Qwen2

> **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	√ (need to run the small parameter version)	√ (need to run the small parameter version)

Alibaba Qwen2 is an advanced open source large-scale language model developed by Alibaba, designed to provide powerful natural language processing capabilities.

1. Model scale

Model	parameter
Qwen2	0.5B
Qwen2	1.5B
Qwen2	7B
Qwen2	72B

2. Performance

	Qwen2-7B Instruct	Llama3-8B Instruct	GLM4-9B Chat
AlignBench	**7.21**	6.20	7.01
MT-Bench	**8.41**	8.05	8.35
MMLU	70.5	68.4	72.4
GSM8K	**82.3**	79.6	79.6
MATH	49.6	30.0	50.6
HumanEval	**79.9**	62.2	71.8
C-Eval	**77.2**	45.9	75.6

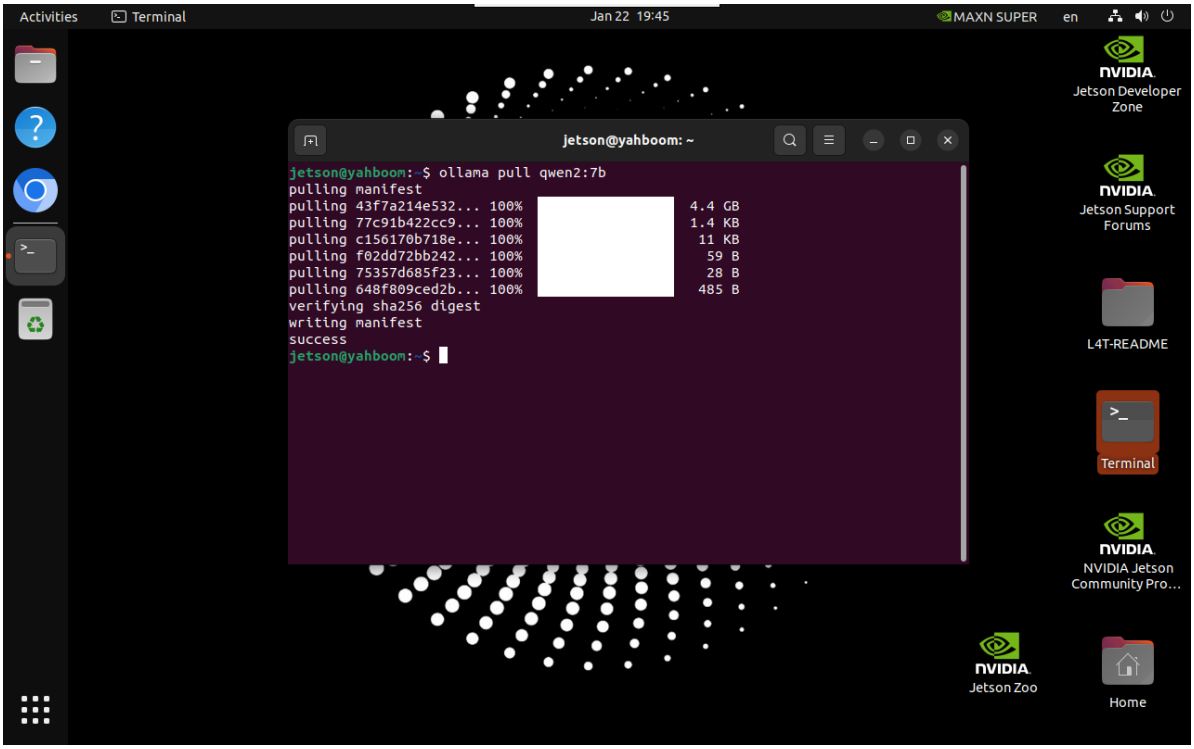
3. Pull Qwen2

Use the pull command to automatically pull the model of the Ollama model library:

```
ollama pull qwen2:7b
```

Small parameter version model: motherboards with 8G or less memory can run this

```
ollama pull qwen2:1.5b
```



4. Use Qwen2

4.1. Run Qwen2

If the system does not have a running model, the system will automatically pull Qwen2 7B model and run:

```
ollama run qwen2:7b
```

