



# Ollama

Mukesh Mittal



# Ollama

- ❖ **Ollama** is an open-source framework
- ❖ designed to facilitate the deployment of large language models on local environments.
- ❖ It aims to simplify the complexities involved in running and managing these models,
- ❖ providing a seamless experience for users across different operating systems.




## Get up and running with large language models.

Run [Llama 3.1](#), [Phi 3](#), [Mistral](#), [Gemma 2](#), and other models. Customize and create your own.

Download ↓

Available for macOS, Linux, and  
Windows (preview)



# Download Ollama



macOS



Linux



Windows

**Download for Windows (Preview)**

Requires Windows 10 or later



Most popular

## llama3

Meta Llama 3: The most capable openly available LLM to date

8B 70B

↓ 5.7M Pulls ↗ 68 Tags ⌚ Updated 3 months ago

## gemma

Gemma is a family of lightweight, state-of-the-art open models built by Google DeepMind. Updated to version 1.1

2B 7B

↓ 4M Pulls ↗ 102 Tags ⌚ Updated 4 months ago

## qwen

Qwen 1.5 is a series of large language models by Alibaba Cloud spanning from 0.5B to 110B parameters

0.5B 1.8B 4B 32B 72B 110B

↓ 3.9M Pulls ↗ 379 Tags ⌚ Updated 2 months ago

## mistral

The 7B model released by Mistral AI, updated to version 0.3.

Tools 7B

↓ 3.3M Pulls ↗ 84 Tags ⌚ Updated 4 weeks ago

## phi3

Phi-3 is a family of lightweight 3B (Mini) and 14B (Medium) state-of-the-art open models by Microsoft.

3B 14B

↓ 2.3M Pulls ↗ 72 Tags ⌚ Updated 2 weeks ago

<https://github.com/ollama/ollama>

github.com/ollama/ollama/tree/main/examples

ollama / examples /

go-generate-streaming

go-generate

go-http-generate

go-multimodal

go-pull-progress

jupyter-notebook

kubernetes

langchain-python-rag-document

langchain-python-rag-privategpt

langchain-python-rag-websummary

langchain-python-simple

langchain-typescript-simple

streaming



## Tool support

July 25, 2024

Ollama now supports tool calling with popular models such as Llama 3.1. This enables a model to answer a given prompt using tool(s) it knows about, making it possible for models to perform more complex tasks or interact with the outside world.

---

## Google Gemma 2

June 27, 2024

Gemma 2 is now available on Ollama in 3 sizes - 2B, 9B and 27B.

---

## An entirely open-source AI code assistant inside your editor

May 31, 2024

Continue enables you to easily create your own coding assistant directly inside Visual Studio Code and JetBrains with open-source LLMs.

## REST API

Ollama has a REST API for running and managing models.


### Generate a response

```
curl http://localhost:11434/api/generate -d '{
  "model": "llama3.1",
  "prompt": "Why is the sky blue?"
}'
```

### Chat with a model

```
curl http://localhost:11434/api/chat -d '{
  "model": "llama3.1",
  "messages": [
    { "role": "user", "content": "why is the sky blue?" }
  ]
}'
```





### Usage:

```
ollama [flags]
ollama [command]
```

### Available Commands:

serve	Start ollama
create	Create a model from a Modelfile
show	Show information for a model
run	Run a model
pull	Pull a model from a registry
push	Push a model to a registry
list	List models
ps	List running models
cp	Copy a model
rm	Remove a model
help	Help about any command

### Flags:

-h, --help	help for ollama
-v, --version	Show version information

Use "ollama [command] --help" for more information



# Open WebUI (Formerly Ollama WebUI) 🖐️

Stars 36k Forks 4.1k Watchers 189

repo size 111 MB languages 11 svelte 55.3% last commit yesterday

hits 7694 / 3907023

Discord Open WebUI Sponsor

Open WebUI is an [extensible](#), feature-rich, and user-friendly self-hosted WebUI designed to operate entirely offline. It supports various LLM runners, including Ollama and OpenAI-compatible APIs. For more information, be sure to check out our [Open WebUI Documentation](#).

