Mistral & Mixtral Guide: Every Model Worth Running Locally

February 3, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Mistral Nemo 12B is the Mistral model to run in 2026 — 128K context, Apache 2.0 license, fits on a 16GB GPU at Q4, and holds its own against Llama 3.1 8B despite being larger. The original Mistral 7B is dated but still useful on tight VRAM (~4GB at Q4). Mixtral 8x7B pioneered MoE for local AI but needs ~26-32GB even quantized — most people are better off with Qwen 3 32B on the same hardware. Codestral is a solid code model but has a non-commercial license, so Qwen 2.5 Coder is usually the better choice. Mistral was the benchmark leader in 2023-2024; in 2026, they're a solid option but no longer the default recommendation over Llama 3 or Qwen 3.

📚 **More on this topic:** Llama 3 Guide · Qwen Models Guide · DeepSeek Models Guide · VRAM Requirements

Mistral AI burst onto the scene in late 2023 with a 7B model that embarrassed much larger competitors. Mixtral 8x7B introduced Mixture of Experts to the open-source world. For a while, Mistral was the default answer to “what should I run locally?”

That’s no longer true. Llama 3 and Qwen 3 have caught up and passed Mistral on most benchmarks. But Mistral models are still solid — particularly Mistral Nemo 12B with its 128K context window — and understanding the lineup helps you make informed choices.

This guide covers what’s worth running, what’s been superseded, and where Mistral still makes sense.

The Mistral Lineup

Every Mistral model relevant for local use:

Model	Parameters	Context	VRAM (Q4)	License	Status
Mistral 7B	7B	8K	~4 GB	Apache 2.0	Dated but usable
	12B	128K	~8 GB	Apache 2.0	**Recommended**

Model	Parameters	Context	VRAM (Q4)	License	Status
Mistral Nemo 12B					
Codestral 22B	22B	32K	~12 GB	Non-commercial	Good, but license limits use
Mixtral 8x7B	46.7B (12.9B active)	32K	~26-32 GB	Apache 2.0	Superseded by dense models
Mixtral 8x22B	141B (39B active)	64K	~66 GB	Apache 2.0	Requires serious hardware

Not locally runnable: Mistral Small, Medium, Large, and Large 2 are API-only or require datacenter hardware. Skip them for local use.

Why Mistral Still Matters

Mistral occupies a specific niche in 2026:

Apache 2.0 licensing — Unlike some competitors, Mistral's open models are genuinely open. No usage restrictions, no commercial limitations (except Codestral).

European AI — Based in Paris, Mistral represents European AI development. If you care about geographic diversity in AI, that matters.

Mistral Nemo's context — 128K tokens with Apache 2.0 licensing is still relatively rare. The collaboration with NVIDIA produced a model specifically optimized for single-GPU deployment.

Legacy compatibility — Many existing projects and fine-tunes are built on Mistral. If you're using a specific Mistral-based model, understanding the base helps.

What Mistral doesn't have: benchmark leadership. Qwen 3 and Llama 3 outperform Mistral at most comparable sizes. That's the honest assessment.

Mistral 7B: The Original

Released September 2023, Mistral 7B was a breakthrough — a 7B model outperforming Llama 2 13B on most benchmarks. It introduced sliding window attention for efficient long-context handling.

Specs

- **Parameters:** 7 billion
- **Context:** 8K (sliding window)
- **VRAM:** ~4 GB at Q4_K_M, ~15 GB at FP16
- **License:** Apache 2.0

The Reality in 2026

Mistral 7B is dated. The benchmarks that impressed in 2023 are now beaten by newer 7-8B models:

Model	MMLU	Notes
Qwen 3 8B	73.8%	Current leader
Llama 3.1 8B	73.0%	Strong all-rounder
Mistral 7B	~62%	Shows its age

When Mistral 7B still makes sense:

- You have exactly 4 GB VRAM and need the smallest viable model
- You're running a fine-tune specifically based on Mistral 7B
- Cost matters more than quality (62.5% cheaper than Llama 3 on AWS Bedrock)

When to use something else:

- Any scenario where you can fit Qwen 3 8B or Llama 3.1 8B instead

```
ollama run mistral:7b
```

Mistral Nemo 12B: The Current Pick

Mistral Nemo is a collaboration between Mistral AI and NVIDIA, released July 2024. It's specifically designed for efficient single-GPU deployment with a 128K context window.

Specs

- **Parameters:** 12 billion
- **Context:** 128K tokens
- **Tokenizer:** Tekken (trained on 100+ languages)
- **VRAM:** ~8 GB at Q4, ~24 GB at FP16
- **License:** Apache 2.0

What Makes Nemo Special

Quantization-aware training — Mistral trained Nemo with FP8 inference in mind. Unlike most models where quantization hurts quality, Nemo maintains performance at reduced precision. This is huge for local deployment.

128K context at 12B — Most 12B models top out at 32K. Nemo's 128K window means you can feed it entire codebases, long documents, or extended conversations without truncation.

