Mixtral 8x7B & 8x22B VRAM Requirements

February 5, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Mixtral 8x7B needs ~26GB VRAM at Q4_K_M — it won't fit on a single 24GB GPU without cutting context. At Q3_K_M it squeezes into 24GB with short context. Mixtral 8x22B needs ~86GB at Q4_K_M, requiring dual 48GB cards or aggressive quantization. The catch: all 46.7B parameters (8x7B) or 141B parameters (8x22B) must load into VRAM, even though only 2 experts run per token. In 2026, dense models like Qwen 3 32B deliver similar quality to 8x7B while fitting comfortably on 24GB at Q4. Mixtral 8x22B still has a niche for its 64K context window, but you need serious hardware.

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

Mixtral models have some of the most confusing VRAM requirements in local AI. “8x7B” sounds like it should need 7B worth of memory. It doesn’t. It needs closer to 47B worth. And “8x22B” isn’t a 22B model — it’s 141B parameters that all need to live in VRAM simultaneously.

This guide gives you exact numbers at every quantization level so you can figure out whether your GPU can actually run these models, or whether you’re better off with a dense alternative.

Why MoE VRAM Is Confusing

How Mixture of Experts Works

Mixtral uses a Mixture of Experts (MoE) architecture. Instead of one monolithic feedforward network, the model has 8 separate “expert” networks. For each token, a router selects 2 experts to process it. The other 6 sit idle.

This means Mixtral 8x7B has 46.7 billion total parameters but only activates 12.9 billion per token. Inference speed scales with active parameters — so it runs roughly as fast as a 13B dense model.

Here’s the catch that trips everyone up: **all 46.7B parameters must be loaded into VRAM.** The router can’t predict which experts it’ll need for the next token, so every expert stays resident in memory. You’re paying 47B worth of VRAM for 13B worth of compute.

The Numbers That Matter

	Mixtral 8x7B	Mixtral 8x22B
Total parameters	46.7B	141B
Active per token	12.9B (2 of 8 experts)	39B (2 of 8 experts)
Context window	32K	64K
VRAM behavior	Loads like a ~47B dense model	Loads like a ~141B dense model
Speed behavior	Runs like a ~13B dense model	Runs like a ~39B dense model

This is the fundamental tradeoff of MoE: you get the speed of a smaller model and the quality of a larger one, but the VRAM cost of the larger one.

Mixtral 8x7B VRAM Requirements

Every quantization level, with real file sizes from GGUF builds:

Quantization	Bits per Weight	Model Size	VRAM Needed*	Quality Impact
FP16	16	~93 GB	~95 GB	Baseline (reference)
Q8_0	8	49.6 GB	~52 GB	Negligible loss
Q6_K	6	38.4 GB	~41 GB	Minimal
Q5_K_M	5	32.2 GB	~35 GB	Minor
Q4_K_M	4	26.4 GB	~29 GB	Noticeable on complex tasks
Q3_K_M	3	20.4 GB	~23 GB	Significant degradation
Q2_K	2	15.6 GB	~18 GB	Severe — not recommended

*VRAM needed = model size + ~2-3GB for KV cache (4K context) + framework overhead. Longer context adds more.

What GPU Can Run 8x7B?

GPU Tier	Best Quantization	Context Limit	Verdict
8GB (RTX 4060)	None	—	Won't fit. Not even Q2_K.
12GB (RTX 3060)	None	—	Won't fit at any useful quantization.
16GB (RTX 4060 Ti 16GB)	Q2_K (barely)	~2K tokens	Technically possible, practically useless. Severe quality loss.
24GB (RTX 3090/4090)	Q3_K_M	~4K tokens	Tight fit with degraded quality. Workable for short conversations.
32GB (RTX 5090)	Q4_K_M	~8K tokens	The real sweet spot. Good quality with decent context.
48GB (dual 24GB / A6000)	Q6_K	~16K tokens	Comfortable. Full quality with room for context.

