

Qwen2

Qwen2

- Model size
- Performance
- Pull Qwen2
- Use Qwen2
 - Run Qwen2
 - Have a conversation
 - End the conversation
- References

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	Ollama	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 language model developed by Alibaba, designed to provide powerful natural language processing capabilities.

Model size

Model	Parameters
Qwen2	0.5B
Qwen2	1.5B
Qwen2	7B
Qwen2	72B

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

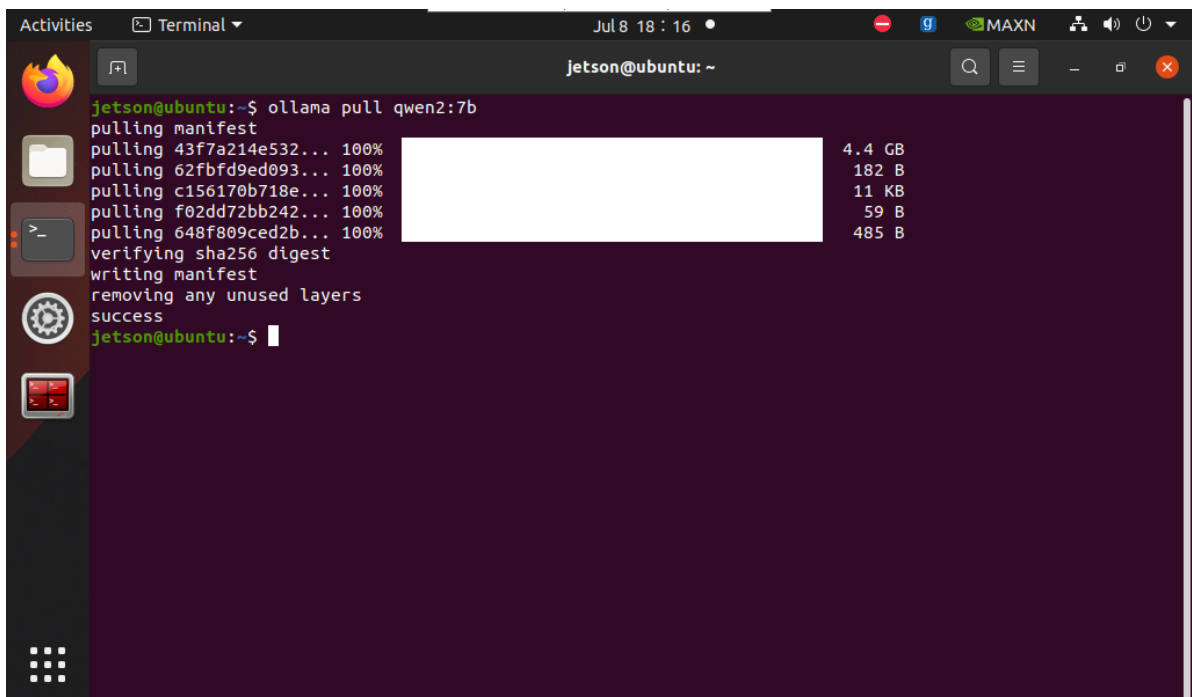
Pull Qwen2

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

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

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

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



```
jetson@ubuntu:~$ ollama pull qwen2:7b
pulling manifest
pulling 43f7a214e532... 100% 4.4 GB
pulling 62fbfd9ed093... 100% 182 B
pulling c156170b718e... 100% 11 KB
pulling f02dd72bb242... 100% 59 B
pulling 648f809ced2b... 100% 485 B
verifying sha256 digest
writing manifest
removing any unused layers
success
jetson@ubuntu:~$
```

