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

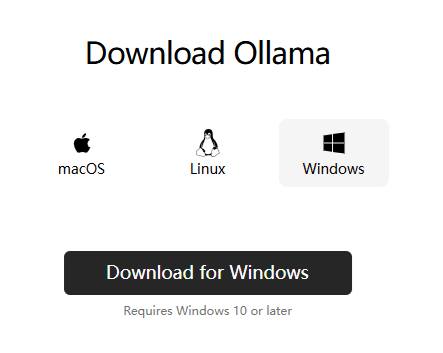
**Run LLaMA locally using Ollama**

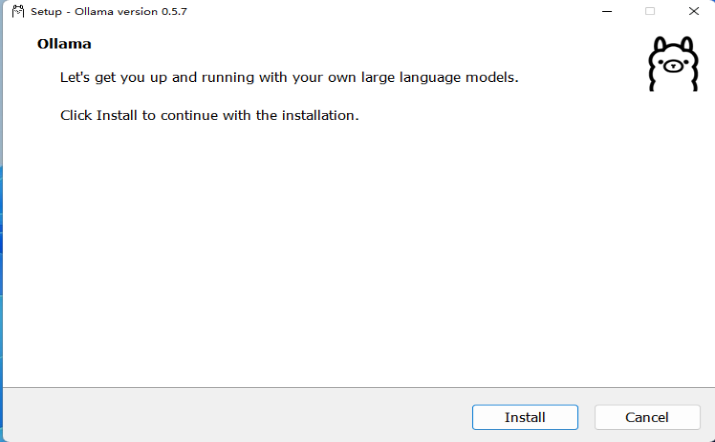
OLLAMA is a simple local LLM (Large Language Model) inference tool. It can efficiently run models such as LLaMA on the CPU or GPU, and supports Mac, Linux, and Windows.

