

# Run your own private AI.

Right now, every word you type into AI is sitting on **someone else's servers**. Your ideas. Your numbers. Your clients.

That is your edge, leaving the building. This guide shows you how to take it back — with AI that runs on your own machine. Private. Unlimited. Yours.

The whole thing takes about 20 minutes. No coding. A Mac or Windows computer is all you need, and every tool here is free. Follow the four steps. Each one has a check so you always know it worked.

by Drey Thomas • [dreythomas.com/private-ai](https://dreythomas.com/private-ai)

1

## Install Ollama

Ollama is the free app that runs AI on your own computer. Go to [ollama.com](https://ollama.com), download it for your machine, and install it like any other app.

**You'll know it worked:** a small llama icon shows up in your menu bar (Mac) or system tray (Windows). **If not:** reopen the app you downloaded and let it finish installing.

2

## Turn it on

Open the Terminal app (Mac) or Command Prompt (Windows) and type one line. It downloads the AI once, then runs it right on your machine.

```
> ollama run gemma3
```

**You'll know it worked:** after a download bar, you'll see a blinking `>>>` prompt waiting for you. **If it says "command not found":** Ollama isn't installed yet — go back to step 1.

3

## Talk to it

Type any question and hit enter. That answer came from your own computer — nothing left the building. Type `/bye` to stop. Want a clean chat screen instead of the terminal? Install `Open WebUI` (free) for a ChatGPT-style window.

**You'll know it worked:** you get a full answer back, even with your wifi turned off. That's the proof it's truly running on your machine.

4

## Put it to work

Hand it the everyday busywork — private, and with no limits:

- First drafts, rewrites, and outlines
- Summaries of long notes, calls, or documents
- Sorting and cleaning up messy lists
- Reading and pulling text out of screenshots

**The smart split:** run the everyday work on your private AI, and keep a paid top model only for the heavy, high-stakes swings. Most of your day stays private; you bring in the firepower when it counts.

Tools: Ollama (ollama.com) • models gemma3 / llama3.2 / qwen3 • Open WebUI (openwebui.com). All free, all open source.

## You did the easy part. Here's the real part.

Getting a model running is 20 minutes. Turning it into a system that actually runs your business is where the edge lives — and it's the part most people get wrong:

- **The right models** for your actual work, sized to your machine, so it's fast not sluggish.
- **Your workflows wired in** — your docs, your templates, your repeat tasks, not a blank prompt.
- **Security done right** so client and financial data stays locked to your machine.
- **A clean chat setup** your whole team can use, no terminal required.

DONE FOR YOU

**Or skip all of it. I'll set it up for you.**

I install private AI systems for founders — dialed in for your business, your data, your team. You get the edge without spending your afternoon in a terminal.

Book a call:

[discoverycall.dreythomas.com](https://discoverycall.dreythomas.com)

Run your own private AI · a guide by Drey Thomas · [dreythomas.com/private-ai](https://dreythomas.com/private-ai)