RUN FRONTIER AI ON YOUR OWN MAC

No cloud. No API bills. No one watching your tokens.

A step-by-step guide to setting up EXO on Mac

What is EXO?

EXO is a free, open-source tool that lets you cluster multiple Apple Silicon Macs into one giant AI brain. Because Macs use **unified memory**, pooling them together gives you one massive memory bank — big enough to run trillion-parameter frontier models that normally live inside a data center.

Models You Can Run

Model	Size	What It Does
DeepSeek V3	671B params	Coding + reasoning beast
Llama 3.3	70B params	Fast, versatile assistant
Qwen 2.5	72B params	Multilingual + math
Mistral Large	123B params	Instruction following

What You Need

Mac with Apple Silicon (M1 / M2 / M3 / M4)	macOS 13.0 Ventura or later
Python 3.12+ installed	Homebrew (optional but recommended)
At least 16 GB unified memory per Mac	Good Wi-Fi or ethernet between Macs

How to Install EXO — Step by Step

01

Open Terminal

Press Cmd + Space → type Terminal → hit Enter.

02

Install Python 3.12+ (if not already)

Run: brew install python@3.12 Verify with: python3 --version (should show 3.12 or higher)

03

Install EXO via pip

Run this single command: pip install exo-explore This downloads everything automatically. Takes 1-3 minutes.

04

Start EXO on your first Mac

Run: exo EXO starts up, prints your local IP, and waits for peers.

05

Add more Macs (optional but powerful)

On each additional Mac, install EXO the same way and run: exo They discover each other automatically on the same Wi-Fi. No config needed.

06

Download a model

In a new Terminal window, run: exo run llama-3.2-3b Replace with any supported model (see table above). First run downloads the weights.

07

Start chatting

Once the model loads, EXO opens a local web UI at: http://localhost:52415 Open it in your browser and start using your private AI — no internet, no API key.

Official GitHub Repository

github.com/exo-explore/exo

Star the repo to get updates. All source code, model support lists, and advanced clustering configs live here.

Pro Tips

More RAM = bigger models	A single M2 Max (96 GB) can run 70B models solo. Pool two and you hit 192 GB.
Ethernet > Wi-Fi	Wired connection between Macs dramatically speeds up model splitting.
Private by default	Nothing leaves your machine. Point it at private docs and run wild all night.
No ongoing cost	Pay once for the hardware. Zero per-token fees. Ever.

Follow @thevibefounder for more 100X builds

Comment EXO on the video to get this guide sent to you directly.