THE LOCAL AI CHEAT SHEET · OFFLINE

Run a real AI on your laptop. Free. Private. Offline.

You don't need a $1,000 GPU — you need to understand **three numbers**. Get them right and you'll pick the perfect model in ten seconds, then run it with one command. This is the whole thing on five pages.

Measured on an Apple Silicon laptop: Qwen2.5-Coder 7B — 4.7 GB on disk, ~48 tokens/sec, 100% on-device, $0. No API key, no bill, no data leaving the machine.

THE ONLY THING THAT MATTERS

It's not your CPU. It's memory — and 3 numbers.

1

Memory ceiling

~0.6 GB of RAM per 1B parameters. A 7B model ~ 4–5 GB; a 14B ~ 8–10 GB. You can't use all your RAM — the OS needs a few GB. The whole model must fit at once.

2

Quantization

Rounds the model's numbers smaller — like a compressed JPEG vs a RAW file. "Q4" cuts memory ~70% with barely any quality loss. Ollama downloads a sensible quant by default.

3

Context window

Your chat + the model's replies also live in RAM, on top of the model. Long chats / pasted files eat memory fast and can crash it. A smaller model with headroom beats a bigger one choking.

The formula: model size + context + system < your RAM. Stay under it and everything stays smooth.

PART 1 · PICK YOUR MODEL BY RAM

Match the model to your memory.

YOUR RAM	MODEL SIZE	WHAT TO RUN	BEST FOR
8 GB	3B – 4B	**Gemma 4B · Qwen 3B · Llama 3B**	Chat, simple code
16 GB	7B – 8B	**Qwen2.5-Coder 7B » Llama 8B**	Real coding — sweet spot
24–32 GB	14B	**Qwen2.5-Coder 14B · Gemma 12B**	Serious work
48 GB+	30B+	**Qwen 32B · larger MoE models**	The big leagues

» **Top pick for most people (16 GB):** Qwen2.5-Coder 7B — punches far above its size for code, and it's the one measured in this guide.

Why "go one size down" usually wins

A 7B that fits entirely in fast memory beats a 14B that spills to disk and crawls. Speed and headroom matter more than raw parameter count for everyday work.

PART 2 · SET IT UP IN ~5 MINUTES

Three steps. One command each.

1 · Install Ollama

Download the free app from ollama.com — install like any other program. (Mac / Windows / Linux.)

```
ollama.com
```

2 · Run a model

First run downloads a few GB, once. After that it's instant and fully offline.

```
$ ollama run qwen2.5-coder:7b
```

3 · Just type

Ask it to write code, explain an error, refactor. That's the whole learning curve.

```
$ explain this error / refactor this
```

HANDY COMMANDS

List installed models	`ollama list`
Swap to another model	`ollama run gemma`
Leave the chat	`/bye`
See what's loaded in RAM	`ollama ps`
Remove a model	`ollama rm`

PART 3 · TOY → DAILY DRIVER

Four upgrades, all free.

Plug into your editor

Continue or the Cline extension drops a local model into VS Code — free autocomplete + chat where you work.

On a Mac? Use MLX builds

Models tagged "MLX" are tuned for Apple Silicon and run noticeably faster than the generic build.

Turn on KV cache quant

Compresses the context memory so you fit much longer chats in the same RAM. Set OLLAMA_KV_CACHE_TYPE=q8_0 + OLLAMA_FLASH_ATTENTION=1.

Pick the right size

When in doubt, go one size DOWN. A 7B that runs fast beats a 14B that swaps to disk and crawls.

Got value?

Comment **OFFLINE** on the video and we'll send this straight to you · also at **hyperautomationlabs.co/free/offline**

OFFLINE

Go deeper — the playbooks behind this channel

The Complete Claude Code Guide — $19
gumroad.com/l/claude-code-guide

The OpenAI Codex Guide — $19
gumroad.com/l/codex-guide

Claude for Cowork & Sales — $19
gumroad.com/l/claude-cowork-sales

Claude Certified Architect — Prep Kit — $29
gumroad.com/l/claude-certified-architect-prep

Follow Hyperautomation Labs on **YouTube**, **Instagram** and **Facebook** — we do the boring testing so you get the simple, honest version.