<https://github.com/open-webui/open-webui>

## Installation with Default Configuration

- If Ollama is on your computer, use this command:

```
docker run -d -p 3000:8080 --add-host=host.docker.internal:host-gateway -v open-webui:/app/backend/
```

- If Ollama is on a Different Server, use this command:

To connect to Ollama on another server, change the `OLLAMA_BASE_URL` to the server's URL:

```
docker run -d -p 3000:8080 -e OLLAMA_BASE_URL=https://example.com -v open-webui:/app/backend/data -
```

- To run Open WebUI with Nvidia GPU support, use this command:

```
docker run -d -p 3000:8080 --gpus all --add-host=host.docker.internal:host-gateway -v open-webui:/a
```

```
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
```

👤 New Chat



llama3:latest ▾ +

🔗 Workspace

🔍 Search

Today

New Chat



what is open webui?



llama3:latest

🔍 Searching the web for 'what is open webui?'



👤 Timothy J. Baek


+ Send a Message



LLMs can make mistakes. Verify important information.



## Made by Open WebUI Community

1  @hub 2  @blnx 3  @darkstorm2150 4  @hotnikq 5  @coco21 6  @nicluckie 7  @mesharu

Models [+ Create](#)

## New Functions

[See All](#)

#1 **PIPE** **Mixture of Agents Pipe**  
moa\_pipe Mixture of Agents pipe

[VIEW](#)

@arossito79

#2 **FILTER** **v0.1 Auto Context**  
auto\_context Gives auto context informations like date, time plus some user infos...

[VIEW](#)

@webmarka

#3 **PIPE** **v0.1.1 Perplexity**  
perplexity Perplexity Manifold Pipe

[VIEW](#)

@sastineab

#4 **PIPE** **v0.1.0 Azure OpenAI**  
azure Pipe to connect with Azure OpenAI models

[VIEW](#)

@snompy

## Featured Models

[See All](#)

#1  
**Carl Jung Gpt-4o**  
Carl Jung

[VIEW](#)

@adudeandhis...

#2  
**Finn**  
This a model like no other, This model it supposed to act like my...

[VIEW](#)

@fishesarethings

#3  
**St. Augustine Of...**  
St. Augustine of Hippo the Church Father

[VIEW](#)

@adudeandhis...

#4  
**Navi**  
WEB Version: Saints Security Group LLC's main cybersec model...

[VIEW](#)

@alexkollar

#5  
**Alienar**  
A text adv where the narrates a

## New Tools

[See All](#)

#1 **TOOL** **v0.1.1 Outil de génération d'éléments de récit**  
outil\_generation\_delements\_de\_recit Générer un large éventail d'éléments...

[VIEW](#)

@monseur...

#2 **TOOL** **v1.1.0 Paperless**  
paperless Tool to interact with paperless-ngx documents

[VIEW](#)

@joine

#3 **TO...** **v0..... Youtube Transcript Provider (Langchain...**  
youtube\_transcript\_provider\_lo A youtube transcript provider without RAG. Uses...

[VIEW](#)

@thearyadev

#4 **TOOL** **v0.0.4 Enhanced Web Scrape**  
web\_scrape An improved web scraping tool that extracts text content using Jina...

[VIEW](#)

@wholybird

## New Prompts

[See All](#)

#1 **analyze It And Provide Specific Recomm**  
@um... /analyze-it-and-provide-specific-...

#2 **AI-RedTeam-Assistant**  
@sramelyk /ai-redteam-assistant

#3 **testt**  
@worshiperworshiper /testt

#4 **test**  
@worshiperworshiper /test

#5 **Linux Command Expert**  
@kennikvik /linux-command

#6 **optStratEnterprise**  
@yenn789 /optstratentreprise

#7 **etymology**  
@caico /etymology

#8 **TutorGPT (German)**  
@tt12sp /tutorgpt



# Questions?



# Thanks