1. **In Windows command lines:**

* wsl --install -d Ubuntu
* wsl --set-default-version 2

1. **Insert these lines inside WSL UBUNTU – Basic dependencies:**

* sudo apt update
* sudo apt install -y python3-pip python3-venv git build-essential

1. **Insert these lines inside WSL UBUNTU – create a python venv and install vllm dependencies :**

* python3 -m venv ~/vllm-venv
* source ~/vllm-venv/bin/activate
* pip install --upgrade pip
* pip install "vllm>=0.5.5" "transformers>=4.43" "accelerate" "einops" "autoawq>=0.2.7"

1. **Insert these lines inside WSL UBUNTU – Activate vllm + download and run LLM**

* source ~/vllm-venv/bin/activate
* python -m vllm.entrypoints.openai.api\_server \

--host 0.0.0.0 --port 8000 \

--model Qwen/Qwen2.5-7B-Instruct-AWQ \

--dtype auto \

--quantization awq \

--max-model-len 8192 \

--gpu-memory-utilization 0.90 \

--enforce-eager \

--swap-space 8

1. **Copy this code and Test your vllm setup:**

import requests

import json

q = "در مورد هوش مصنوعی عامل یاایجنتیک توضیح بده"

url = "http://localhost:8000/v1/chat/completions"

payload = {

    "model": "Qwen/Qwen2.5-7B-Instruct-AWQ",

    "messages": [

        {"role": "user", "content": q}

    ],

    "stream": True

}

headers = {

    "Content-Type": "application/json"

}

with requests.post(url, json=payload, headers=headers, stream=True) as r:

    for line in r.iter\_lines():

        if line:

            decoded = line.decode("utf-8")

            if decoded.startswith("data: "):

                data = decoded[len("data: "):]

                if data.strip() == "[DONE]":

                    break

                try:

                    content = json.loads(data)["choices"][0]["delta"].get("content", "")

                    print(content, end="", flush=True)

                except Exception as e:

                    pass

print("\n--- Stream finished ---")