MISTRAL AI

Mistral Models

Efficient · Open · Enterprise Which Model to Use When

Mistral Large 2 | Mistral Small 3 | Codestral | Mixtral 8x22B | Mistral NeMo | Pixtral Large

A Practical Decision Guide · 2025 | Open Weights & Commercial API

The Mistral Model Family

Six models spanning frontier intelligence, coding, vision, and efficient edge deployment

🔥 FLAGSHIP & LANGUAGE MODELS

Most Capable

Mistral Large 2

Frontier-class reasoning & instruction following

- Top-tier reasoning & analysis
- 32K context, 123B params
- Strong multilingual (29 langs)
- Function calling & JSON mode

MoE Powerhouse

Mixtral 8x22B

Sparse Mixture-of-Experts 141B total params

- Active 39B params per token
- Beats LLaMA 2 70B on most evals
- Long 64K context window
- Best open-weight efficiency

Speed & Value

Mistral Small 3

Fast, affordable everyday language tasks

- 3× cheaper than Large
- Sub-100ms response latency
- Classification & extraction
- Ideal for high-volume pipelines

⚡ SPECIALIZED MODELS

Best for Code

Codestral

Purpose-built 22B coding model

- 80+ programming languages
- Fill-in-middle (FIM) support
- Repo-level code completion

Vision+Language

Pixtral Large

Frontier multimodal images & text

- Native image understanding
- Chart, doc & screenshot analysis
- Multiple images per prompt

Edge & On-Prem

Mistral NeMo

12B model optimized for local deployment

- Runs on a single GPU
- Apache 2.0 self-hostable
- Strong reasoning at 12B scale

Which Mistral Model Should You Use?

Match your task across language, code, vision, and edge deployment open weights & enterprise API

USE CASE / SCENARIO	MODEL	WHY IT FITS
Complex reasoning, Q&A & enterprise tasks	Large 2	Frontier 123B model top-tier instruction following & multilingual
High-efficiency open-weight deployments	Mixtral 8x22B	Sparse MoE frontier quality with only 39B active params per call
High-volume, cost-sensitive API workloads	Small 3	3× cheaper than Large fast latency ideal for pipelines at scale
Code generation, completion & review	Codestral	Purpose-built coding model 80+ languages, beats GPT-4 on HumanEval
Image & document understanding (multimodal)	Pixtral Large	Native vision+language reads charts, screenshots & multi-image prompts
On-device, local & private deployments	NeMo 12B	Runs on a single GPU Apache 2.0, fully self-hostable, no API needed
Agentic workflows with function calling	Large 2	Native JSON mode + function calling best Mistral model for agents

💡 Pro tip: Default to Mistral Small 3 for cost-efficient tasks, Large 2 for frontier quality, Codestral for code. Use NeMo 12B to self-host on a single GPU with Apache 2.0 license.