Install Ollama

on Windows & macOS | Javathoughts.com

What is Ollama?

Ollama is an open-source tool that lets you run large language models (LLMs) locally on your machine, without any cloud dependencies or API keys. It supports models like Llama 2, Mistral, CodeLlama, Phi-2, and many more.

System Requirements

- macOS 13 (Ventura) or later / Windows 10 or later
- 8 GB RAM minimum (16 GB recommended)
- GPU recommended for faster inference (Apple Silicon / NVIDIA)
- ~4-8 GB free disk space per model

macOS Installation

```
# Download and install Ollama
curl -fsSL https://ollama.com/install.sh | sh

# Verify installation
ollama --version

# Pull and run a model
ollama pull llama2
ollama run llama2
```

Windows Installation

- Download the installer from https://ollama.com/download
- Run the .exe installer and follow the on-screen instructions
- Open PowerShell or Command Prompt
- Verify with: ollama --version
- Pull a model: ollama pull llama2

Using the REST API

Ollama exposes a local HTTP API on port 11434:

```
cur
  "model": "llama2",
  "prompt": "Explain microservices in 3 sentences"
}'
```

Troubleshooting

- Port conflict: Change port with OLLAMA_HOST=0.0.0.0:8080
- Slow inference: Ensure GPU drivers are up to date
- Model not found: Run ollama list to see available models