Llama3

> Demonstration environment

Development board: RDK X5 series board

sd card: 64G

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

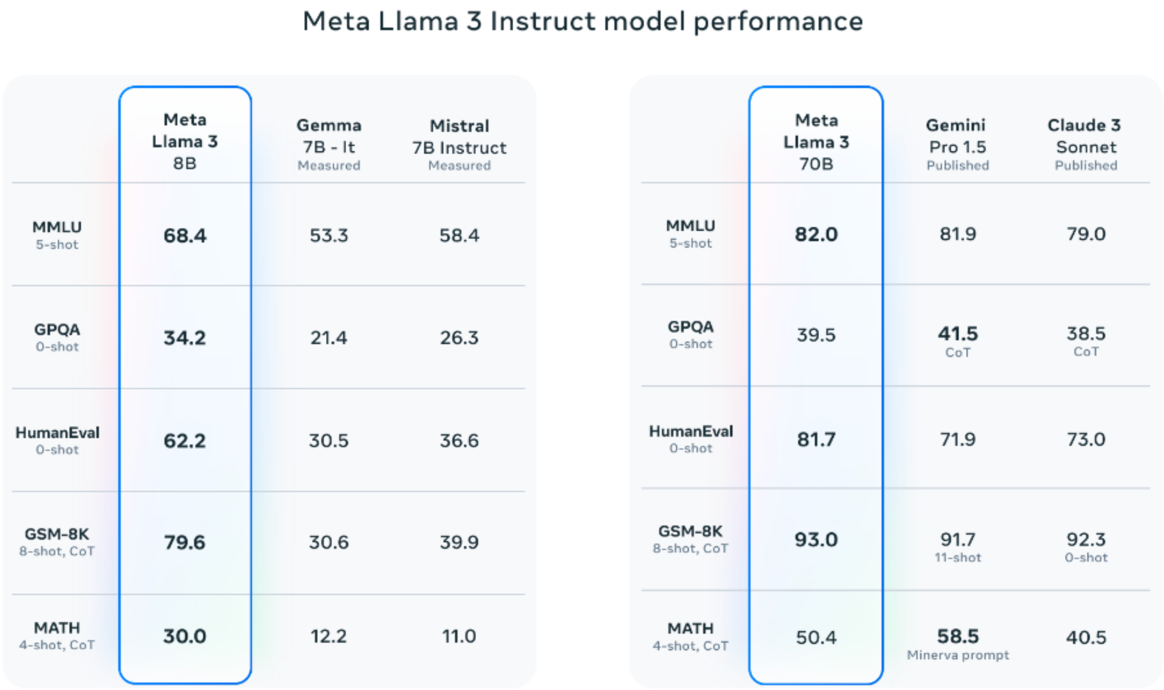
Motherboard model	Ollama direct operation
rdk x5 8GB	√
rdk x5 4GB	×

Meta Llama3 is a family of advanced open source Large Language Models (LLMs) developed by the Meta AI division.