Multilingual strength — The Tekken tokenizer handles 100+ languages better than most competitors. Strong on English, French, German, Spanish, Italian, Portuguese, Chinese, Japanese, Korean, Arabic, and Hindi.

Benchmarks vs Competition

Benchmark	Mistral Nemo 12B	Llama 3.1 8B
MMLU	68%	73%
HellaSwag	83.5%	—
TriviaQA	73.8%	—
Coding (practical)	Strong	Strong

Llama 3.1 8B beats Nemo on MMLU despite being smaller. But Nemo has 128K context vs Llama's 128K, and Nemo's quantization behavior is better. For long-context tasks on limited hardware, Nemo is often the better choice.

Setup

```
ollama run mistral-nemo
```

Custom Modelfile for long context:

```
FROM mistral-nemo

PARAMETER num_ctx 32768
PARAMETER temperature 0.7

SYSTEM "You are a helpful assistant with access to a large context window."
```

Mixtral 8x7B: The MoE Pioneer

Mixtral 8x7B introduced Mixture of Experts (MoE) to mainstream local AI. Eight expert networks, two active per token, for a total of 46.7B parameters but only 12.9B active during inference.

How MoE Works

Instead of one massive feedforward network, Mixtral has 8 smaller expert networks. A router decides which 2 experts handle each token. You get the knowledge of a 47B model with the compute cost of a 13B model.

The catch: **you still need VRAM for all 47B parameters**. MoE saves compute, not memory.

VRAM Requirements

This is where Mixtral becomes impractical for most local users:

Precision	VRAM	Hardware
FP16	~90 GB	2x A100 80GB
Q5_0	~32 GB	Beyond single RTX 3090/4090
Q4_K_M	~26-28 GB	Still exceeds 24 GB

Even at aggressive quantization, Mixtral 8x7B doesn't fit on a single consumer GPU. You need dual RTX 3090s, a Mac with 48GB+ unified memory, or substantial CPU offloading (which kills speed).

Performance When You Can Run It

If you have the hardware:

Hardware	Speed
Dual RTX 4090	~59 tok/s
Dual RTX 3090	~54 tok/s
Mac M2 Ultra	~44 tok/s
Mac M3 Max	~22 tok/s

Quantization tip: Use Q5_0, not K-quants. Community testing shows Q5_K_M causes more rambling and inconsistency with Mixtral specifically. Q5_0 performs better, especially for coding tasks.

The Honest Take

In 2024, Mixtral was exciting. In 2026, if you have 32 GB+ VRAM:

- **Qwen 3 32B** (dense) gives similar or better quality at ~20 GB
- **Llama 3.3 70B** at heavy quantization gives better quality
- **DeepSeek R1-Distill-32B** beats Mixtral on reasoning tasks

Mixtral pioneered MoE for local AI, but dense models have caught up. Unless you specifically need MoE behavior or have a Mixtral-based fine-tune, the hardware requirements aren't justified by the performance.

```
ollama run mixtral:8x7b
```

Codestral: The Code Specialist

Codestral is Mistral's dedicated coding model — 22B parameters, 32K context, trained on 80+ programming languages.

Specs

- **Parameters:** 22 billion
- **Context:** 32K tokens
- **Languages:** 80+ including Python, Java, C, C++, JavaScript, Bash
- **Fill-in-the-middle:** Supported
- **License: Mistral AI Non-Production License** (commercial license available separately)

The License Problem

Codestral's biggest issue isn't performance — it's licensing. The Non-Production License means you can't use it for commercial projects without paying Mistral separately. Compare:

Model	License	HumanEval
Qwen 2.5 Coder 32B	Apache 2.0	88.4%
Codestral 22B	Non-commercial	81.1%
DeepSeek Coder 33B	Permissive	77.4%

Qwen 2.5 Coder is fully open, smaller, AND scores higher on HumanEval. For most users, there's no reason to accept Codestral's license restrictions.

When Codestral Makes Sense

- **Long-range code completion:** Codestral's 32K context gives it an edge on RepoBench (34.0% vs 28.4% for DeepSeek) — useful when you need to understand large codebases
- **Personal projects:** If you're not building commercial software, the license doesn't matter
- **IDE integration:** Codestral works well with Continue and similar tools

```
ollama run codestral
```

The Better Choice

For most coding needs, Qwen 2.5 Coder is the recommendation. Higher benchmarks, no license restrictions, and active development.

Mixtral 8x22B: Serious Hardware Only

The big Mixtral — 141B total parameters, 39B active per token, 64K context.

VRAM Requirements

Precision	VRAM
FP16	~260-300 GB
Q4	~66 GB

This is multi-A100 or 4x RTX 4090 territory. For home users, it's not realistic.

If you have access to this hardware, Mixtral 8x22B delivers:

- 90% on GSM8K (maj@8)
- Strong coding (HumanEval)
- Competitive with much larger models

But at this VRAM tier, you're comparing against Llama 3.3 70B and considering the full DeepSeek R1. The MoE advantage doesn't justify the complexity for most use cases.

