YESBRAINER

Getting Started with Ollama

Your Complete Onboarding Manual

1 Core · 7 Lenses · ∞ Forever · 8+ Tools

Version 2.0 · An Offline.Ltd product

1. What is Ollama?

You have probably used AI assistants before — tools like ChatGPT, Gemini, or Claude that run in your web browser and send everything you type to a company's servers. Ollama is different. It lets you run AI models directly on your own computer, like any other application.

Think of it this way: if a cloud AI assistant is like going to a restaurant, Ollama is like having a private chef in your own kitchen. The ingredients never leave your house.

In practical terms, Ollama is a free, open-source program that you install on your Mac, Windows, or Linux computer. Once installed, it runs AI language models locally — meaning the AI brain lives on your hard drive, not on someone else's server. It requires no account, no subscription, and no internet connection once set up.

What Ollama does for Yesbrainer

Job	What It Does	Model
Thinking	Reads your documents, analyses scenarios through multiple lenses, finds tensions, identifies blind spots, runs adversarial attacks	llama3.2
Understanding	Turns your text into mathematical patterns so the Source Vault can find the most relevant passages	nomic-embed-text

You do not need to understand how these models work. Ollama handles everything. Just install it and pull the models.

2. Why Offline AI Matters for Thinkers

Your ideas stay yours

When you paste your research into a cloud AI, that text travels to a company’s servers. With Ollama, your work never leaves your computer. Not a single word. This isn’t just a feature — it’s the precondition for honest thought.

No one is watching your thinking process

Early thinking is messy. Half-formed. Sometimes offensive or absurd. That’s the point — you’re stress-testing ideas, not publishing them. A local AI has no content policy applied to your private analysis.

You work on your schedule

Ollama works on a plane. It works in a cabin with no Wi-Fi. It works at 3am when cloud servers are busy. Once the models are downloaded, your thinking partner is available whenever you are.

No ongoing cost

Cloud AI subscriptions cost €20–40 per month. Ollama is free. The models are free. You can run as many analyses as your computer can handle, forever.

You choose the brain

With Ollama, you can try different models and pick the one that works best for your domain. Yesbrainer lets you switch models in any lens with a single text field.

3. What You Will Need

Component	Minimum	Recommended
Operating System	macOS 12+, Windows 10+, or Linux	Latest version of your OS
RAM (Memory)	8 GB	16 GB or more
Disk Space	5 GB free	15 GB free (for multiple models)
Processor	Any modern CPU	Apple Silicon (M1–M4) or GPU with 8GB+ VRAM
Browser	Chrome 90+, Firefox 88+, Edge 90+, Safari 15+	Chrome or Firefox, latest

If your computer was purchased in the last five years, it almost certainly meets the minimum requirements.

A note on speed

Local AI is slower than cloud AI. Responses typically take 5 to 30 seconds depending on your hardware. Yesbrainer streams the response as it generates, so you will see words appearing in real time rather than waiting for the full answer.

4. Installing Ollama

Installation takes about two minutes.

macOS

1. Download Ollama from **ollama.com** — it detects your platform automatically.
2. Open the .dmg file, drag Ollama to Applications.
3. Launch Ollama from Applications or Spotlight. A llama icon appears in your menu bar.
4. Open Terminal, type: `ollama --version`

Windows

1. Download from **ollama.com** for Windows.
2. Run OllamaSetup.exe and follow the prompts.
3. Open Command Prompt (Win+R, type cmd). Type: `ollama --version`

Linux

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

Then start the service with: `ollama serve`

5. Pulling Your First AI Models

Ollama is the engine, but it needs a brain to run. "Pulling" a model means downloading it to your computer. You only need to do this once per model.

The two essential models

`ollama pull llama3.2` — The thinking model (~2 GB)

`ollama pull nomic-embed-text` — The understanding model (~275 MB)

Optional additional models

Model	Pull Command	Size	Why Try It
gemma2:9b	ollama pull gemma2:9b	~5 GB	Strong alternative thinking model
mistral	ollama pull mistral	~4 GB	Excellent at structured analysis
llama3.1:8b	ollama pull llama3.1:8b	~4.7 GB	Larger, more capable

6. Connecting Ollama to Yesbrainer

Why you need a local server

For security reasons, browsers do not allow pages opened directly from your hard drive (file:// address) to communicate with other programs. A local web server solves this with a single command. Nothing goes to the internet.