1、Model scale

Model	Parameter
Llama3	8B

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

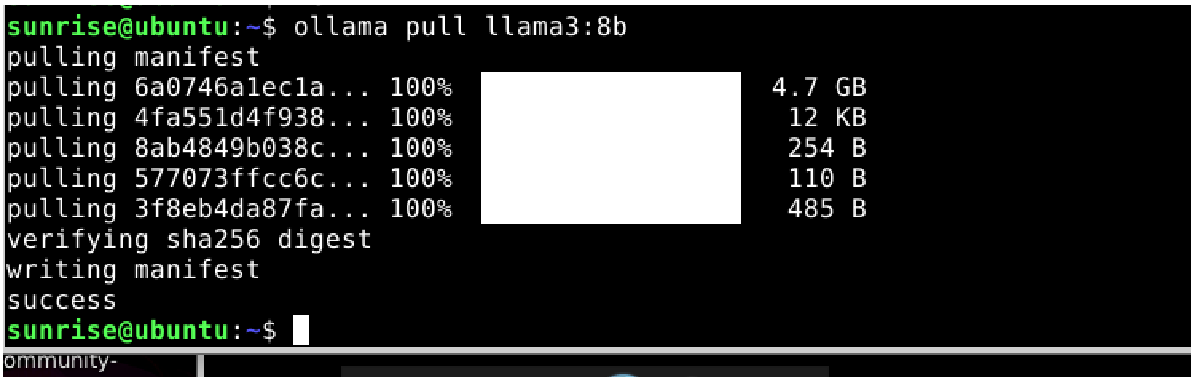
	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	**82.0**	81.9	79.0
GPQA 0-shot	39.5	**41.5** CoT	38.5 CoT
HumanEval 0-shot	**81.7**	71.9	73.0
GSM-8K 8-shot, CoT	**93.0**	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	**58.5** Minerva prompt	40.5

3. Run Llama3

Only **rdk x5 8Gb motherboard** can run this model

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

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



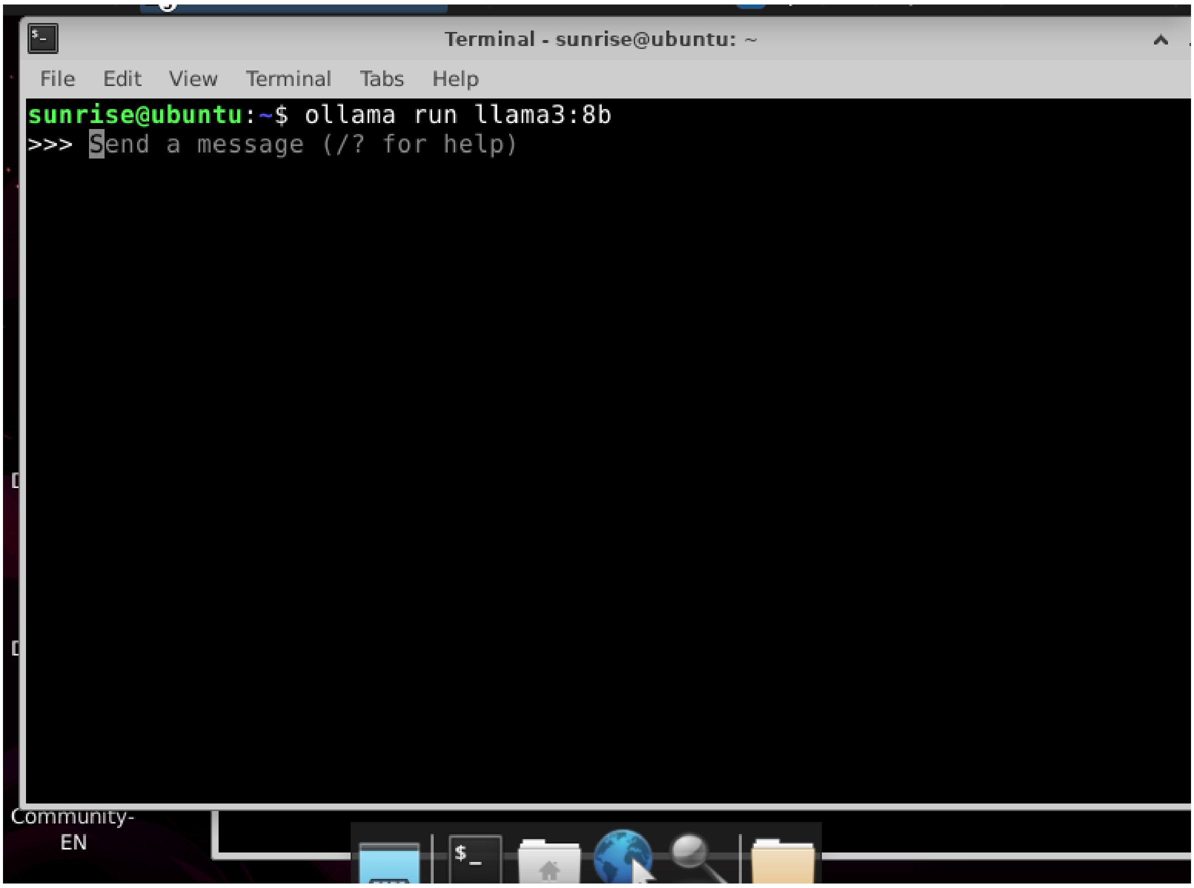
```
sunrise@ubuntu:~$ ollama pull llama3:8b
pulling manifest
pulling 6a0746a1ec1a... 100%                  4.7 GB
pulling 4fa551d4f938... 100%                   12 KB
pulling 8ab4849b038c... 100%                  254 B
pulling 577073ffcc6c... 100%                  110 B
pulling 3f8eb4da87fa... 100%                  485 B
verifying sha256 digest
writing manifest
success
sunrise@ubuntu:~$
```

