Llama3

> **Demo Environment**

Development board: Jetson Orin series motherboard

SSD: 128G

> **Tutorial application scope**: Whether the motherboard can run is related to the available memory of the system. The user's own environment and the programs running in the background may cause the model to fail to run.

Motherboard model	Run directly with Ollama	Run with Open WebUI
Jetson Orin NX 16GB	√	√
Jetson Orin NX 8GB	√	×
Jetson Orin Nano 8GB	√	×
Jetson Orin Nano 4GB	×	×

Meta Llama3 is a series of advanced open source large language models (LLMs) developed by the Meta AI department.

1. Model size

Model	Parameters
Llama3	8B
Llama3	70B

2. Performance

Meta Llama 3 Instruct model performance

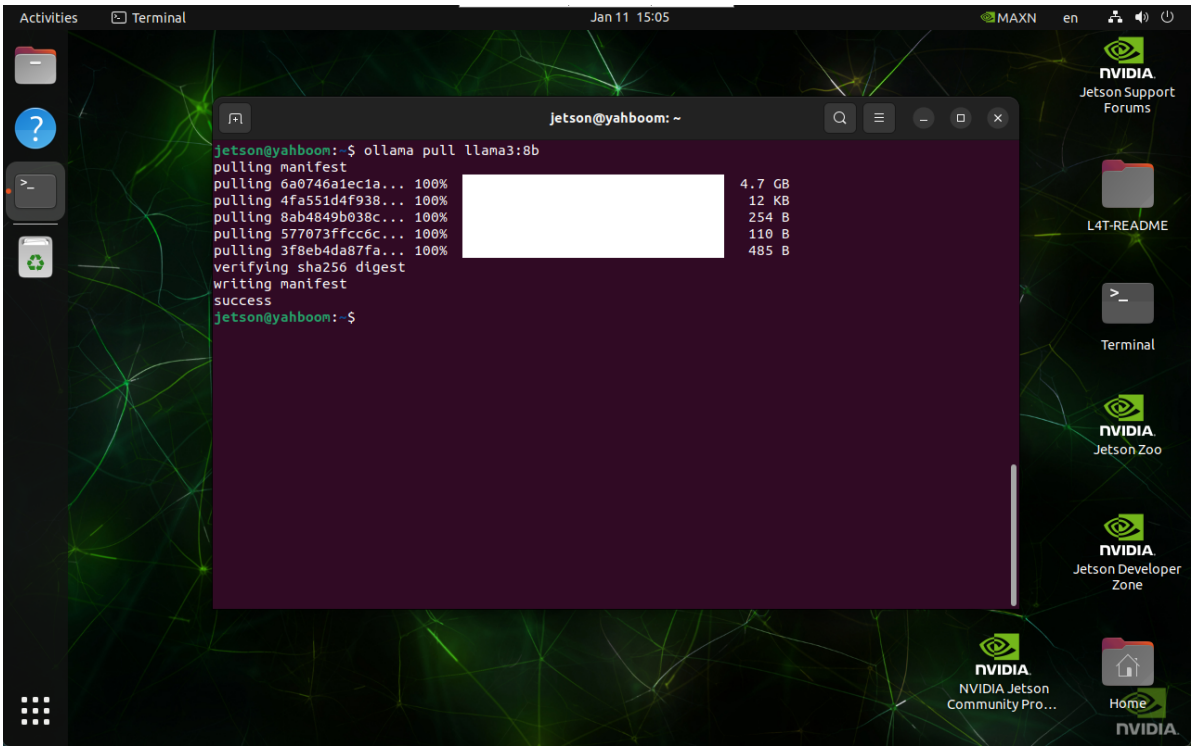
	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	68.4	53.3	58.4
GPQA 0-shot	34.2	21.4	26.3
HumanEval 0-shot	62.2	30.5	36.6
GSM-8K 8-shot, CoT	79.6	30.6	39.9
MATH 4-shot, CoT	30.0	12.2	11.0

