

# WHAT IS LOCAL AI?

OWNED, NOT RENTED

**Local AI** means running open-weight AI models on hardware you own. No API call. No per-token billing. No vendor in the loop. Your prompt leaves your laptop, runs on your hardware, and the answer comes back, all in the same room.

The phrase that makes this possible is "**open-weight**." A model's **weights** are the billions of numerical settings inside the neural network: the actual "knowledge" it learned during training. Most frontier labs lock those weights behind their API. Open-weight labs publish them, so you can download them, run them on a laptop, fine-tune them on your own data, quantize them to fit cheaper hardware, or run them offline on a plane. That single distinction, locked weights vs. published weights, is the **entire difference** between frontier and local. Everything else flows from it.

## The local model ecosystem in 2026

| FAMILY         | MAKER / LICENSE      | SIZES YOU WILL SEE                    |
|----------------|----------------------|---------------------------------------|
| Gemma 4        | Google · Apache 2.0  | 12B 26B-A4B 31B E2B / E4B             |
| Qwen 3.5 / 3.6 | Alibaba · Apache 2.0 | 27B 35B-A3B 122B-A10B 397B-A17B       |
| GLM-5.2        | Z.ai / Zhipu · MIT   | 744B-A40B 1M context coding-first 5.2 |
| gpt-oss        | OpenAI · Apache 2.0  | 20B 120B OpenAI-API compatible        |

On the name "gpt-oss": it stands for "GPT open-source software." The weights are real and genuinely capable, and being API-compatible means existing OpenAI code swaps over by changing one URL.

## Why people chase local AI

- 1 Privacy.** Prompts never leave your hardware. For a lawyer, doctor, or journalist that is the requirement, not a nice-to-have.
- 2 Cost ceiling.** Once you own the box, one more query costs **fractions of a cent** of electricity. Local bills are flat.
- 3 No rate limits.** Your hardware, your limits. No vendor throttling you during business hours.
- 4 No vendor lock-in.** Weights are weights. Swap Qwen for Gemma by changing one line in your config.
- 5 Offline operation.** Works on a plane, in a bunker, on a boat. Anywhere with electricity and no internet.
- 6 Customization.** Fine-tune on your own proprietary data without asking anyone's permission.

local · the one-line truth

**\$ Local = you paid the hardware bill. The inference is free.**

The symmetric mirror of the frontier line (page 5). Frontier: someone else paid training, you pay inference. Local: **you paid hardware, inference is essentially free.** The trade-off is that your hardware caps your capability, which is exactly what page 12 is about.