NIB KNOWS
AI EXPLAINER

Comment LOCAL
episode 1 setup guide

Run a free AI model on your own computer.



A short beginner guide to Ollama and Gemma 4 (e2b). No account or API key needed for local use. Official steps checked 25 Nov 2026.

1 Install Ollama

Windows

Windows 10 (22H2) or newer. The easy way: download **OllamaSetup.exe** from `ollama.com/download` and run it — no Administrator needed.

Or in PowerShell:

```
irm https://ollama.com/install.ps1
| iex
```

Mac

macOS Sonoma (14) or newer. Download **Ollama.dmg** from `ollama.com/download`, then drag it to Applications.

Or in Terminal:

```
curl -fsSL
https://ollama.com/install.sh | sh
```

Apple silicon uses the GPU; Intel Macs run on the CPU.

Linux

In a terminal:

```
curl -fsSL
https://ollama.com/install.sh | sh
```

If the app isn't running yet, start the server with `ollama serve`.

2 Run the model

Open a terminal (PowerShell on Windows) and type this exact command:

```
❯ ollama run gemma4:e2b
```

The first time, it downloads the model — about 7–8 GB — then a chat starts. Ask a question, press Enter, and wait for the reply. Type `/bye` when you want to leave.

3 The 8 GB catch

This starter model wants about **8 GB of graphics memory** (VRAM), or unified memory on a Mac. That's the official recommendation.

Got less? You can still try it. Ollama can use ordinary system RAM instead — it works, just slower. You'll also want about 8 GB of free disk space for the download. Bigger models need a much beefier machine.

Short on memory?

Run it anyway and give it a moment. If replies crawl, that's the RAM fallback doing its best — not a broken install.

4 Try these first

PROMPT 1

Explain how a heat pump works, in two short paragraphs, as if I've never heard of one.

PROMPT 2

Give me a simple 20-minute dinner I can cook with pasta, a tin of tomatoes, and whatever's in the fridge.

5 Where your messages go

If you run it locally, once the model is downloaded, it stays on your computer. Ollama's own FAQ says: **"Ollama runs locally. We don't see your prompts or data when you run locally."** The local server listens on your machine only, by default (`localhost:11434`).

This guide is for **local** use only. Cloud models (names ending in `:cloud`) are different — those requests are processed on Ollama's servers. Skip anything with `:cloud` if you want to keep it on your computer.

Official pages (checked 25 Nov 2026): ollama.com/download · ollama.com/library/gemma4:e2b docs.ollama.com/quickstart · docs.ollama.com/faq · Windows / macOS / Linux docs on the same site.

NIB KNOWS
A free local setup guide. No income claims. Gemma 4 is a Google model; Ollama is the app that runs it.