Use Qwen2

Run Qwen2

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
```

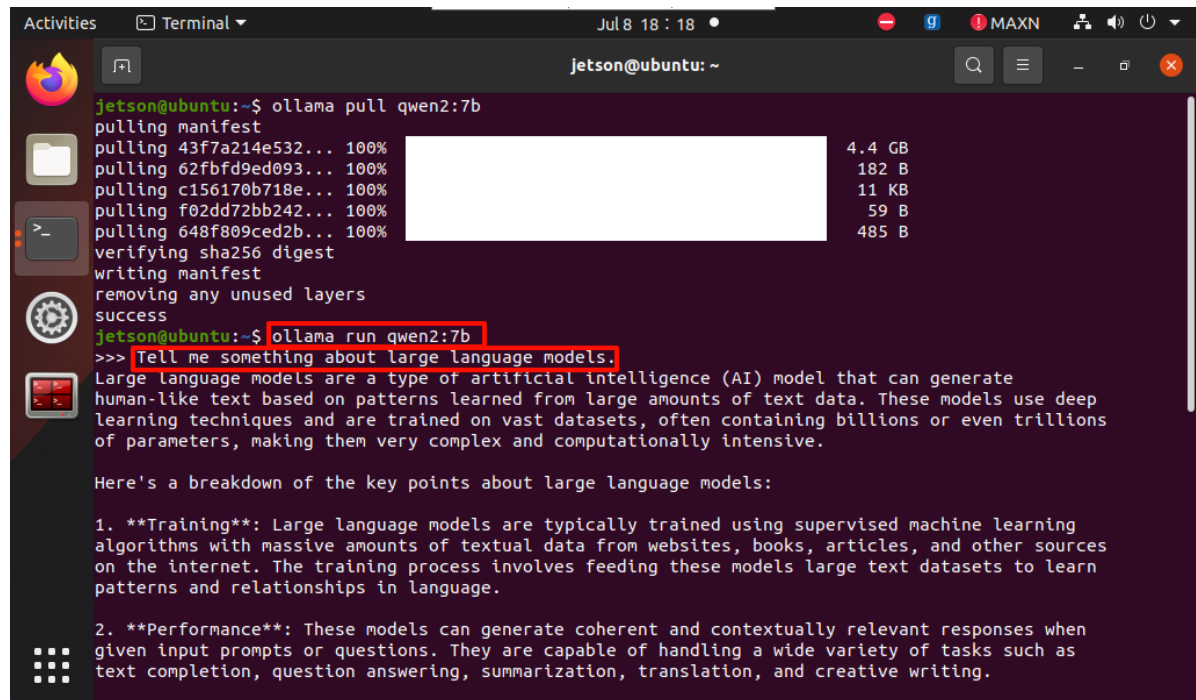
Small parameter version model: motherboards with 8G or less memory can run this

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

Have a conversation

Tell me something about large language models.

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



```
jetson@ubuntu:~$ ollama pull qwen2:7b
pulling manifest
pulling 43f7a214e532... 100% 4.4 GB
pulling 62fbfd9ed093... 100% 182 B
pulling c156170b718e... 100% 11 KB
pulling f02dd72bb242... 100% 59 B
pulling 648f809ced2b... 100% 485 B
verifying sha256 digest
writing manifest
removing any unused layers
success
jetson@ubuntu:~$ ollama run qwen2:7b
>>> Tell me something about large language models.
Large language models are a type of artificial intelligence (AI) model that can generate human-like text based on patterns learned from large amounts of text data. These models use deep learning techniques and are trained on vast datasets, often containing billions or even trillions of parameters, making them very complex and computationally intensive.


Here's a breakdown of the key points about large language models:

1. Training: Large language models are typically trained using supervised machine learning algorithms with massive amounts of textual data from websites, books, articles, and other sources on the internet. The training process involves feeding these models large text datasets to learn patterns and relationships in language.

2. Performance: These models can generate coherent and contextually relevant responses when given input prompts or questions. They are capable of handling a wide variety of tasks such as text completion, question answering, summarization, translation, and creative writing.
```

End the conversation

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



```
text completion, question answering, summarization, translation, and creative writing.

3. Limitations: Despite their impressive capabilities, large language models still have limitations. They might occasionally produce outputs that are factually incorrect, biased due to the training data they were exposed to, or simply because of the inherent stochastic nature of these models. Moreover, their computational requirements can be quite high and not feasible for many applications without significant resources.

4. Applications: Large language models find use in a variety of fields including natural language processing (NLP), customer service chatbots, content generation, text summarization, and more. They are also used as foundational components or building blocks in other AI systems that require language understanding and reasoning capabilities.

5. Ethical Considerations: The development and deployment of large language models raise ethical concerns related to issues such as bias in the training data, privacy violations due to the processing of personal information contained within the text, and the impact on job displacement by automating tasks traditionally done by human workers.

6. Advancements: Research into improving these models is ongoing, with a focus on making them more efficient, less biased, and capable of handling more complex language tasks like understanding sarcasm, humor, and context-specific nuance.

In summary, large language models are powerful tools that can help us better understand human language in ways that were previously impossible. However, it's important to be aware of their limitations and the ethical implications of their use and societal implications.

jetson@ubuntu:~$ ollama run qwen2:7b
>>> Use Ctrl + d or /bye to exit.
jetson@ubuntu:~$
>>> /bye
jetson@ubuntu:~$
```

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>