 8x7B was groundbreaking when it launched in December 2023. A MoE model that matched Llama 2 70B on benchmarks while running at 13B speed. Nothing else could do that.

In 2026, dense models have caught up. The numbers tell the story:

Model	MMLU	Active Params	VRAM (Q4)	Release
Mixtral 8x7B	70.6	12.9B	~29GB	Dec 2023
Qwen3-14B	81.1	14B	~9GB	Apr 2025
Llama 3.3 70B	86.0	70B	~40GB	Dec 2024
Mixtral 8x22B	77.8	39B	~90GB	Apr 2024

Qwen3-14B scores 81.1 on MMLU versus Mixtral 8x7B's 70.6 — and needs 9GB of VRAM instead of 29GB. On a 24GB card, you can run Qwen3-14B at Q8 with room to spare, while Mixtral 8x7B barely fits at Q3.

When Mixtral 8x7B still wins:

- Inference speed. At Q3 on 24GB, Mixtral generates tokens faster than a dense 32B model because only 12.9B params activate per token. If you prioritize speed over benchmark scores, MoE has an edge.
- Long context. Mixtral's 32K window is larger than many dense alternatives at similar quality.
- Multi-turn conversation. Some users report Mixtral handles complex dialogue flows better than similarly-sized dense models, though this is subjective.

When to use something else:

- If quality per VRAM matters, Qwen3-14B is the better model at the 24GB tier. Better benchmarks, less VRAM, and a newer training dataset.
- If you have 48GB+, skip Mixtral 8x7B entirely. Run Qwen3-32B or DeepSeek-R1-32B instead.
- If you're considering Mixtral 8x22B, you need hardware that could also run Llama 3.3 70B, which outscores it on most benchmarks.

Running Mixtral Locally

Mixtral works with all major inference tools:

```
# Ollama (easiest)
ollama run mixtral:8x7b    # Q4_K_M by default

# Specify quantization
ollama run mixtral:8x7b-instruct-v0.1-q3_K_M

# For 8x22B (if you have the VRAM)
ollama run mixtral:8x22b
```

For text-generation-webui, download GGUF files from bartowski or mradermacher on HuggingFace. The Q3_K_M and Q4_K_M variants get the most downloads.

If you're using ExLlamaV2 on NVIDIA, EXL2 versions of Mixtral are available from turboderp and LoneStriker. ExLlamaV2 is particularly effective with MoE models because its custom CUDA kernels handle the expert routing efficiently.

Bottom Line

Mixtral 8x7B needs 23-29GB VRAM at practical quantization levels. It fits on a 24GB card at Q3_K_M. The 8x22B needs 72-90GB and is strictly multi-GPU or high-end Mac territory.

The MoE architecture means great inference speed for its quality level, but the VRAM cost is the full parameter count — not the active parameter count. That's the fact that trips everyone up.

If you're buying hardware specifically for Mixtral, don't. Buy for the model you actually want to run, and if Mixtral fits in that VRAM budget, great. In 2026, Qwen3-14B does more with less VRAM on most tasks. Mixtral's edge is speed and context length, not raw quality.

Get notified when we publish new guides.

Subscribe — free, no spam

Source: https://insiderllm.com/guides/mixtral-vram-requirements/

Free guides for running AI locally