

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

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
```

The screenshot shows a terminal window titled 'jetson@yahboom: ~' with the following output from the command 'ollama pull qwen2:7b':

```
jetson@yahboom:~$ 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
jetson@yahboom:~$
```

The desktop background is dark green with a network-like pattern. On the right side, there are several NVIDIA-related icons: 'NVIDIA Jetson Support Forums', 'L4T-README', 'Terminal', 'NVIDIA Jetson Zoo', 'NVIDIA Jetson Developer Zone', 'NVIDIA NVIDIA Jetson Community Pro...', and 'Home'.

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
```

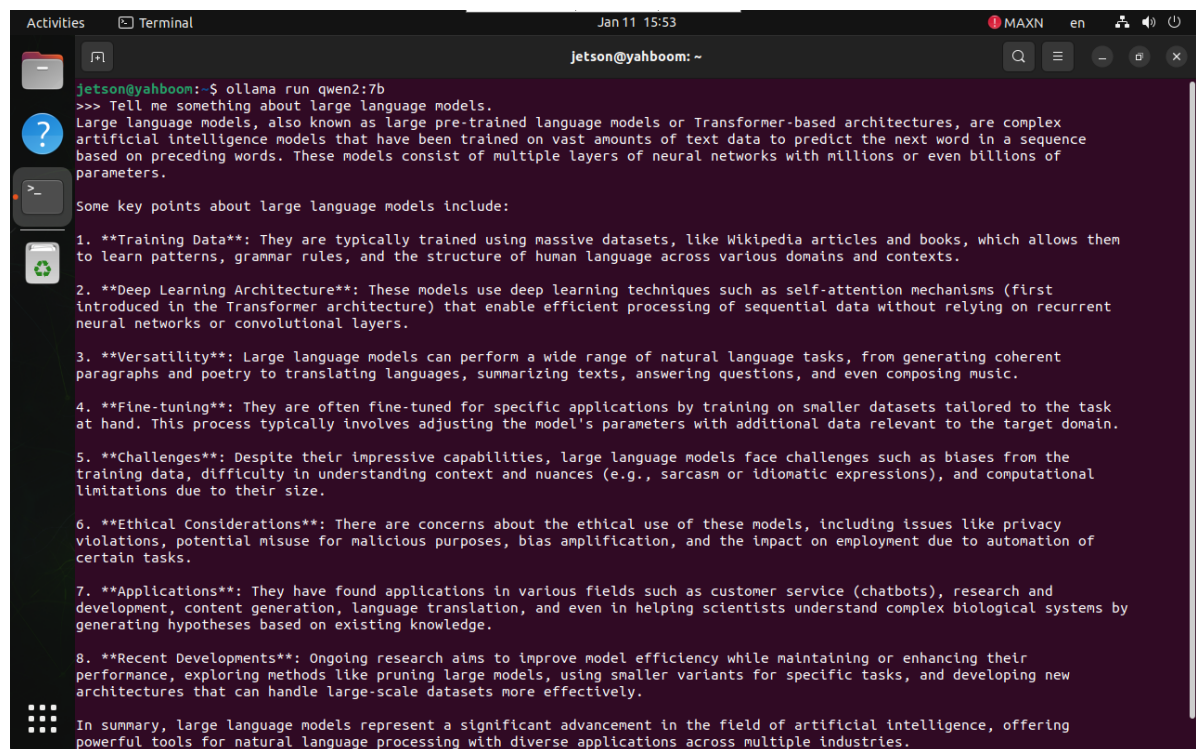
Small parameter version model: motherboards with 8G memory or less can run this

