

# Qwen2

## Qwen2

- 1、Model scale
  - 2、Performance
  - 3. Pull Qwen2
  - 4. Use Qwen2
    - 4.1. Run Qwen2
    - 4.2. Have a conversation
    - 4.3. End the conversation
- References

### Demonstration environment

**Development board:** rdk x5 motherboard

**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	√（需运行小参数版本）

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

## 2、Performance

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

### 3. Pull Qwen2

Using the pull command will automatically pull the model from the Ollama model library.

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

```
sunrise@ubuntu:~$ ollama pull qwen2:7b
pulling manifest
pulling 43f7a214e532... 100% 4.4 GB
pulling 77c91b422cc9... 100% 1.4 KB
pulling c156170b718e... 100% 11 KB
pulling f02dd72bb242... 100% 59 B
pulling 75357d685f23... 100% 28 B
pulling 648f809ced2b... 100% 485 B
verifying sha256 digest
writing manifest
success
```

Small parameter version model.

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

```
Applications 2 Xfce Terminal
Terminal - sunrise@ubuntu: ~
File Edit View Terminal Tabs Help
sunrise@ubuntu:~$ ollama pull qwen2:1.5b
pulling manifest
pulling 405b56374e02... 100% 934 MB
pulling 62fbfd9ed093... 100% 182 B
pulling c156170b718e... 100% 11 KB
pulling f02dd72bb242... 100% 59 B
pulling c9f5e9ffbc5f... 100% 485 B
verifying sha256 digest
writing manifest
success
sunrise@ubuntu:~$
```

## 4. Use Qwen2

### 4.1. Run Qwen2

8gb motherboard runs the following parameters and smaller models:

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
```

```
sunrise@ubuntu:~$ ollama run qwen2:7b
>>> hello
Hello! How can I help you today? If you have any questions or need
assistance with something, feel free to ask. Whether it's about
technology, science, history, language, or anything else within my
knowledge base, I'm here to provide information and guidance.
>>> Send a message (/? for help)
```

A motherboard with 4 GB of RAM can run this model and smaller models.

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

