Running AI Locally: Complete Setup Guide for Ollama and Open WebUI

A Step-by-Step Tutorial for Beginners

Table of Contents

Introduction

This guide provides complete, beginner-friendly instructions for installing and running large language models (LLMs) locally on your Mac or Windows computer using Ollama. You will also learn how to optionally install Open WebUI, which provides a ChatGPT-like interface for your local models.

Running AI locally gives you complete privacy, full control, and independence from corporate surveillance and subscription fees. No data leaves your computer, and you can use AI models without an internet connection after the initial download.

Why Run AI Locally?

Running AI models locally on your own computer provides several critical advantages:

- **Privacy and Security**: Your conversations, documents, and data never leave your computer. No corporate servers log your queries or analyze your usage patterns. This is especially important for sensitive work, personal information, or confidential documents.
- **Complete Control**: You decide which models to use, when to update them, and how to configure them. There are no usage limits, no throttling, and no sudden policy changes that restrict what you can do.
- **Cost Savings**: Once you download a model, you can use it unlimited times with no subscription fees, no per-query charges, and no hidden costs. The only expense

is the electricity your computer uses.
- **Independence**: You are not dependent on internet connectivity or corporate service availability. Your AI tools work offline, anytime, anywhere. If a company shuts down its AI service, your local models continue working.
- **Learning and Experimentation**: Running models locally demystifies AI technology. You can experiment freely, understand how models work, and develop genuine AI literacy without worrying about costs or usage restrictions.

System Requirements

Minimum Requirements

For macOS:

- macOS 11 (Big Sur) or later
- Apple Silicon (M1/M2/M3) or Intel processor
- 8 GB RAM (16 GB recommended)
- 10 GB free disk space (more for larger models)

For Windows:

- Windows 10 or later (64-bit)
- Modern CPU (Intel Core i5 or AMD Ryzen 5 or better)
- 8 GB RAM (16 GB recommended)
- 10 GB free disk space (more for larger models)

Recommended Hardware

For the best experience, especially with larger models:

- 16 GB RAM or more
- SSD (solid-state drive) for faster model loading
- Dedicated GPU (optional, but improves performance for larger models)

Note: Smaller models like **Gemma** (2B parameters) run well on modest hardware, while larger models like **Llama 3.1 70B** require more powerful computers.

Part 1: Installing Ollama

macOS Installation

Installing Ollama on macOS is straightforward and takes less than 5 minutes.

Step 1: Download Ollama

1. Open your web browser and navigate to **ollama.ai**
2. Click the **"Download for macOS"** button
3. The installer file (`Ollama.dmg`) will download to your **Downloads** folder

Step 2: Install Ollama

1. Open the downloaded `Ollama.dmg` file by double-clicking it
2. A new window will appear showing the Ollama icon
3. **Drag the Ollama icon** into your **Applications** folder
4. Wait for the installation to complete (usually takes 10-20 seconds)
5. Eject the installer disk image by right-clicking the Ollama disk icon on your desktop and selecting **"Eject"**

Step 3: Open Terminal

Ollama runs from the command line, so you need to open Terminal:

1. Open **Finder**
2. Navigate to **Applications** → **Utilities** → **Terminal**
    - *Alternatively*, press **Cmd + Space** to open Spotlight Search, type "Terminal", and press **Enter**
3. A Terminal window will open with a command prompt

Step 4: Verify Installation

To confirm Ollama is installed correctly, type the following command in Terminal and press **Enter**:

```
ollama --version
```

You should see output showing the Ollama version number, such as `ollama version 0.1.23`. If you see this, Ollama is successfully installed!

Windows Installation

Installing Ollama on Windows is also quick and straightforward.

Step 1: Download Ollama

1. Open your web browser and navigate to **ollama.ai**
2. Click the **"Download for Windows"** button
3. The installer file (`OllamaSetup.exe`) will download to your **Downloads** folder

Step 2: Install Ollama

1. Open the downloaded `OllamaSetup.exe` file by double-clicking it
2. If Windows asks "Do you want to allow this app to make changes to your device?", click **"Yes"**
3. The Ollama installer will launch and begin installation automatically
4. Wait for the installation to complete (usually takes 30-60 seconds)
5. Click **"Finish"** when the installation is complete

Step 3: Open Command Prompt or PowerShell

Ollama runs from the command line, so you need to open Command Prompt or PowerShell:

Option 1: Command Prompt

1. Press **Windows Key + R** to open the Run dialog
2. Type `cmd` and press **Enter**

Option 2: PowerShell

1. Press **Windows Key + X**
2. Select **"Windows PowerShell"** or **"Terminal"** from the menu

Step 4: Verify Installation

To confirm Ollama is installed correctly, type the following command and press **Enter**:

```
ollama --version
```

You should see output showing the Ollama version number, such as `ollama version 0.1.23`. If you see this, Ollama is successfully installed!

