AI Tips Weekly
Practical, sourced AI tips. No hype.

Which local AI fits your RAM?

Sizes checked on ollama.com on Oct 17, 2026

Rule of thumb

- The download size is roughly the memory a model needs to load.
- You also need room for context and for your OS and apps, so leave a few GB spare.
- On a Mac, the GPU can't use all of your RAM: Apple's figures for M1 Pro/Max MacBook Pros are **21 GB usable on 32 GB and 48 GB on 64 GB**.

Your memory	What fits	Try first (Ollama tag — download size — maker)
8 GB	Small models (about 2–4B), short context	`llama3.2:3b` — 2.0 GB — Meta `phi4-mini` — 2.5 GB — Microsoft `gemma3:4b` — 3.4 GB — Google
16 GB	About 7–12B	`gemma4:e4b` — 6.6–9.5 GB — Google DeepMind `ministral-3:8b` — 6.0 GB — Mistral AI `gemma4:12b` — 7.7–8.0 GB — Google DeepMind
24–32 GB	About 20–30B, or mixture-of-experts	`gpt-oss:20b` — 14 GB — OpenAI `mistral-small3.2:24b` — 15 GB — Mistral AI `gemma4:26b` — 16–19 GB — Google DeepMind **tight on 24 GB**
64 GB+	30B+ with longer context	`gemma4:31b` — 19–20 GB — Google DeepMind · plenty of room for long context `llama3.1:70b` — 43 GB — Meta **tight on a 64 GB Mac**

Beyond 64 GB: `gpt-oss:120b` — 65 GB — OpenAI **needs more than 64 GB** (Ollama: fits a single 80 GB GPU). **Sizes change, so check the listing before you download.**

Pick a tool

Tool	Best for	Notes
Ollama	The simplest start	Install, then `ollama run gemma4:e4b`. Serves a local API at `localhost:11434`.
LM Studio	Clicking instead of typing	Desktop app with a model browser. Mac needs Apple Silicon and macOS 14+; Windows x64 needs AVX2.
llama.cpp	Tinkerers	The open-source engine behind many apps. Runs GGUF files on CPU, NVIDIA (CUDA), Apple Metal and Vulkan.
MLX	Apple Silicon Macs	Apple's framework; `mlx-lm` runs models natively on a Mac. Ollama also lists `-mlx` tags.

When "local" really is private and offline

- **Private:** a downloaded model running on your own machine. Ollama: "We don't see your prompts or data when you run locally." LM Studio: "Nothing you enter into LM Studio when chatting with LLMs leaves your device."
- **Not local:** Ollama tags ending in `cloud` (for example `gemma4:cloud`) and Ollama's web search run on Ollama's servers. To stay local-only, set `OLLAMA_NO_CLOUD=1` and restart Ollama.
- **Offline:** only after the download. Searching for and downloading models needs the internet.
- **Watch out:** if you expose the server to your network (Ollama binds to `127.0.0.1` by default), or connect agents and plugins that use the web, data can leave your machine.

Also popular: Alibaba's Qwen 3.5 (`qwen3.5:4b` 3.3–4.0 GB, `qwen3.5:9b` 6.6–8.9 GB, `qwen3.5:27b` 17–20 GB). See the safety note below.

> **Safety note: Qwen and DeepSeek (Chinese providers).** Qwen (Alibaba) and DeepSeek come from Chinese companies. Running the weights offline keeps your prompts on your machine, but built-in censorship stays: one study found DeepSeek R1 refuses some China-sensitive questions even when run locally. Their cloud apps follow the provider's rules; DeepSeek's privacy policy says personal data is stored in the PRC. Download only from official sources.

Sources (checked Oct 17, 2026): ollama.com/library (gemma4, gemma3, llama3.2, llama3.1, phi4-mini, ministral-3, mistral-small3.2, gpt-oss, qwen3.5)
docs.ollama.com/faq
docs.ollama.com/cloud
lmstudio.ai/docs/app/system-requirements
lmstudio.ai/docs/app/offline
github.com/ggml-org/llama.cpp
github.com/ml-explore/mlx-lm
developer.apple.com/videos/play/tech-talks/10580
arxiv.org/abs/2505.12625
DeepSeek privacy policy (cdn.deepseek.com)