

Qwen2 model

Qwen2 model

[Model scale](#)

[Performance](#)

[Got Qwen2](#)

[Use Qwen2](#)

[Run Qwen2](#)

[Have a conversation](#)

[End the conversation](#)

[References](#)

Demo environment

Development boards: Jetson Nano

SD(TF) card: 64G

It is recommended to run the 4B and below parameter models

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

Model scale

Model	Parameter
Qwen2	0.5B
Qwen2	1.5B
Qwen2	7B
Qwen2	72B

Jetson Nano: Tested with Qwen2 model with parameters 1.5B and below

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

Use the pull command to automatically obtain the model of the Ollama model library.

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

```

jetson@jetson-desktop: ~
jetson@jetson-desktop:~$ 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
removing any unused layers
success
jetson@jetson-desktop:~$

```

Use Qwen2

Run Qwen2

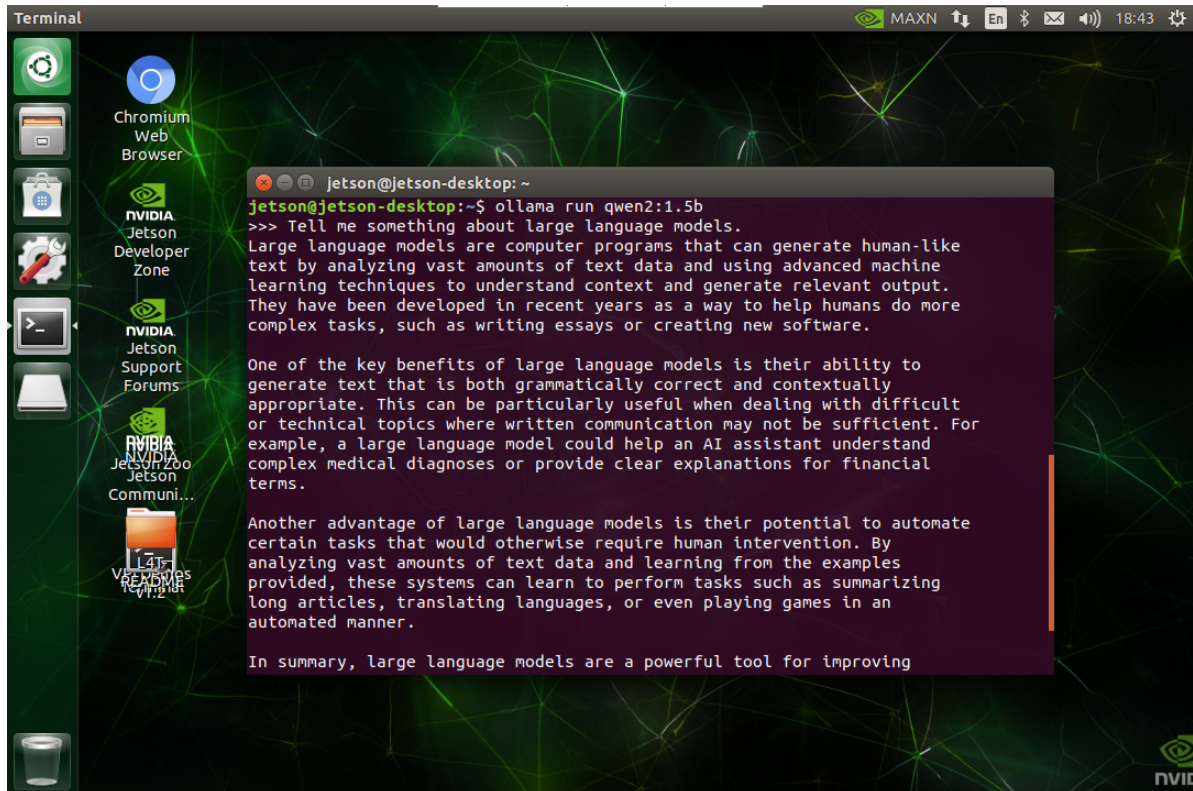
If the system does not have a running model, the system will automatically pull the Qwen2 1.5B model and run it.

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

Have a conversation

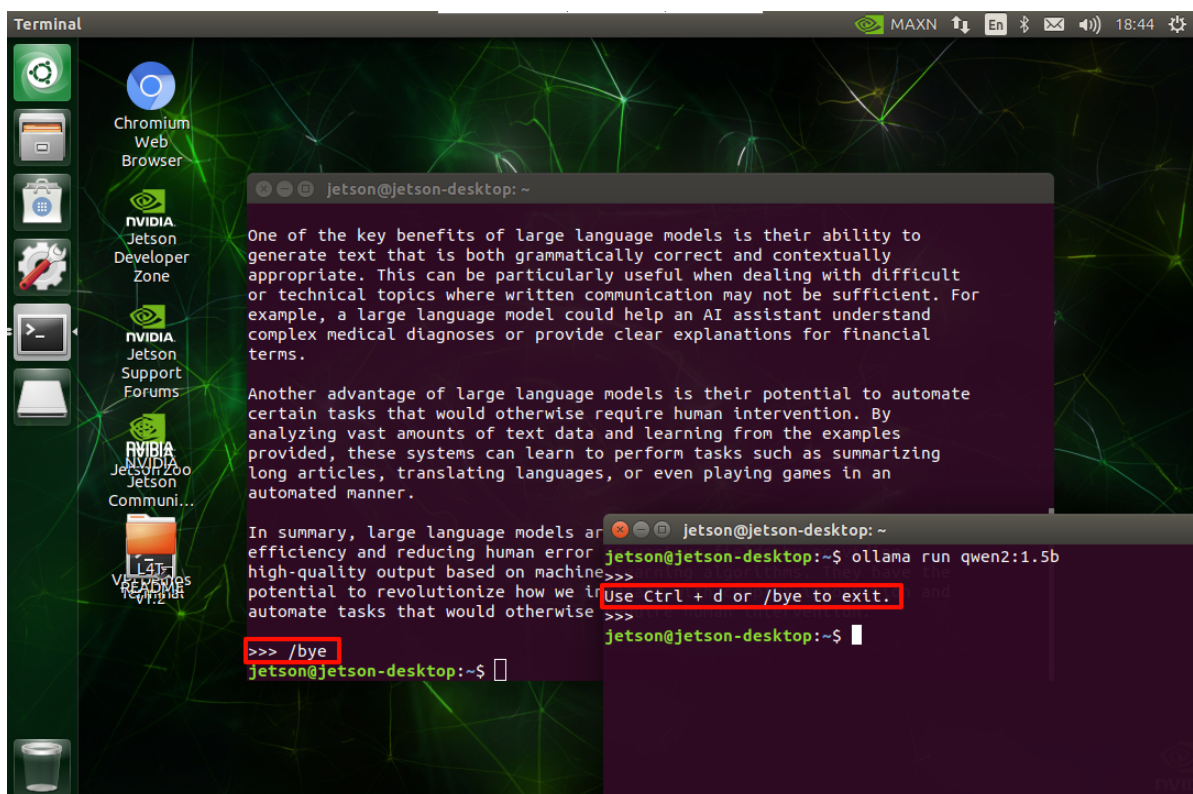
Tell me something about large language models.

The response time is related to the hardware configuration, please wait patiently.



End the conversation

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



References

Ollama

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>