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	82.0	81.9	79.0
GPQA 0-shot	39.5	41.5 CoT	38.5 CoT
HumanEval 0-shot	81.7	71.9	73.0
GSM-8K 8-shot, CoT	93.0	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	58.5 Minerva prompt	40.5

3. Pull Llama3

Using the pull command will automatically pull the model of the Ollama model library:

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



4. Use Llama 3

4.1. Run Llama 3

If the system does not have a running model, the system will automatically pull the Llama3 8B model and run it:

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

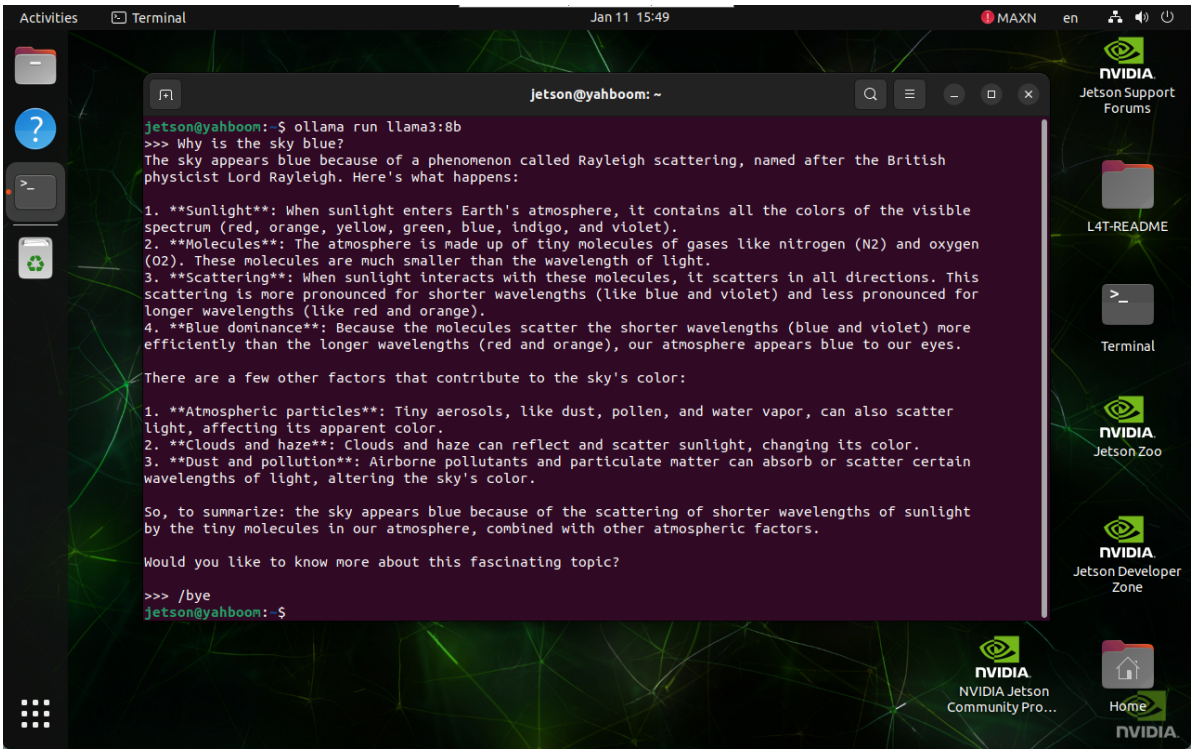
4.2. Have a conversation

```
Why is the sky blue?
```

The time to answer the question depends on the hardware configuration, so be patient!

4.5. End the conversation

Use the `Ctrl+d` shortcut key or `/bye` to end the conversation!



References

> **Ollama**

Official website: https://ollama.com/

GitHub: https://github.com/ollama/ollama

> **Llama 3**

GitHub: https://github.com/meta-llama/llama3

Ollama corresponding model: https://ollama.com/library/llama3