**1. Install Ollama**

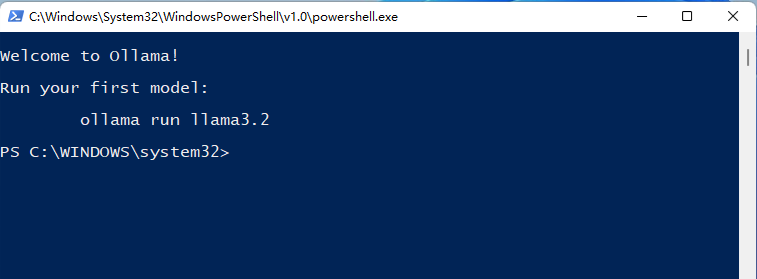
**（1）Download and install Ollama:**

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）Download LLaMA 2:**

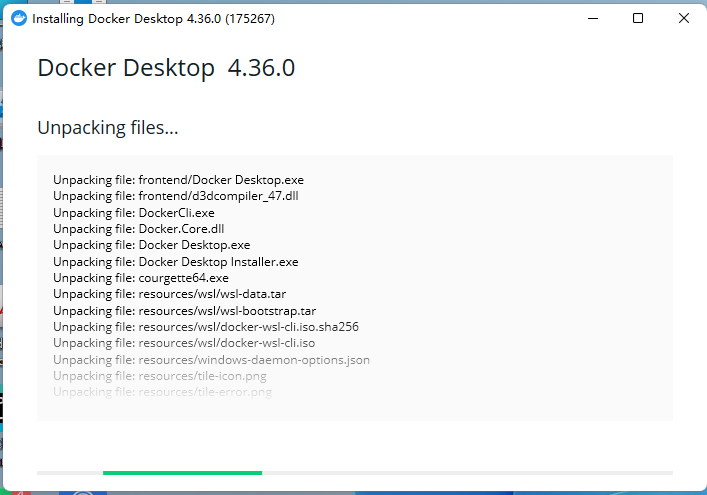
bash

ollama pull llama2

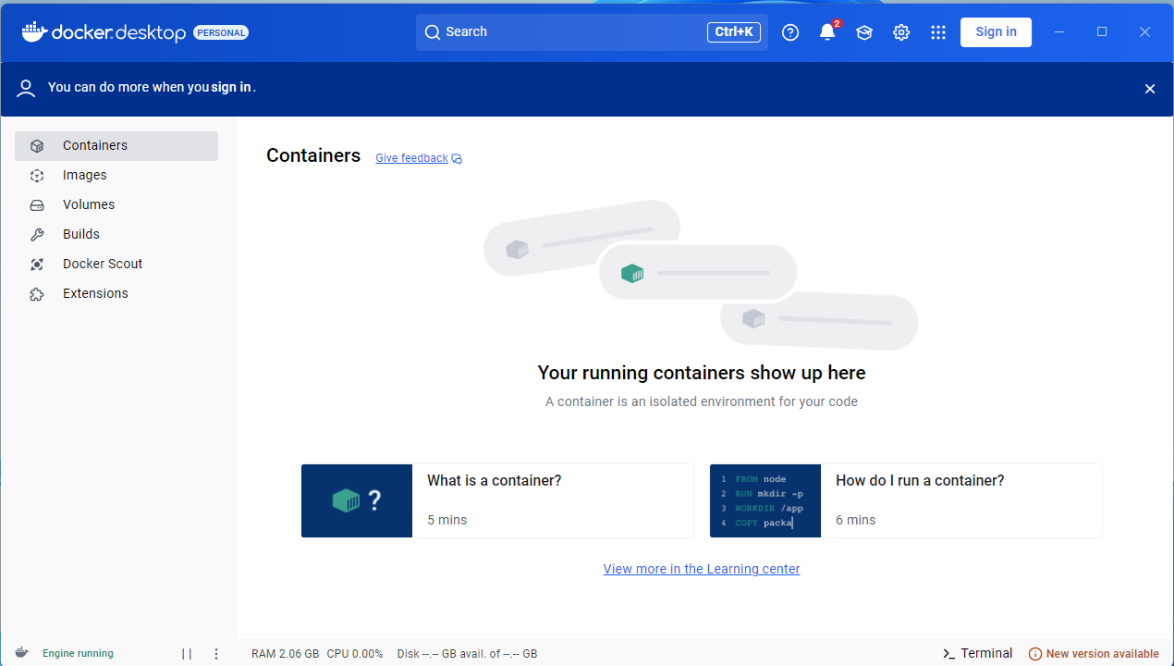


**2. Running LLaMA 2 in Docker**

**（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/)

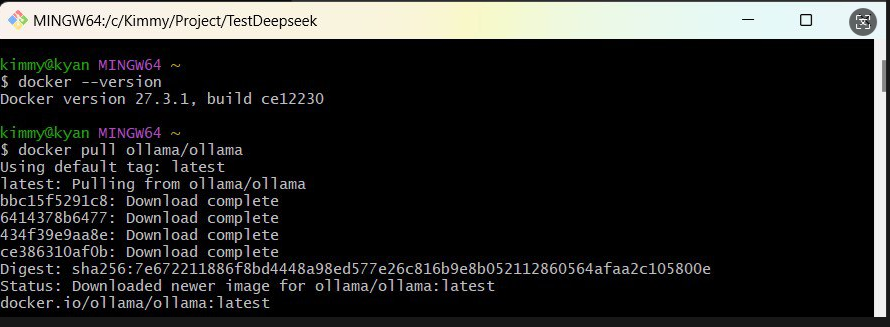
After the Docker is installed successfully, it will show as below:



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

bash

docker pull ollama/ollama

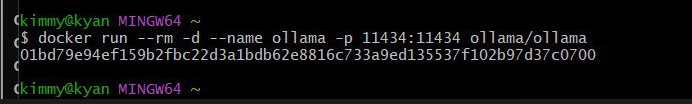


**（3）Run the Ollama container**

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

bash

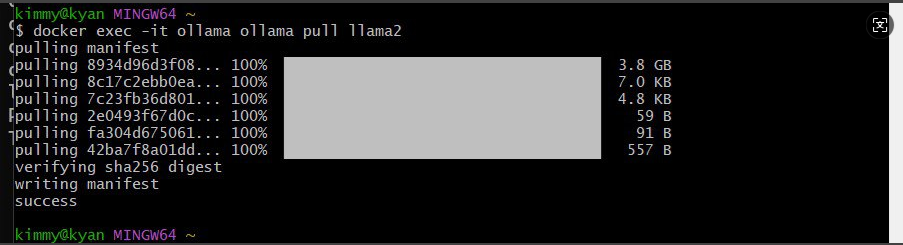
docker run --rm -d --name ollama -p 11434:11434 ollama/ollama



（4）Download LLaMA 2 model

bash

docker exec -it ollama ollama pull llama2



**3. Python code calling LLaMA 2**

Use requests to call the Ollama API directly:import requests

# Ollama API 地址

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

# Request parameters

payload = {

"model": "llama2",

"prompt": "Tell me a joke.",

"stream": False # Do not use streaming returns

}

# Send POST request

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

# Parse response

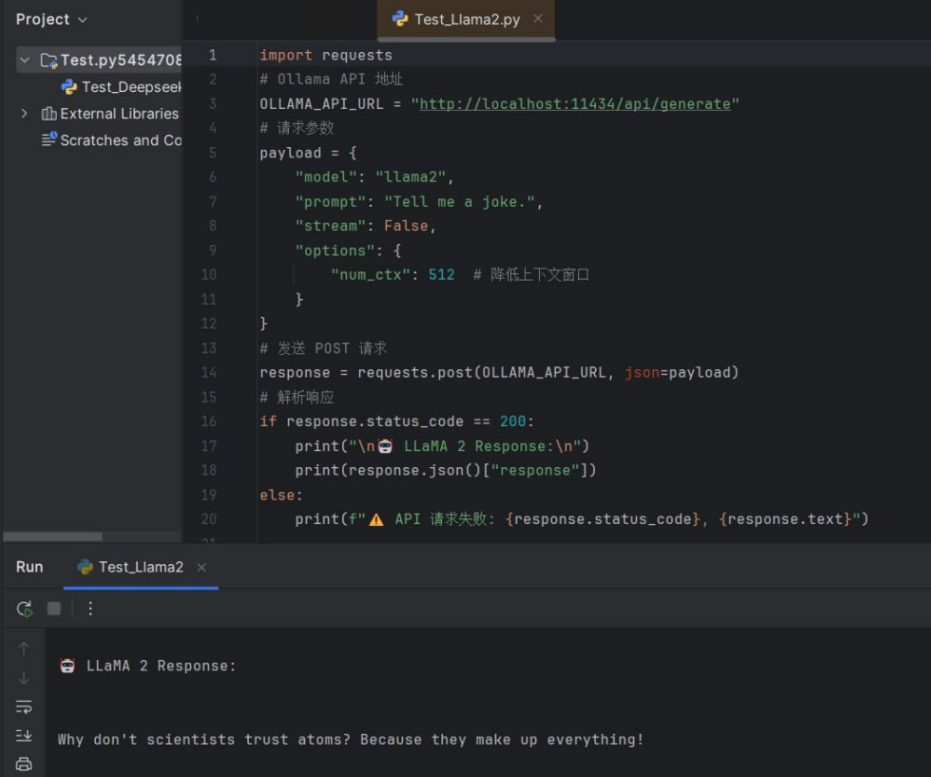
if response.status\_code == 200:

print("\n🤖 LLaMA 2 Response:\n")

print(response.json()["response"])

else:

print(f"⚠️ API request failed: {response.status\_code}, {response.text}")



**4. Conclusion**

After the above steps are completed, you can run LLaMA 2 locally on Windows and deploy it through Docker containerization, while using Python code to call its API for inference.

✅ Ollama’s Docker version is completely free

✅ No API Key required, all calculations are local