Option A — Launcher Script (Recommended)

On macOS/Linux: `chmod +x start-yesbrainer.sh` then `./start-yesbrainer.sh`

On Windows: Double-click `start-yesbrainer.bat`

Option B — Manual

1. Put all files in one folder.
2. Open a terminal in that folder.
3. `python3 -m http.server 8090`
4. Open `http://localhost:8090/yesbrainer-core.html`
5. Check for the Ollama status indicator in the Core header.
6. Open each lens in its own tab from the same localhost URL.

Always open the lenses through localhost:8090, never by double-clicking the HTML files.

Lens URLs

Lens	URL
Core	http://localhost:8090/yesbrainer-core.html
Economic	http://localhost:8090/economic-lens.html
Political	http://localhost:8090/political-lens.html
Social	http://localhost:8090/social-lens.html
Cultural	http://localhost:8090/cultural-lens.html
Technological	http://localhost:8090/technological-lens.html
Environmental	http://localhost:8090/environmental-lens.html
Ethical / Religious	http://localhost:8090/ethical-lens.html

7. Your First Session

1. In the Core, go to the Source Vault tab. Drag .txt or .md files onto the upload area.

2. Set the authority tier — Primary for your main research, Secondary or Background for the rest.

3. Click INDEX ALL SOURCES. The embedding model processes each chunk.

4. Switch to the Steering Dashboard. Set lens weights for your scenario.

5. Open the Synthesis Engine tab. Type your idea or scenario.

6. Click Synthesize. Watch the multi-lens analysis stream in.

7. Try the Arena — click Assassinate This Idea to stress-test your synthesis.

8. Try the Weaver — click Weave Futures to project your idea across time horizons.

9. Open individual lenses in separate tabs for deep dives.

The first query after starting Ollama may take extra seconds as the model loads. Subsequent queries will be faster.

8. Choosing the Right Model

Hardware	Recommended Model	Experience
8 GB RAM (older laptop)	llama3.2 (default)	Solid quality, may be slow on long syntheses
16 GB RAM (modern laptop)	llama3.2 or llama3.1:8b	Smooth experience
Apple Silicon (M1–M4)	llama3.2 or gemma2:9b	Excellent performance
Desktop with GPU (8GB+ VRAM)	llama3.1:8b or mistral	Fast, high-quality output

Every lens has an “Ollama Model” text field. Type the name of the model you want. The change takes effect on the next AI query. You can use different models in different lenses.

9. Troubleshooting

Ollama status shows "Disconnected"

Ollama is not running. On Mac, open the Ollama app. On Windows, check the system tray. On Linux, run `ollama serve`. Then refresh.

Page shows file:// in the address bar

You opened the HTML file directly. Use http://localhost:8090/, not a file:// address.

"Model not found" error

The model is not downloaded. Run: `ollama pull llama3.2` (or the model name shown in the error).

Responses are very slow

Normal for local AI. Close memory-intensive apps. On 8 GB machines, stick to llama3.2.

Source Vault indexing fails

The embedding model is probably missing. Run: `ollama pull nomic-embed-text`

Lenses can’t connect to Core

All tabs must be from the same localhost:8090 address.

Python command not found

Try `python` instead of `python3`. If neither works, install from python.org.

10. Glossary

Term	Definition
AI Model	A program trained on text to understand and generate language. Runs on your computer.
BroadcastChannel	A browser feature that lets tabs on the same website talk to each other.
Embedding	A mathematical representation of text for semantic search.
IndexedDB	A database built into your browser for storing large data.
llama3.2	The default thinking model. Created by Meta and freely available.
Local server	A tiny web server on your computer. Nothing is exposed to the internet.
localStorage	Simple browser storage where Yesbrainer saves state.
nomic-embed-text	The understanding model that creates embeddings.
Ollama	A free, open-source application that runs AI models on your computer.
Pull	The Ollama term for downloading a model. Works offline once pulled.
RAG	Retrieval-Augmented Generation. Finds relevant passages and gives them to the AI.
Stream	When a response appears word by word rather than all at once.
Terminal	A text-based interface. On Mac: Terminal. On Windows: Command Prompt or PowerShell.

YESBRAINER

Your ideas. Your machine. Your rules.