The honest assessment: **Mixtral 8x7B is awkward on consumer hardware.** It's too big for 24GB at good quality, and if you're buying 32GB+ hardware, dense models like Qwen 3 32B give you better quality at Q4_K_M (~20GB) with VRAM to spare.

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

Mixtral 8x22B VRAM Requirements

The big one. Here's what you're looking at:

Quantization	Bits per Weight	Model Size	VRAM Needed*	Quality Impact
FP16	16	~282 GB	~285 GB	Baseline
Q8_0	8	149 GB	~152 GB	Negligible loss
Q6_K	6	116 GB	~119 GB	Minimal
Q5_K_M	5	100 GB	~103 GB	Minor
Q4_K_M	4	85.6 GB	~88 GB	Noticeable on complex tasks
Q3_K_M	3	67.8 GB	~71 GB	Significant degradation
Q2_K	2	52.1 GB	~55 GB	Severe — not recommended

*Assumes 4K context. 8x22B's 64K context window at full length adds substantially more.

What GPU Can Run 8x22B?

GPU Setup	Best Quantization	Context Limit	Verdict
24GB single	None	—	Not happening.
2x 24GB (48GB total)	Q2_K	~4K tokens	Barely. Quality is bad.
48GB (A6000 / L40)	Q3_K_M	~4K tokens	Tight. Degraded quality.
2x 48GB (96GB total)	Q4_K_M	~8K tokens	The minimum serious setup.
80GB (A100 / H100)	Q4_K_M	~16K tokens	Comfortable single-GPU option.
128GB+	Q6_K+	~32K tokens	Full quality with good context.

Mixtral 8x22B is a datacenter model that people try to run on consumer hardware. Unless you have 96GB+ of GPU memory across multiple cards, look at Llama 3.1 70B instead — it fits on 2x 24GB GPUs at Q4_K_M (~40GB) and delivers comparable quality.

How Context Length Eats Your VRAM

The tables above assume short context (~4K tokens). But Mixtral 8x7B supports 32K and 8x22B supports 64K. The KV cache grows linearly with context length and eats into your available VRAM.

Here's the key insight: **KV cache scales with active parameters, not total parameters.** Since only 2 experts run per token, the KV cache for 8x7B scales like a ~13B model, and for 8x22B like a ~39B model. This is one area where MoE actually helps.

KV Cache VRAM by Context Length

Context Length	Mixtral 8x7B KV Cache	Mixtral 8x22B KV Cache
2K tokens	~0.3 GB	~0.8 GB
4K tokens	~0.5 GB	~1.5 GB
8K tokens	~1.0 GB	~3.0 GB
16K tokens	~2.0 GB	~6.0 GB

Context Length	Mixtral 8x7B KV Cache	Mixtral 8x22B KV Cache
32K tokens	~4.0 GB	~12.0 GB
64K tokens	N/A	~24.0 GB

For a deeper dive into how context length affects VRAM, see our context length explainer.

At 32K context on Mixtral 8x7B with Q4_K_M, you need ~26GB (model) + ~4GB (KV cache) + ~2GB (overhead) = **~32GB total.** That's exactly one RTX 5090. On a 24GB card, you're limited to roughly 4K-8K tokens before running out of memory.

Mixtral vs Dense Models: The Honest Comparison

This is where people should pay close attention. MoE models made sense when there weren't good dense alternatives at the same quality tier. In 2026, the landscape has changed.

Mixtral 8x7B vs Dense Alternatives

Mixtral 8x7B performs roughly like a dense 30B model on benchmarks. Here's how the VRAM compares:

Model	Quality Tier	VRAM at Q4_K_M	Fits 24GB?
Mixtral 8x7B	~30B dense	~29 GB	No
Qwen 3 32B	32B dense	~20 GB	Yes, with room
DeepSeek-R1-Distill-32B	32B dense	~20 GB	Yes, with room
Llama 3.3 70B	70B dense	~40 GB	No (needs 2x 24GB)

Qwen 3 32B and DeepSeek-R1-Distill-32B both fit on a single 24GB GPU at Q4_K_M with VRAM left for context. They match or beat Mixtral 8x7B on most benchmarks. Unless you specifically need Mixtral's architecture for a fine-tune or have legacy infrastructure, **dense 32B models are the better choice on the same hardware.**

For more on these alternatives, see our Qwen Models Guide and DeepSeek Models Guide.