Part 2: Using Ollama

Now that Ollama is installed, you can download and run AI models on your computer.

Downloading Your First Model

Ollama makes it incredibly easy to download models. We recommend starting with **Llama 3.1 8B**, which offers excellent performance and runs well on most modern computers.

Download Llama 3.1 8B

In your Terminal (macOS) or Command Prompt/PowerShell (Windows), type:

```
ollama pull llama3.1:8b
```

Press **Enter**. Ollama will begin downloading the model. You will see a progress bar showing the download status:

```
pulling manifest
pulling 8934d96d3f08... 100% ▕████████████████▏ 4.7 GB
pulling 8c17c2ebb0ea... 100% ▕████████████████▏ 7.0 KB
pulling 7c23fb36d801... 100% ▕████████████████▏ 4.8 KB
pulling 2e0493f67d0c... 100% ▕████████████████▏   59 B
pulling fa304d675061... 100% ▕████████████████▏   91 B
pulling 42ba7f8a01dd... 100% ▕████████████████▏  557 B
verifying sha256 digest
writing manifest
success
```

Note: The download size is approximately **4.7 GB**, so it may take several minutes depending on your internet connection speed. Once downloaded, the model is stored locally and you will never need to download it again.

Running Models

Once a model is downloaded, you can start using it immediately.

Start an Interactive Chat Session

To start chatting with the Llama 3.1 model, type:

```
ollama run llama3.1:8b
```

Press **Enter**. You will see a prompt that looks like this:

```
>>>
```

You can now type questions or prompts, and the model will respond. For example:

```
>>> What is the capital of France?
```

```
The capital of France is Paris.

>>> Explain photosynthesis in simple terms.

Photosynthesis is the process by which plants use sunlight, water, and carbon dioxide
to create oxygen and energy in the form of sugar. It's how plants make their own food!
```

Exit the Chat Session

To exit the interactive chat session, type:

```
/bye
```

Or press **Ctrl + D** (macOS/Linux) or **Ctrl + Z** then **Enter** (Windows).

Essential Ollama Commands

Here are the most important Ollama commands you will use:

List Downloaded Models

To see all models you have downloaded:

```
ollama list
```

Example output:

```
NAME                ID              SIZE      MODIFIED
llama3.1:8b         8934d96d3f08    4.7 GB    2 hours ago
mistral:latest      f974a74358d6    4.1 GB    1 day ago
```

Pull (Download) a Model

To download a new model:

```
ollama pull <model-name>
```

Examples:

```
ollama pull mistral
ollama pull gemma:2b
ollama pull llama3.1:70b
```

Run a Model

To start an interactive chat session with a model:

```
ollama run <model-name>
```

Examples:

```
ollama run llama3.1:8b
ollama run mistral
ollama run gemma:2b
```

Delete a Model

To remove a model you no longer need (to free up disk space):

```
ollama rm <model-name>
```

Example:

```
ollama rm gemma:2b
```

Get Help

To see all available Ollama commands:

```
ollama --help
```

Part 3: Installing Open WebUI (Optional)

What is Open WebUI?

Open WebUI is a free, open-source web interface that provides a ChatGPT-like experience for your local Ollama models. Instead of using the command line, you can interact with your models through a familiar, user-friendly web interface with features like:

- Conversation history
- Easy model switching
- Markdown rendering and code highlighting
- Document upload and analysis
- Multiple chat sessions
- Completely private and offline

Note: Open WebUI is **optional**. Ollama works perfectly fine from the command line, and you do not need Open WebUI to use Ollama. However, if you prefer a graphical interface, Open WebUI is an excellent choice.