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
```

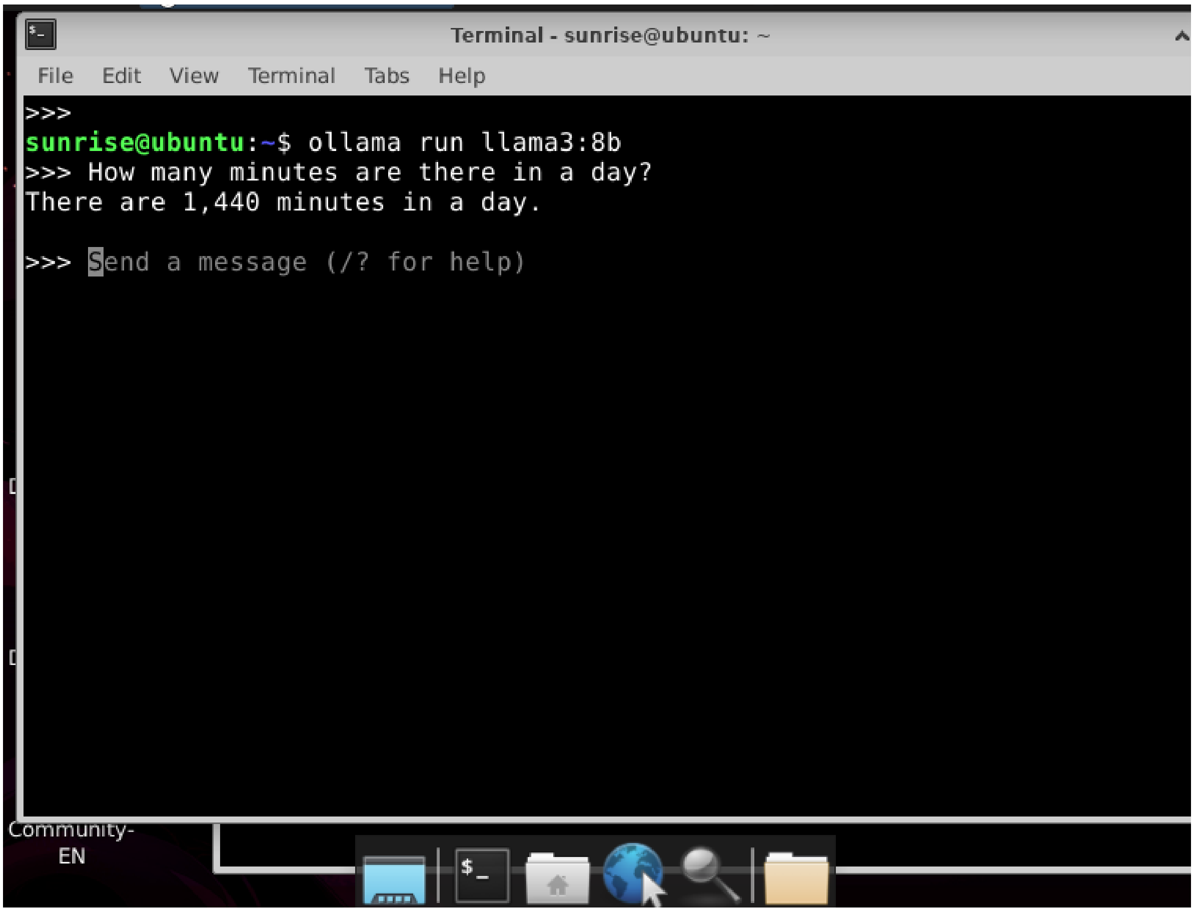


