Qwen2 model

> **Demonstration environment**

Development Board : Raspberry Pi 5B

SD(TF)card : 64G (Above 16G, the larger the capacity, the more models can be experienced)

```
Raspberry Pi 5B (8G RAM): Run 8B and below parameter models
Raspberry Pi 5B (4G RAM): Run 3B and below parameter models
Raspberry Pi 5B (2G RAM): Run 0.5B and below parameter models
```

Model scale

Model	参数
Qwen2	0.5B
Qwen2	1.5B
Qwen2	7B
Qwen2	72B

```
Raspberry Pi 5B (8G RAM): Test Qwen2 model with parameters of 7B and below.
Raspberry Pi 5B (4G RAM): Test Qwen2 model with parameters of 1.5B and below.
```
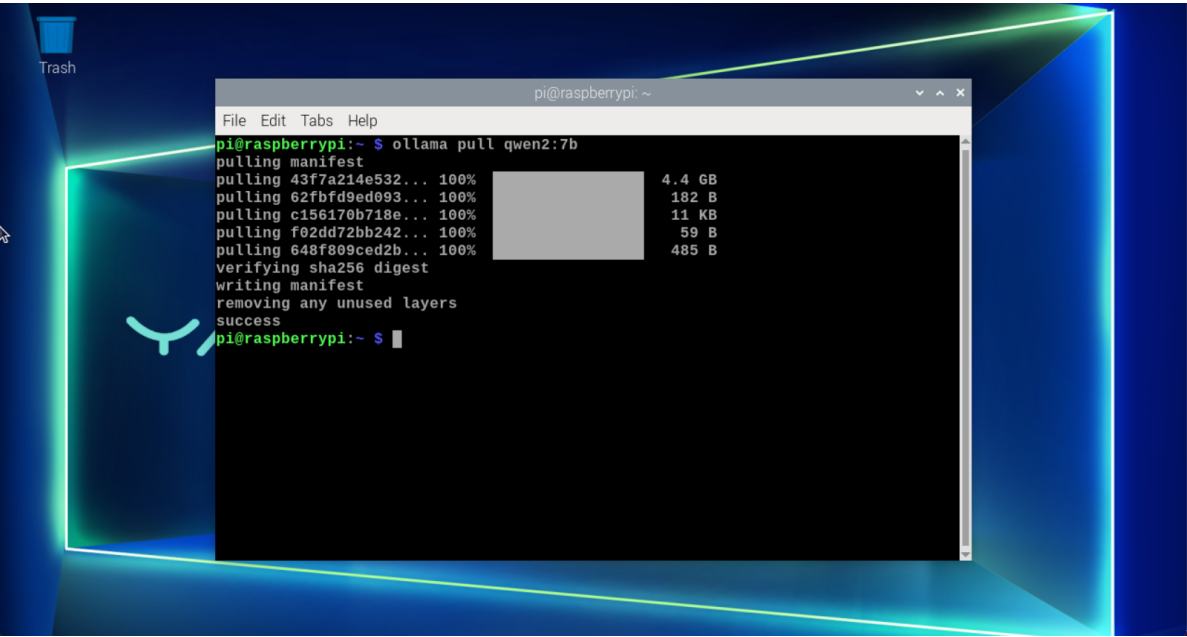
Performance performance

	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

Got Qwen2

Using the pull command will automatically pull the models from the Ollama model library.
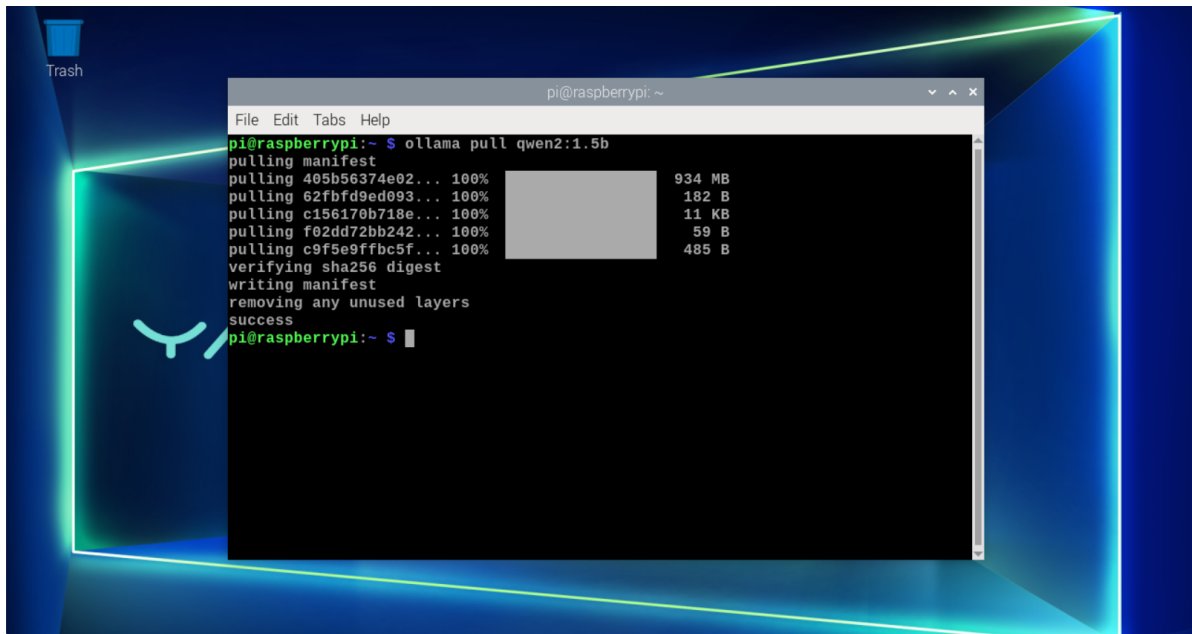
> **Raspberry Pi 5B (8G RAM)**

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


> **Raspberry Pi 5B (4G RAM)**

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

Use Qwen2

Run Qwen2

> **Raspberry Pi 5B (8G RAM)**

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

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

> **Raspberry Pi 5B (4G RAM)**

If the system does not have a running model, the system will automatically pull the Qwen2 1.5B model and run it.

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

Dialogue

> **Raspberry Pi 5B (8G RAM)**

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

The response time to the question is related to the hardware configuration, please be patient and wait.

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

The response time to the question is related to the hardware configuration, please be patient and wait.

End conversation

You can end the conversation by using the shortcut key 'Ctrl+d' or '/bye'.

Raspberry Pi 5B (4G RAM)

Reference material

Ollama

Website： https://ollama.com/

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

Qwen2

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

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