Installing Docker

Open WebUI runs inside a **Docker container**, so you need to install Docker first.

macOS: Install Docker Desktop

1. Navigate to **docker.com/products/docker-desktop**
2. Click **"Download for Mac"**
3. Choose the appropriate version:
    - **Apple Silicon** (M1/M2/M3): Download "Mac with Apple chip"
    - **Intel**: Download "Mac with Intel chip"
4. Open the downloaded `Docker.dmg` file
5. Drag the **Docker** icon to your **Applications** folder
6. Open Docker from your Applications folder
7. Docker will ask for permissions—click **"OK"** to grant them
8. Wait for Docker to start (you will see a Docker icon in your menu bar)
9. Docker Desktop is now running

Windows: Install Docker Desktop

1. Navigate to docker.com/products/docker-desktop
2. Click **"Download for Windows"**
3. Open the downloaded `Docker Desktop Installer.exe` file
4. If prompted, click **"Yes"** to allow the installer to make changes
5. Follow the installation wizard:
    - Check **"Use WSL 2 instead of Hyper-V"** (recommended)
    - Click **"OK"** to proceed
6. Wait for the installation to complete
7. Click **"Close and restart"** to restart your computer
8. After restarting, Docker Desktop will launch automatically
9. Accept the Docker Subscription Service Agreement
10. Docker Desktop is now running

Verify Docker Installation

To confirm Docker is installed and running, open Terminal (macOS) or Command Prompt/PowerShell (Windows) and type:

```
docker --version
```

You should see output like:

```
Docker version 24.0.6, build ed223bc
```

If you see this, Docker is successfully installed!

Installing Open WebUI

Once Docker is installed and running, installing Open WebUI is a single command.

Step 1: Run the Open WebUI Container

In your Terminal (macOS) or Command Prompt/PowerShell (Windows), type the following command and press **Enter**:

```
docker run -d -p 3000:8080 --add-host=host.docker.internal:host-gateway -v open-webui:/app/backend/data --name open-webui --restart always ghcr.io/open-webui/open-webui:main
```

What this command does:

- Downloads the Open WebUI Docker image (first time only)
- Runs Open WebUI in the background
- Makes it accessible at `http://localhost:3000`
- Connects to your local Ollama installation
- Saves your data persistently
- Automatically restarts Open WebUI when you restart your computer

The first time you run this command, Docker will download the Open WebUI image (approximately 1-2 GB). This may take a few minutes. You will see output like:

```
Unable to find image 'ghcr.io/open-webui/open-webui:main' locally
main: Pulling from open-webui/open-webui
...
```

```
Status: Downloaded newer image for ghcr.io/open-webui/open-webui:main
a7f3c5d8e9b2c1f4e6d8a9b3c5d7e9f1a2b4c6d8e0f2a4b6c8d0e2f4a6b8c0d2
```

The long alphanumeric string at the end is the container ID, confirming that Open WebUI is now running.

Step 2: Access Open WebUI

1. Open your web browser
2. Navigate to **http://localhost:3000**
3. You will see the Open WebUI interface

Step 3: Create an Account

The first time you access Open WebUI, you need to create a local account:

1. Click **"Sign up"**
2. Enter your desired username and password
    - **Note**: This account is stored **only on your computer**. It is not sent to any server.
3. Click **"Create Account"**
4. You will be logged in automatically

Step 4: Start Chatting

1. You will see a chat interface similar to ChatGPT
2. In the top-left corner, you can select which Ollama model to use from the dropdown menu
3. Type your message in the text box at the bottom and press **Enter**
4. The model will respond in the chat window

Congratulations! You now have a fully functional, private, local AI assistant running on your computer.

Managing Open WebUI

Stop Open WebUI

If you want to stop Open WebUI (to free up system resources), run:

```
docker stop open-webui
```

Start Open WebUI Again

To start Open WebUI after stopping it:

```
docker start open-webui
```

Then navigate to **http://localhost:3000** in your browser.

Check if Open WebUI is Running

To see if the Open WebUI container is running:

```
docker ps
```

You should see a line with `open-webui` in the output if it's running.

Update Open WebUI

To update Open WebUI to the latest version:

```
docker stop open-webui
docker rm open-webui
docker pull ghcr.io/open-webui/open-webui:main
docker run -d -p 3000:8080 --add-host=host.docker.internal:host-gateway -v open-webui:/app/backend/data --name open-webui --restart always ghcr.io/open-webui/open-webui:main
```

Recommended Models

Here are some of the most popular and useful models you can run with Ollama, organized by size and use case.

Best All-Around Models

Model	Size	Description	Command
Llama 3.1 8B	4.7 GB	High-quality, versatile model suitable for most tasks. Excellent balance of performance and resource requirements.	`ollama pull llama3.1:8b`
Mistral	4.1 GB	Fast, efficient model with excellent instruction-following capabilities. Great for everyday tasks.	`ollama pull mistral`

Lightweight Models (For Older/Less Powerful Computers)

Model	Size	Description	Command
Gemma 2B	1.4 GB	Lightweight model optimized for consumer hardware. Runs smoothly on less powerful computers.	`ollama pull gemma:2b`
Phi-3 Mini	2.3 GB	Small but capable model from Microsoft. Good for basic tasks on limited hardware.	`ollama pull phi3`