VRAM Requirements Table

Complete reference for all Mistral models:

Model	Q3_K_M	Q4_K_M	Q5_K_M	Q8_0	FP16
Mistral 7B	~3 GB	~4 GB	~5 GB	~7 GB	~15 GB
Mistral Nemo 12B	~6 GB	~8 GB	~9 GB	~13 GB	~24 GB
Codestral 22B	~10 GB	~12 GB	~15 GB	~23 GB	~44 GB
Mixtral 8x7B	~22 GB	~26 GB	~32 GB	~48 GB	~90 GB

Model	Q3_K_M	Q4_K_M	Q5_K_M	Q8_0	FP16
Mixtral 8x22B	~50 GB	~66 GB	~80 GB	~140 GB	~260 GB

What to run on your GPU:

GPU	VRAM	Best Mistral Model
RTX 3060	8/12 GB	Mistral 7B or Nemo Q4
RTX 4060 Ti	16 GB	Mistral Nemo Q5
RTX 3090/4090	24 GB	Mistral Nemo FP16 or Codestral Q4
2x RTX 3090	48 GB	Mixtral 8x7B Q4
Mac M2/M3 Ultra	64-192 GB	Mixtral 8x7B or 8x22B

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

Mistral vs the Competition

Honest positioning of Mistral models against alternatives:

At 7-8B (Tight VRAM)

Model	MMLU	Best For
Qwen 3 8B	73.8%	**Overall winner**
Llama 3.1 8B	73.0%	Fine-tune ecosystem
Mistral 7B	~62%	Legacy compatibility, lowest cost

Verdict: Mistral 7B is outclassed. Use Qwen 3 8B or Llama 3.1 8B unless you have a specific reason not to.

At 12-14B (Mid-Range)

Model	Context	VRAM (Q4)	Notes
Mistral Nemo 12B	128K	~8 GB	Best context-per-VRAM
Qwen 3 14B	32K+	~10 GB	Better benchmarks

Model	Context	VRAM (Q4)	Notes
Llama 3.1 8B	128K	~5 GB	Smaller but competitive

Verdict: Mistral Nemo is competitive here. If you need 128K context on limited VRAM, it's a solid choice. For raw capability, Qwen 3 14B edges it out.

At 30-50B (High-End Consumer)

Model	Type	VRAM (Q4)	Notes
Qwen 3 32B	Dense	~20 GB	**Best on 24 GB**
Mixtral 8x7B	MoE	~26-32 GB	Needs dual GPU
DeepSeek R1-32B	Dense	~18 GB	Best for reasoning

Verdict: Mixtral loses to dense models on practical deployments. The MoE overhead isn't worth it anymore.

For Coding

Model	HumanEval	License	Recommendation
Qwen 2.5 Coder 32B	88.4%	Apache 2.0	**Best overall**
Codestral 22B	81.1%	Non-commercial	Only if license is OK
DeepSeek Coder 33B	77.4%	Permissive	Decent alternative

Verdict: Qwen 2.5 Coder wins. Codestral's license kills it for most use cases.

Setup Guide

Quick Start

```
# The recommendation for most users
ollama run mistral-nemo

# Original 7B for tight VRAM
ollama run mistral:7b
```

```
# MoE if you have the hardware
ollama run mixtral:8x7b

# Coding (check license first)
ollama run codestral
```

Modelfile for Mistral Nemo

```
FROM mistral-nemo

# Maximize context (adjust based on VRAM)
PARAMETER num_ctx 32768

# Slightly lower temperature for consistency
PARAMETER temperature 0.6

# Stop sequences
PARAMETER stop "<|endoftext|>"
PARAMETER stop "</s>"

SYSTEM "You are a helpful, concise assistant."
```

```
ollama create my-nemo -f Modelfile
ollama run my-nemo
```

Mixtral-Specific Settings

If running Mixtral, avoid K-quants:

```
# Pull Q5_0 specifically (community recommendation)
ollama pull mixtral:8x7b-instruct-v0.1-q5_0
```

Bottom Line

Mistral was the model to run in 2023-2024. In 2026, the landscape has shifted:

Still worth running:

- **Mistral Nemo 12B** — Best-in-class for 128K context on limited VRAM, Apache 2.0
- **Mistral 7B** — Only if you have <6 GB VRAM or need the absolute cheapest option

Superseded:

- **Mixtral 8x7B** — Dense models (Qwen 3 32B) give better value at the same VRAM tier
- **Codestral** — Qwen 2.5 Coder beats it AND has a better license

Not practical for local:

- **Mixtral 8x22B** — Datacenter hardware required
- **Mistral Large/Small/Medium** — API-only

If you're starting fresh, Qwen 3 or Llama 3 should be your first choice. If you specifically need long context on modest hardware, Mistral Nemo earns its place.

```
# The Mistral model worth running
ollama run mistral-nemo
```

Source: https://insiderllm.com/guides/mistral-mixtral-guide/