# 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 <u>Llama 3.1</u>, <u>Phi 3</u>, <u>Mistral</u>, <u>Gemma 2</u>, and other models. Customize and create your own.

Download 1

Available for macOS, Linux, and Windows (preview)

## Download Ollama







Download for Windows (Preview)

Requires Windows 10 or later



Filter by name...

Most popular

#### llama3

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

8B 70B

₫ 5.7M Pulls S 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 0 102 Tags 0 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 Al, updated to version 0.3.

Tools 7B

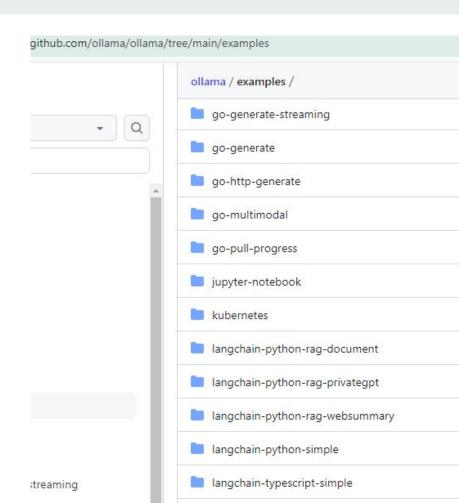
± 3.3M Pulls ○ 84 Tags ○ Updated 4 weeks ago

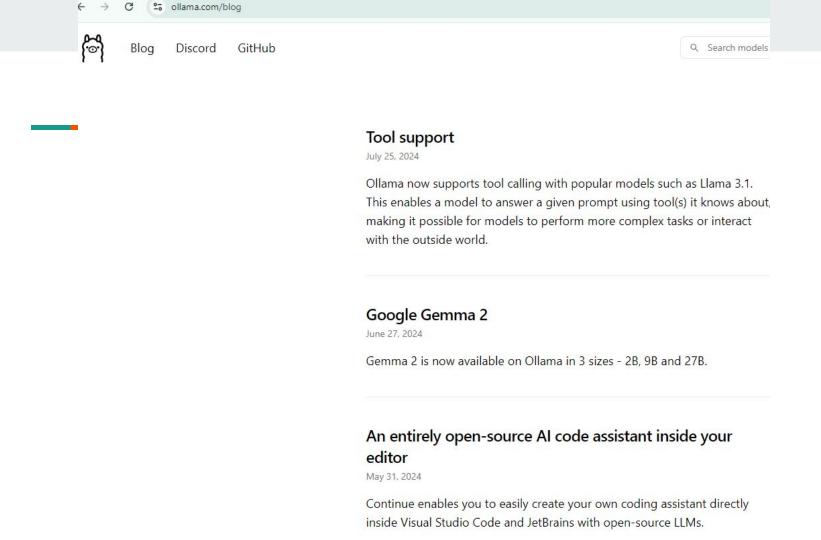
#### phi3

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

3B 14B

# https://github.com/ollama/oll





### **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

```
Usage:
 ollama [flags]
 ollama [command]
Available Commands:
            Start ollama
 serve
 create Create a model from a Modelfile
 show
            Show information for a model
            Run a model
 run
 pull
            Pull a model from a registry
            Push a model to a registry
 push
            List models
 list
```

List running models

rm Remove a model help Help about any command

ps

ср

Flags:

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

Copy a model

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

## Open WebUI (Formerly Ollama WebUI) 🤏





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 OpenAl-compatible APIs. For more information, be sure to check out our Open WebUI Documentation.

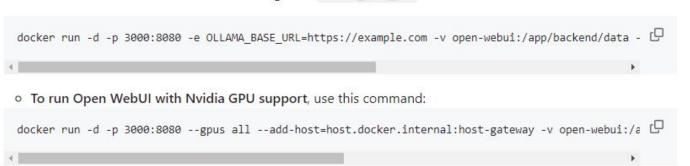
### Installation with Default Configuration

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

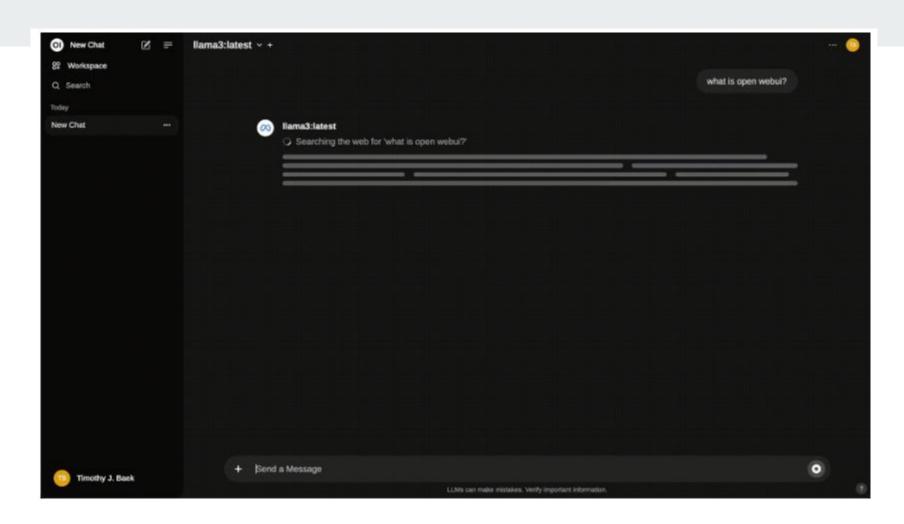
```
docker run -d -p 3000:8080 --add-host=host.docker.internal:host-gateway -v open-webui:/app/backend/
```

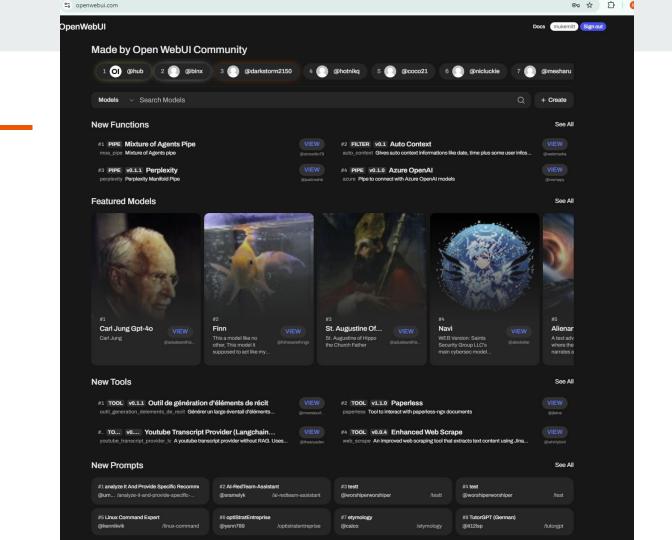
. If Ollama is on a Different Server, use this command:

To connect to Ollama on another server, change the OLLAMA\_BASE\_URL to the server's URL:



docker run -d -p 3000:8080 --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





# Questions?

# Thanks