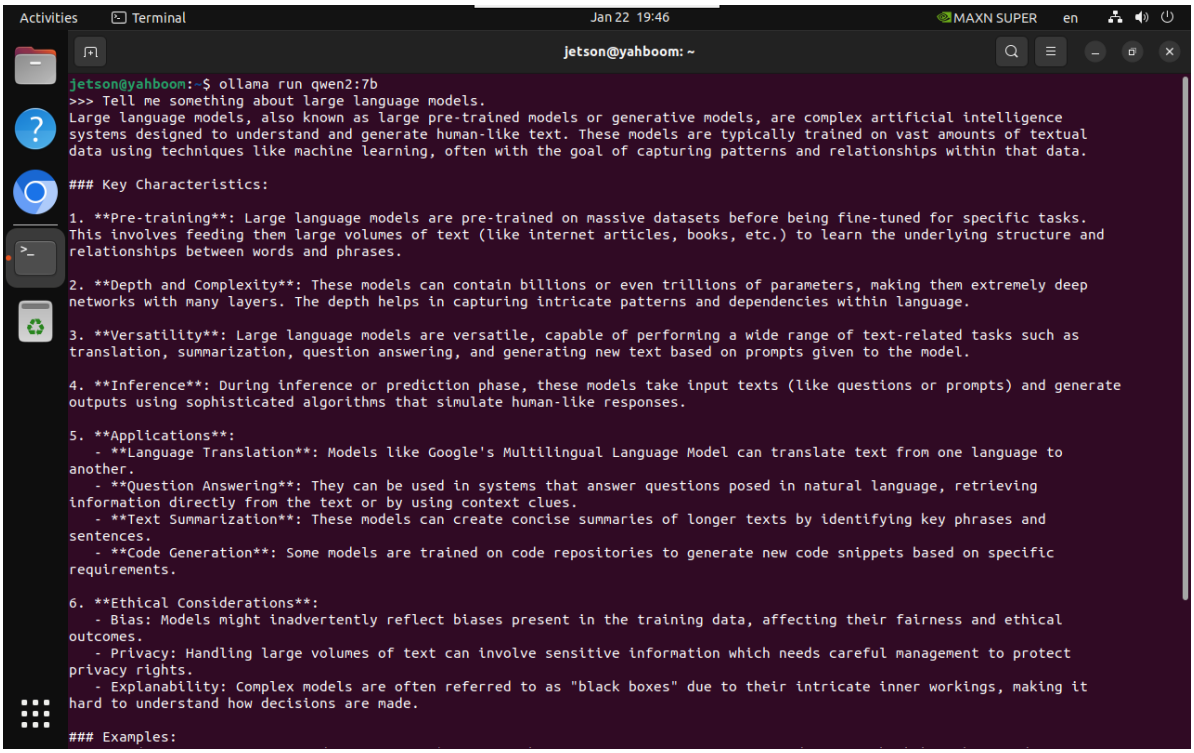
Small parameter version model: motherboards with 8G memory or less can run this

```
ollama run qwen2:1.5b
```

4.2. Have a conversation

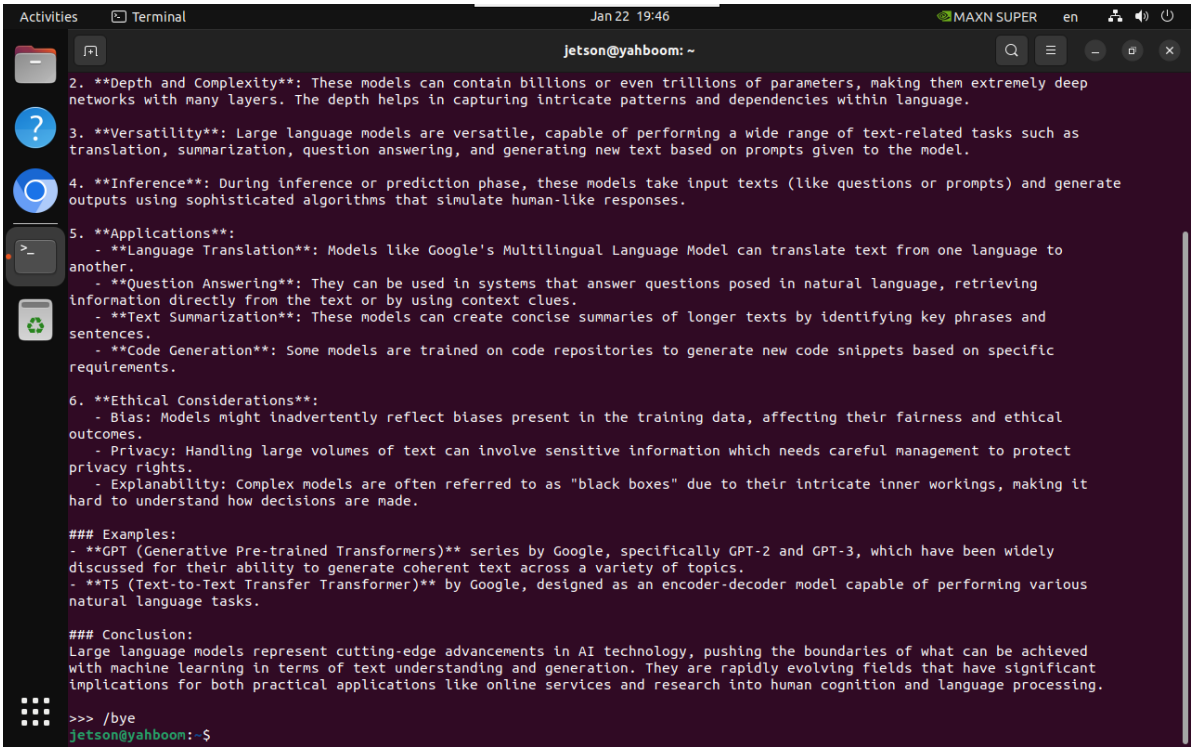
```
Tell me something about large language models.
```

The time to reply to the question depends on the hardware configuration, please be patient!



4.3. 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

> **Qwen2**

GitHub: https://github.com/QwenLM/Qwen2

Ollama corresponding model: https://ollama.com/library/qwen2