```
Terminal - sunrise@ubuntu: ~
File Edit View Terminal Tabs Help
sunrise@ubuntu:~$ ollama run qwen2:1.5b
>>> Who are you?
I am an AI language model, designed to understand and respond to a wide
range of questions and tasks. I use natural language processing
techniques to generate text based on the input I receive. My purpose is
to assist people in finding information, offering help with
language-related tasks, and providing insights into various topics.
>>> Send a message (/? for help)
```

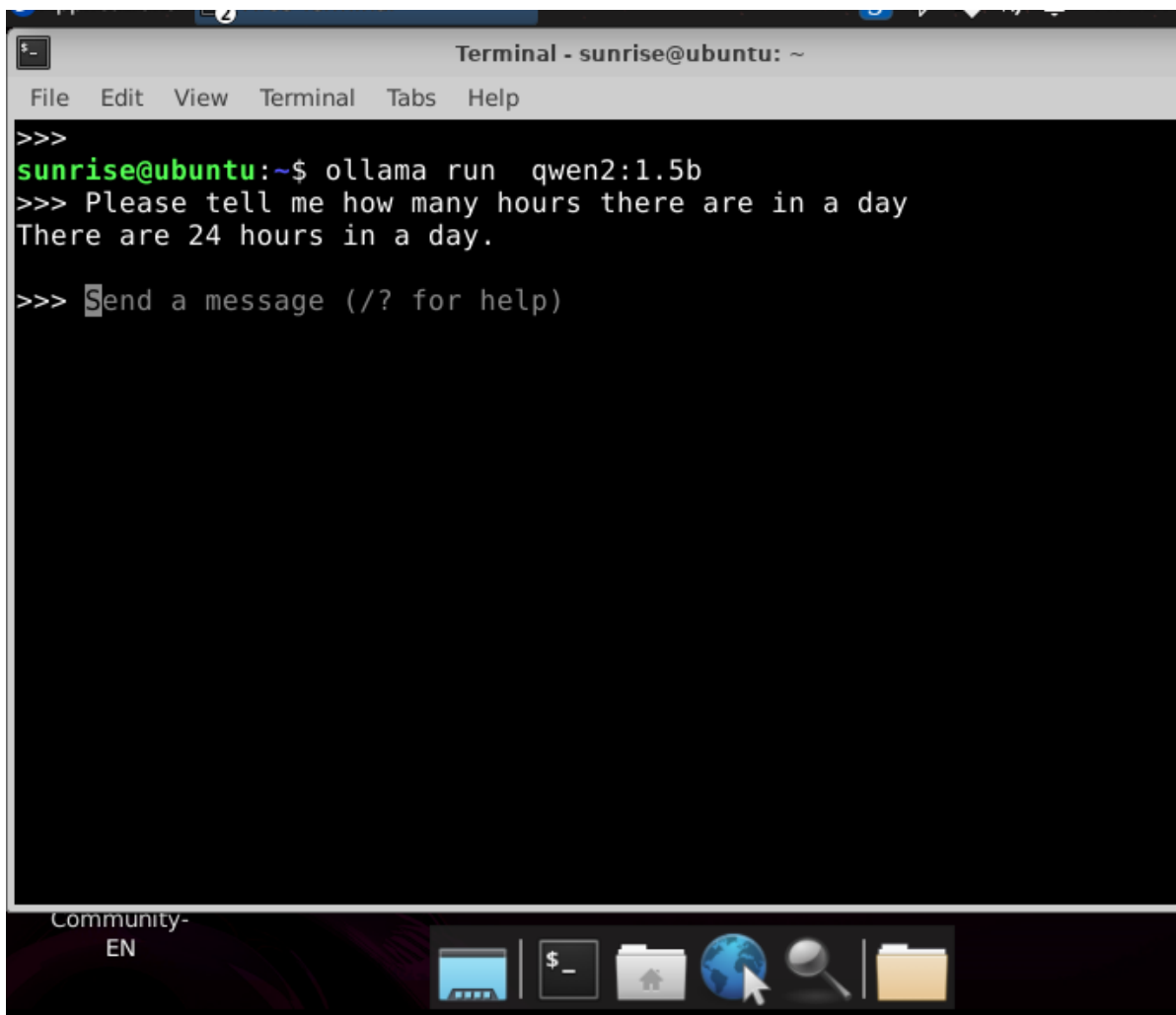
## 4.2. Have a conversation

Please tell me how many hours there are in a day

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

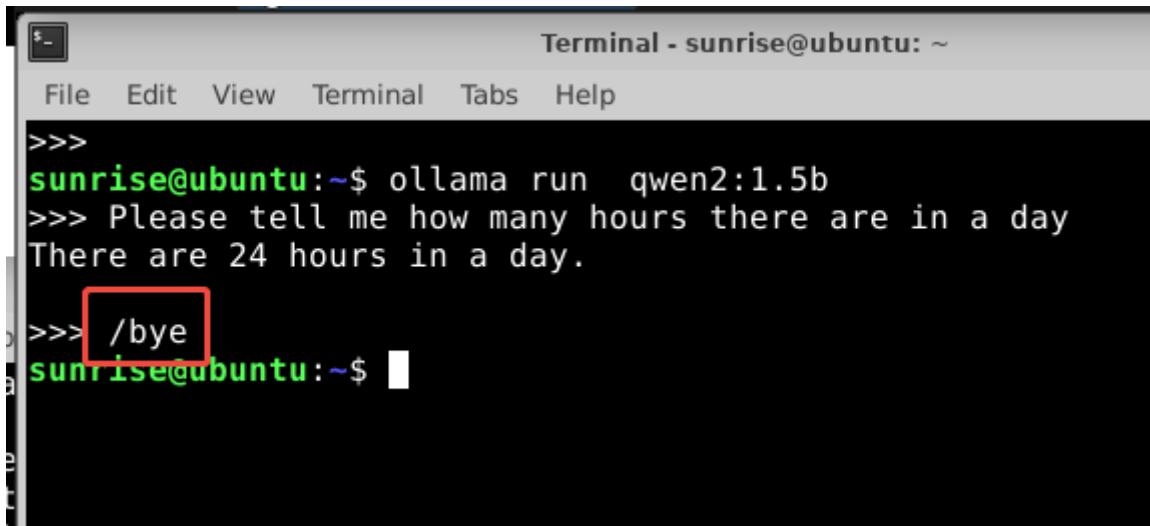
```
Terminal - sunrise@ubuntu: ~
File Edit View Terminal Tabs Help
>>>
sunrise@ubuntu:~$ ollama run qwen2:1.5b
>>> Please tell me how many hours there are in a day
There are 24 hours in a day.
>>> Send a message (/? for help)
```

Community-EN



### 4.3. End the conversation

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

A screenshot of a terminal window titled "Terminal - sunrise@ubuntu: ~". The terminal shows a command prompt where the user has run "ollama run qwen2:1.5b". The model responds with "Please tell me how many hours there are in a day" and "There are 24 hours in a day." The user then enters "/bye", which is highlighted with a red rectangle. The terminal returns to the command prompt "sunrise@ubuntu:~\$".

```
>>>
sunrise@ubuntu:~$ ollama run qwen2:1.5b
>>> Please tell me how many hours there are in a day
There are 24 hours in a day.
>>> /bye
sunrise@ubuntu:~$
```

## References

### 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>