```
sunrise@ubuntu:~$ ollama run llama3:8b
>>> Send a message (/? for help)
```

4.2. Have a conversation

```
How many minutes are there in a day?
```

The response time is related to the hardware configuration, please be patient!

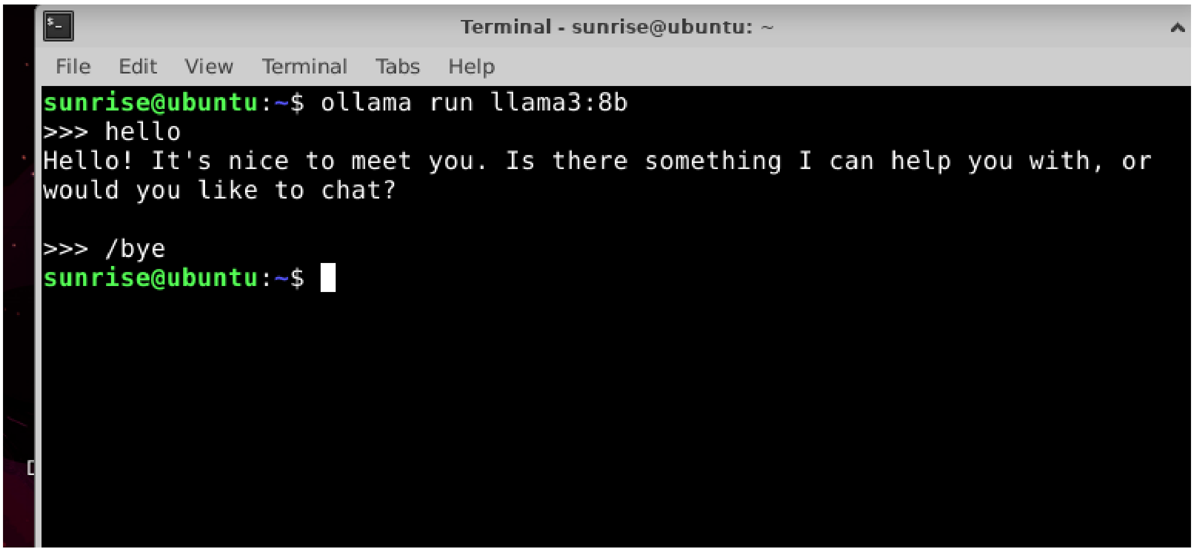


```
>>>
sunrise@ubuntu:~$ ollama run llama3:8b
>>> How many minutes are there in a day?
There are 1,440 minutes in a day.

>>> Send a message (/? for help)
```

4.5. End the conversation

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



Terminal - sunrise@ubuntu: ~
File Edit View Terminal Tabs Help

```
sunrise@ubuntu:~$ ollama run llama3:8b
>>> hello
Hello! It's nice to meet you. Is there something I can help you with, or
would you like to chat?

>>> /bye
sunrise@ubuntu:~$
```

References

> **Ollama**

Website: https://ollama.com/

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

> **Llama 3**

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

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