Advanced Models (Require Powerful Hardware)

Model	Size	Description	Command
Llama 3.1 70B	40 GB	Extremely high-quality model with exceptional reasoning and writing capabilities. Requires 64+ GB RAM.	`ollama pull llama3.1:70b`
Mixtral 8x7B	26 GB	Powerful mixture-of-experts model. Excellent performance, requires 32+ GB RAM.	`ollama pull mixtral`

Specialized Models

Model	Size	Description	Command
CodeLlama	3.8	Optimized for code generation and	`ollama pull`

	GB	programming tasks.	`codellama`
Llama 3.1 Vision	4.7 GB	Multimodal model that can analyze images and text.	`ollama pull llama3.2-vision`

Tip: Start with **Llama 3.1 8B** or **Mistral**. These models provide excellent performance and run well on most modern computers.

Troubleshooting

Ollama Issues

"Command not found: ollama"

Problem: Your system cannot find the Ollama command.

Solution:

- **macOS**: Make sure you dragged Ollama to your Applications folder during installation. Try restarting Terminal.
- **Windows**: Make sure the installation completed successfully. Try restarting Command Prompt/PowerShell. If the problem persists, restart your computer.

"Error: connection refused"

Problem: Ollama is not running in the background.

Solution:

- **macOS**: Open Ollama from your Applications folder. You should see an Ollama icon in your menu bar.
- **Windows**: Ollama should start automatically. If not, search for "Ollama" in the Start menu and launch it.

Model Download Fails or Stops

Problem: The model download is interrupted or fails.

Solution:

- Check your internet connection
- Make sure you have enough disk space (at least 10 GB free)
- Try running the `ollama pull` command again—Ollama will resume the download from where it stopped

Model Runs Very Slowly

Problem: The model takes a long time to respond.

Solution:

- Try a smaller model like **Gemma 2B** or **Phi-3 Mini**
- Close other applications to free up RAM
- Consider upgrading your computer's RAM if you frequently use AI models

---

Docker and Open WebUI Issues

"Cannot connect to the Docker daemon"

Problem: Docker is not running.

Solution:

- **macOS**: Open Docker Desktop from your Applications folder. Wait for the Docker icon to appear in your menu bar.
- **Windows**: Open Docker Desktop from the Start menu. Wait for Docker to finish starting.

"Port 3000 is already in use"

Problem: Another application is using port 3000.

Solution:

- Change the port in the Docker command. Replace `-p 3000:8080` with `-p 3001:8080` (or any other available port)
- Access Open WebUI at `http://localhost:3001` instead

Open WebUI Cannot Connect to Ollama

Problem: Open WebUI shows "Cannot connect to Ollama" error.

Solution:

- Make sure Ollama is running (check for the Ollama icon in your system tray/menu bar)
- Make sure you have at least one model downloaded (`ollama list`)
- Restart the Open WebUI container:

```
docker restart open-webui
```

Open WebUI Shows Blank Page

Problem: Open WebUI loads but shows a blank page.

Solution:

- Clear your browser cache and reload the page
- Try accessing Open WebUI in a different browser
- Restart the Open WebUI container:

```
docker restart open-webui
```

Additional Resources

Official Documentation

- **Ollama Official Website**: ollama.ai
- **Ollama GitHub Repository**: github.com/ollama/ollama
- **Ollama Model Library**: ollama.ai/library
- **Open WebUI GitHub Repository**: github.com/open-webui/open-webui
- **Docker Documentation**: docs.docker.com

Community and Support

- **Ollama Discord Community**: Join the Ollama Discord server for help and discussions

- **Open WebUI Discord Community**: Join the Open WebUI Discord server for support
- **Reddit**: r/LocalLLaMA for discussions about running models locally

Companion Website

All slides, handouts, and resources from the presentation "AI: The Great Illuminator or The Great Imitator?" are available at: pyaim.github.io/AI-Illuminator-Imitator

---

Conclusion

You now have everything you need to run AI models locally on your computer with complete privacy and control. Whether you use Ollama from the command line or prefer the graphical interface of Open WebUI, you have powerful AI tools at your fingertips—no subscriptions, no data collection, no limits.

Remember: AI is a tool, not a replacement for human thinking. Use it to streamline mundane tasks, speed up tedious work, and explore new ideas—but always verify its output, maintain your critical thinking skills, and never delegate your judgment to a machine. Use AI wisely. Question its output. Maintain your critical thinking. Teach others to do the same.

Document Version: 1.0
Last Updated: November 11, 2025
Author: Moez Ben-Azzouz