Mixtral 8x22B vs Dense Alternatives

Mixtral 8x22B competes with Llama 3.1 70B. The 64K context window is its main advantage.

Model	Quality Tier	VRAM at Q4_K_M	Context
Mixtral 8x22B	~70B dense	~88 GB	64K
Llama 3.1 70B	70B dense	~40 GB	128K
Qwen 2.5 72B	72B dense	~42 GB	128K

Llama 3.1 70B needs less than half the VRAM and has double the context window. On raw benchmarks, it's competitive or better. The only scenario where 8x22B wins is if you've invested in the infrastructure and need its specific MoE speed characteristics — inference speed roughly matches a 39B dense model, which is faster than a 70B dense model on the same hardware.

Best Quantization Sweet Spots

For Mixtral 8x7B

Your Hardware	Recommendation
24GB GPU (RTX 3090/4090)	Q3_K_M with 4K context. Functional but not great. Honestly, run Qwen 3 32B at Q4_K_M instead — better quality, better fit.
32GB GPU (RTX 5090)	Q4_K_M with 8K context. The proper sweet spot if you want MoE.
48GB (A6000 or dual 24GB)	Q5_K_M or Q6_K. Full quality with room for long context.

For Mixtral 8x22B

Your Hardware	Recommendation
2x 24GB (48GB total)	Don't. Run Llama 3.1 70B at Q4_K_M instead.
48GB single card	Q3_K_M with very limited context. Marginal experience.
2x 48GB (96GB total)	Q4_K_M with 8K context. First viable setup.
80GB single card (A100)	Q4_K_M with 16K context. Comfortable.

When Mixtral Still Makes Sense

After all that, you might wonder why anyone runs Mixtral in 2026. There are a few valid reasons:

Inference speed. MoE's core advantage is speed at quality. Mixtral 8x7B gives you 30B-class quality at 13B-class speed. If you have the VRAM and need fast responses, that's a real benefit. On a 48GB card, 8x7B at Q6_K generates tokens noticeably faster than a dense 30B at the same quantization.

Fine-tuned variants. There's a large ecosystem of Mixtral fine-tunes for specific tasks — roleplay, coding, instruction following. If a fine-tune exists for your exact use case and nothing comparable exists for dense models, that's a reason.

8x22B's 64K native context. If you need long context on self-hosted infrastructure and have the hardware, 8x22B at Q4 handles it well. But Llama 3.1 70B at 128K context on less VRAM is usually the better choice.

Existing deployments. If your setup already runs Mixtral and works, there's no urgent reason to migrate. The model hasn't gotten worse — the alternatives have just gotten better.

The Bottom Line

Mixtral 8x7B needs ~26-29GB at Q4_K_M. It won't fit on a 24GB GPU with usable context. If you have 32GB+ VRAM and want MoE speed benefits, it's a solid choice. If you have 24GB and want 30B-class quality, skip Mixtral and run Qwen 3 32B.

Mixtral 8x22B needs ~86-88GB at Q4_K_M. It's a datacenter model. If you don't have 96GB+ of GPU memory, run Llama 3.1 70B instead — half the VRAM, comparable quality, more context.

The MoE architecture is genuinely clever, but in 2026, dense models have closed the quality gap while being far more VRAM-efficient. Choose Mixtral for speed-at-quality when you have the memory budget. Choose dense models when VRAM is your constraint — which, for most people reading this, it is.

For GPU buying recommendations at every price point, see our GPU Buying Guide and Used RTX 3090 Buying Guide.

Source: https://insiderllm.com/guides/mixtral-8x7b-8x22b-vram-requirements/

Free guides for running AI locally