LLaMA 3.3 Installation with Docker – Step-by-Step Guide

This guide provides a streamlined approach to installing and running **LLaMA 3.3** using Docker, Ollama, and the Web GUI on **Windows, macOS, and Linux**.

LLaMA 3.3 Model System Requirements

Model	Quantization	Minimum GPU (VRAM)	Recommended GPU (VRAM)	Minimum CPU	Recommended CPU	Minimum RAM
LLaMA 3.3 8B	4-bit	12GB (RTX 3080)	16GB (RTX 3090)	6-core	8-core	24GB
	8-bit	16GB (RTX 3090, 4090)	24GB (A100, H100)	8-core	10-core	32GB
	F32	32GB+ (A100, H100)	40GB+ (A6000, H200)	12-core	16-core	64GB+

1. Install Required Dependencies

For Windows:

1. Install **Docker Desktop** from the official website.
2. Enable **WSL2 backend** (if using Windows 10/11).
3. Open **PowerShell** and verify Docker installation:

```
docker --version
```

For macOS:

1. Install **Homebrew** if not installed:

```
/bin/bash -c "$(curl -fsSL
https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"
```

2. Install **Docker**:

```
brew install --cask docker
```


3. Verify installation:

```
docker --version
```

For Linux (Ubuntu/Debian-based):

1. Update the package list:

```
sudo apt update && sudo apt upgrade -y
```

2. Install Docker:

```
sudo apt install docker.io -y
```

3. Start and enable Docker service:

```
sudo systemctl enable --now docker
```

4. Verify installation:

```
docker --version
```

2. Run the Ollama Docker Container

```
docker run -d -v ollama:/root/.ollama -p 11434:11434 --name ollama
ollama/ollama
```

3. Log in to the Ollama Container

```
docker exec -it ollama /bin/bash
```

4. Pull the LLaMA 3.3 Model (4-bit quantized by default)

```
ollama pull llama3:8b
```

(Replace **llama3:8b** with **llama3:14b** if running a **14B parameter** version.)

5. Run the LLaMA 3.3 Model

```
ollama run llama3:8b
```

Test the model by entering a prompt:

```
>>> What is the future of AI?
```

6. Deploy the Ollama Web GUI

Replace **<YOUR-IP>** with your **local IP address**:

```
docker run -d -p 3000:8080 -e OLLAMA_BASE_URL=http://<YOUR-IP>:11434 \
-v open-webui:/app/backend/data --name open-webui --restart always \
ghcr.io/open-webui/open-webui:main
```

7. Find Your Local IP Address

For Windows:

```
ipconfig
```

For macOS/Linux:

```
ip a | grep inet
```

8. Access the Web Interface

Open a browser and go to: **http://<YOUR-IP>:3000**

9. Stop or Restart Containers (Optional)

To stop the containers:

```
docker stop ollama open-webui
```


To restart:

```
docker start ollama open-webui
```

To remove everything:

```
docker rm -f ollama open-webui
 docker volume rm ollama open-webui
```

Conclusion

Following these steps allows you to run **LLaMA 3.3** efficiently inside a **Docker container** using **Ollama** with an optional **Web GUI**. This setup provides a flexible, portable, and isolated environment for seamless AI interactions.