How to Run Ollama on Local Machine

(Windows + macOS M1)

For Windows 10/11

1. Download and Install

- Go to: https://ollama.com/download
- Download the `.exe` installer
- Run the installer and follow the prompts

2. Open Command Prompt

Search for “Command Prompt” in the Start Menu and open it.

3. Verify the Installation

Run:

```
ollama --version
```

You should see the version number if it's installed correctly.

4. Pull a Model

Example to pull LLaMA 3:

```
ollama pull llama3
```

5. Run the Model

Start interacting:

```
ollama run llama3
```

This opens a local terminal-based chat.

For macOS (Apple Silicon)

1. Install Homebrew (if not already installed)

In the Terminal, run:

```
/bin/bash -c "$(curl -fsSL
https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"
```

2. Install Ollama

Run:

```
brew install ollama
```

3. Verify the Installation

```
ollama --version
```

4. Pull a Model

Example:

```
ollama pull llama3
```

5. Run the Model

Start the chat session:

```
ollama run llama3
```

Apple Silicon (M1/M2/M3/M4) will automatically use GPU acceleration via Metal.

6. Use the API (Optional)

Same as on Windows. REST API is at:

```
http://localhost:11434
```

Common Commands (Works on Both)

Action	Command
List downloaded models	`ollama list`
Remove a model	`ollama rm llama3`
Update Ollama	`ollama update`
Start Ollama server	`ollama serve`
Create custom model	`ollama create mymodel -f Modelfile`