**How to run LLaMA locally with OLLAMA configuration**

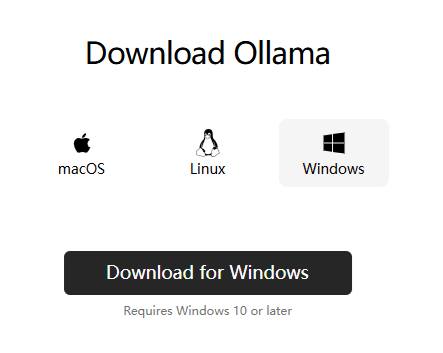
**Run LLaMA locally using Ollama**

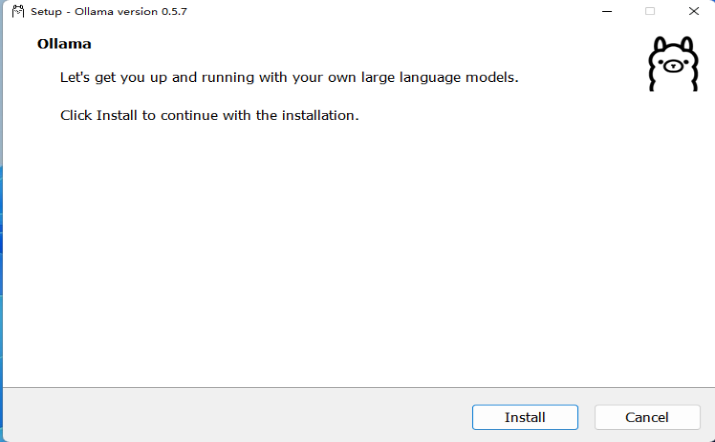
OLLAMA is a simple local LLM (Large Language Model) inference tool. It can efficiently run models such as LLaMA on the CPU or GPU, and supports Mac, Linux, and Windows.

**1. Install Ollama**

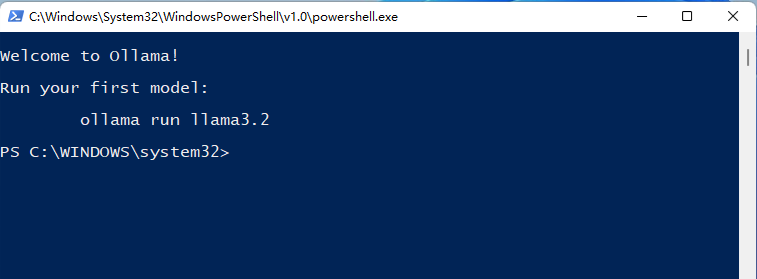
**（1）Download and install Ollama:**

Ollama official version: **https://ollama.com/**, select the windows version to download and install Ollama





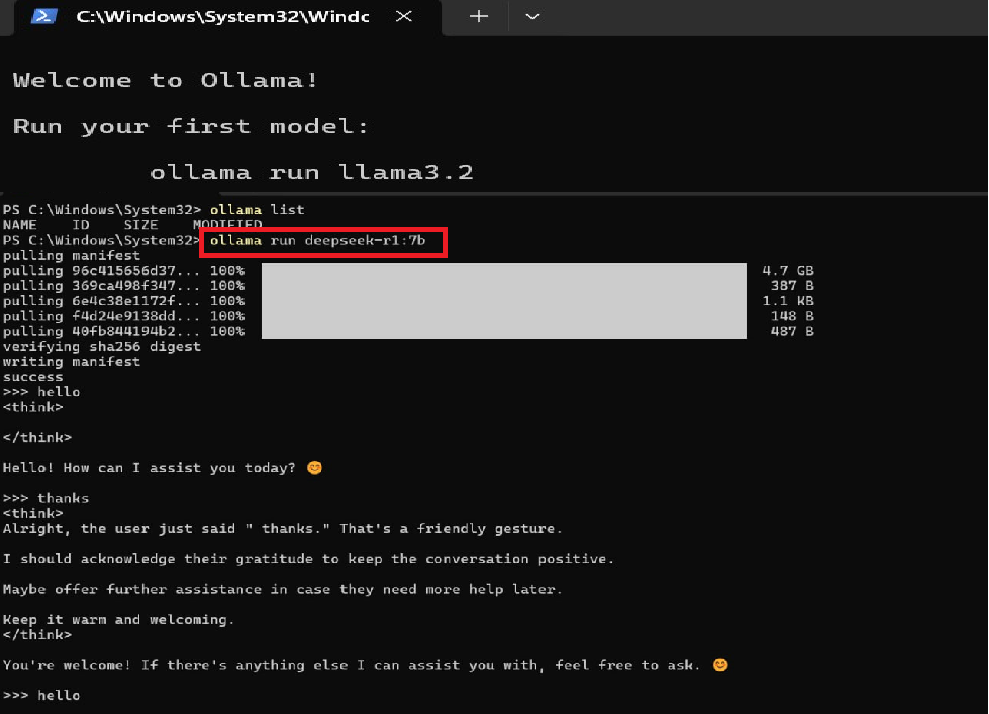
If Ollama is installed successfully, it will show as below:



**（2）Install Deepseek locally (ollama run deepseek-r1:7b):**

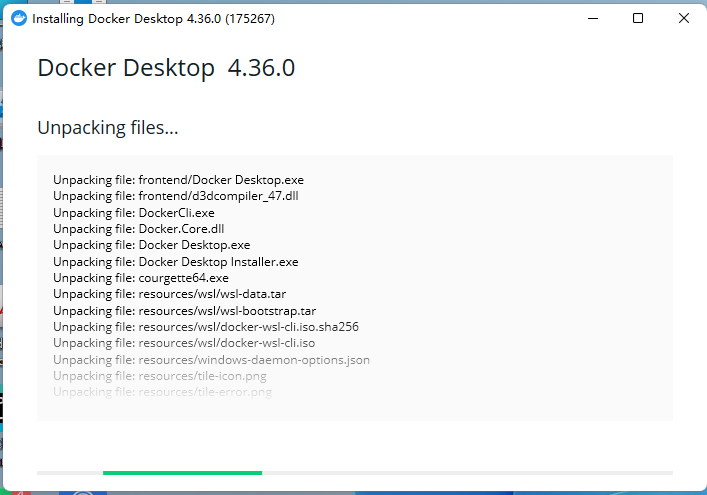
bash

ollama run deepseek-r1:7b

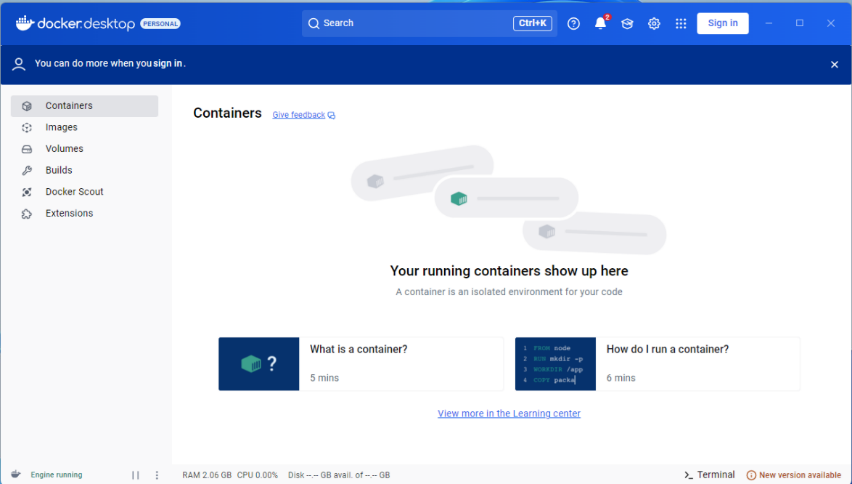


**2. Running deepseek-r1:7b in Docker**

**（1）Install Docker**

Windows users: Download and install Docker Desktop Installer.exe from [**https://www.docker.com/products/docker-desktop/**](https://www.docker.com/products/docker-desktop/)

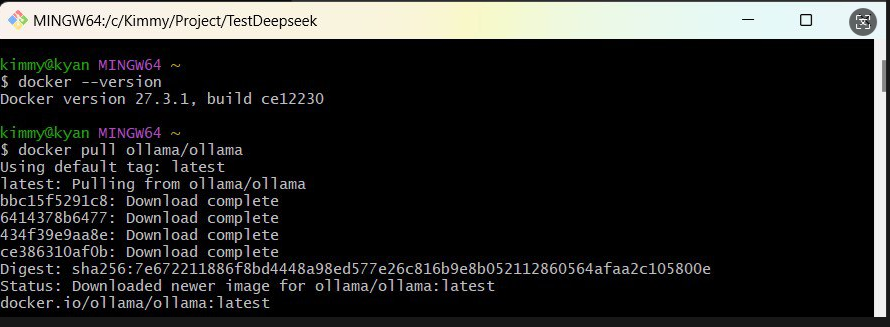
After the Docker is installed successfully, it will show as below:



**（2）Pull the Ollama official Docker image**

bash

docker pull ollama/ollama

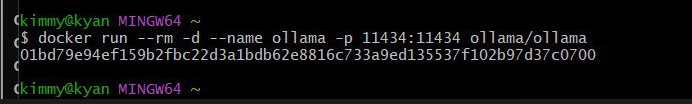


**（3）Run the Ollama container**

**Start Ollama and map the API port:**

bash

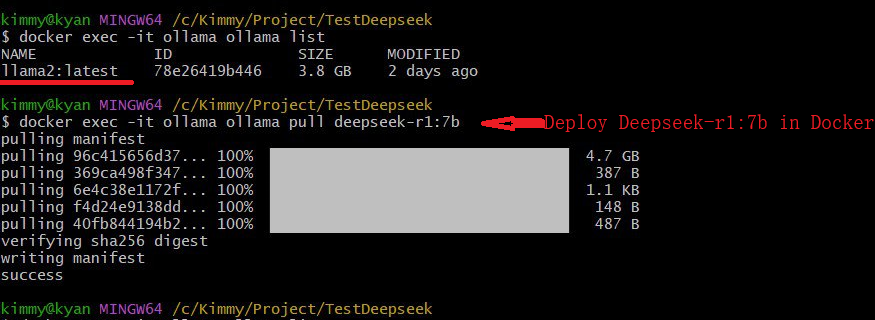
docker run --rm -d --name ollama -p 11434:11434 ollama/ollama



（4）**Enter the Docker container to download DeepSeek-R1:7B**

Bash

docker exec -it ollama ollama pull deepseek-r1:7b



（5）**Test whether the model is installed successfully**

bash

docker exec -it ollama ollama list



**3. Python code calling deepseek-r1:7b**

*import requests*

*# Ollama API endpoint*

*OLLAMA\_API\_URL = "http://localhost:11434/api/generate"*

*# Request data*

*data = {*

*"model": "deepseek-r1:7b",*

*"prompt": "请介绍一下深度学习。",*

*"stream": False # If you want to stream the output, you can change it to True*

*}*

*# Send POST request*

*response = requests.post(OLLAMA\_API\_URL, json=data)*

*# Parse return results*

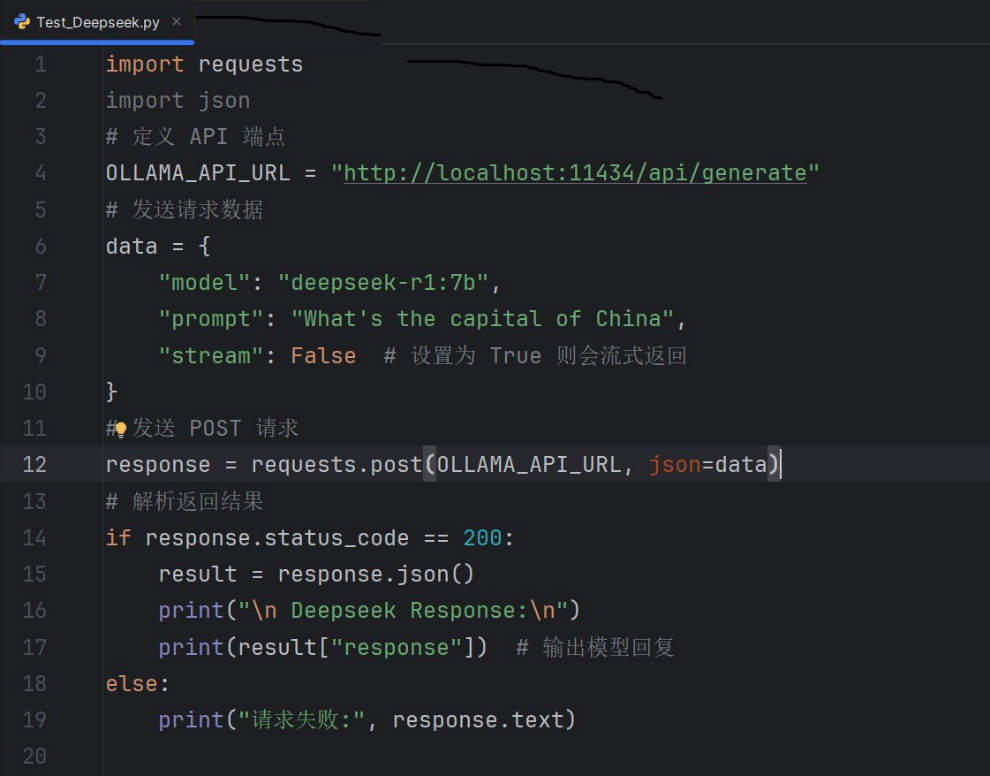
*if response.status\_code == 200:*

*result = response.json()*

*print(result["response"]) # Output model reply*

*else:*

*print("Request failed:", response.text)*



**📌 Description**

**model:** "deepseek-r1:7b" specifies to use DeepSeek-R1:7B

**prompt :** incoming user question

**stream:** False means non-streaming return (change to True if you want streaming return)

**🎯 Summary**

1. Make sure Ollama is running → ollama serve
2. Make sure DeepSeek-R1:7B is installed → ollama pull deepseek-r1:7b
3. Use requests or ollama library to call API
4. Supports streaming output, suitable for long text

**This way, you can call DeepSeek-R1:7B locally in Python and no API Key is required! 🚀**