Local AI Models Decoded

RUN FRONTIER-QUALITY AI ON YOUR OWN HARDWARE. FREE. PRIVATE. FOREVER.

Plex Consulting

Model Naming Decoded

Example: `meta-llama-3.1-70b-instruct-q4_k_s`

Component	Meaning
`meta-llama`	Who made it (Meta)
`3.1`	Version (like iPhone 15 vs 16)
`70b`	70 billion parameters (brain size)
`instruct`	Tuned for conversation (vs "base" for raw completion)
`q4_k_s`	Compression level (quantisation)

Quantisation Guide

Level	Quality	Use Case
`f16`	Full precision	Research, maximum quality, 2x RAM needed
`q8`	Near-perfect	Best balance of quality and size
`q6_k`	Excellent	Slightly smaller, negligible quality loss
`q4_k_m`	Very good	Most popular choice — great quality, half the RAM
`q4_k_s`	Good	Smaller variant, still very capable
`q2_k`	Usable	Maximum compression, noticeable quality drop

Size Guide

Size	Hardware	RAM Needed	Capability
1—3B	Phone / RPi	2—4 GB	Basic tasks, autocomplete
7—8B	Any laptop	8 GB	Good for most tasks
13B	Good laptop	16 GB	Very capable, strong reasoning
32—34B	Workstation	32 GB	Excellent across the board
70B	Server / multi-GPU	48—64 GB	Near frontier quality
405B	Data centre only	200+ GB	Frontier performance

Key Model Families

Model	Maker	Strength
Llama 3.1 / 3.2	Meta	General purpose, coding
Mistral / Mixtral	Mistral AI	Fast, multilingual
Qwen 2.5 / 3	Alibaba	Coding, math
Gemma 2	Google	Lightweight, efficient
Phi-3 / 4	Microsoft	Small but punchy
DeepSeek R1	DeepSeek	Reasoning, math
nomic-embed-text	Nomic AI	Embeddings (search / RAG)
bge-reranker	BAAI	Filters results to top matches

Ollama Commands

Command	What It Does
`ollama pull llama3.1:8b`	Download a model
`ollama run llama3.1:8b`	Start chatting with it
`ollama list`	See all installed models
`ollama serve`	Start the API server
`ollama rm modelname`	Delete a model
`ollama show llama3.1:8b`	View model details

Quick Decision Guide

- **Just starting?** — `ollama pull llama3.1:8b` — works on any machine with 8 GB RAM.
- **Need coding help?** — Qwen 2.5 Coder (7B or 32B) — purpose-built for code generation.
- **Building RAG / search?** — `nomic-embed-text` for embeddings + `bge-reranker` for filtering.
- **Maximum quality locally?** — Llama 3.1 70B (q4_k_m) — needs 48 GB RAM but rivals cloud models.
- **Sensitive data?** — any local model. Data never leaves your machine. Zero API calls.