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

4.2. Have a conversation

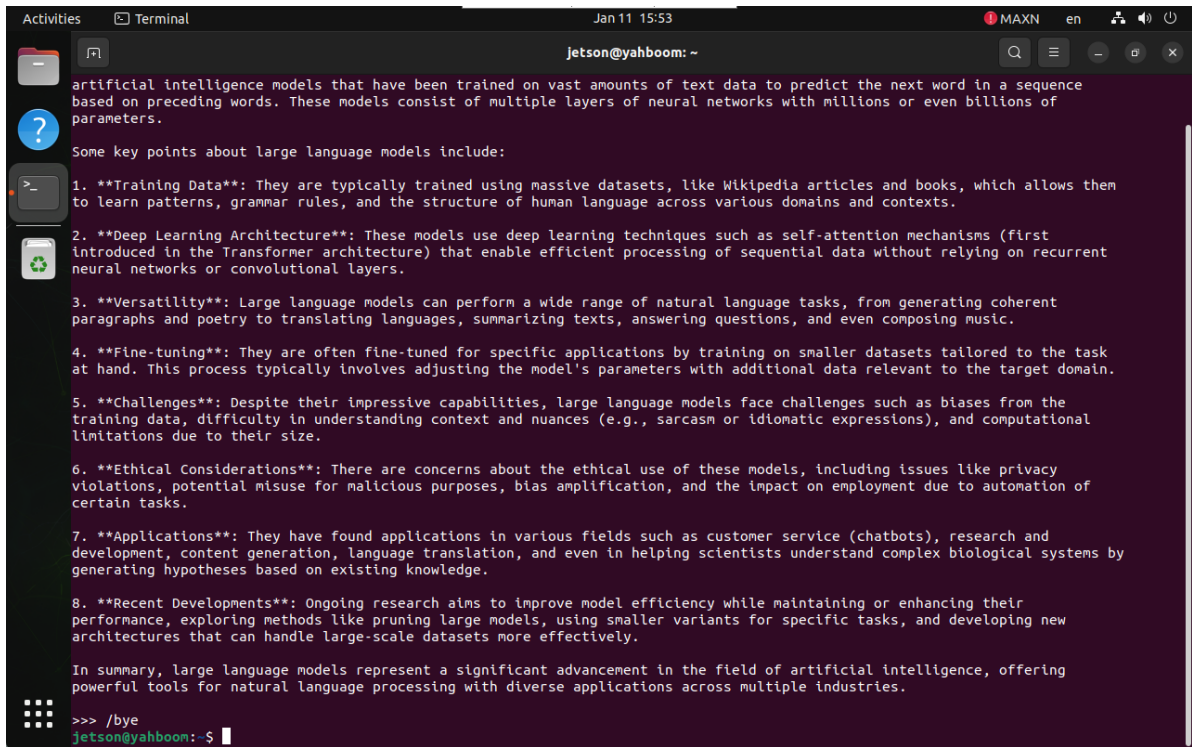
Tell me something about large language models.

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!



```
Activities Terminal Jan 11 15:53 jetson@yahboom: ~
artificial intelligence models that have been trained on vast amounts of text data to predict the next word in a sequence
based on preceding words. These models consist of multiple layers of neural networks with millions or even billions of
parameters.

Some key points about large language models include:

1. **Training Data**: They are typically trained using massive datasets, like Wikipedia articles and books, which allows them
to learn patterns, grammar rules, and the structure of human language across various domains and contexts.

2. **Deep Learning Architecture**: These models use deep learning techniques such as self-attention mechanisms (first
introduced in the Transformer architecture) that enable efficient processing of sequential data without relying on recurrent
neural networks or convolutional layers.

3. **Versatility**: Large language models can perform a wide range of natural language tasks, from generating coherent
paragraphs and poetry to translating languages, summarizing texts, answering questions, and even composing music.

4. **Fine-tuning**: They are often fine-tuned for specific applications by training on smaller datasets tailored to the task
at hand. This process typically involves adjusting the model's parameters with additional data relevant to the target domain.

5. **Challenges**: Despite their impressive capabilities, large language models face challenges such as biases from the
training data, difficulty in understanding context and nuances (e.g., sarcasm or idiomatic expressions), and computational
limitations due to their size.

6. **Ethical Considerations**: There are concerns about the ethical use of these models, including issues like privacy
violations, potential misuse for malicious purposes, bias amplification, and the impact on employment due to automation of
certain tasks.

7. **Applications**: They have found applications in various fields such as customer service (chatbots), research and
development, content generation, language translation, and even in helping scientists understand complex biological systems by
generating hypotheses based on existing knowledge.

8. **Recent Developments**: Ongoing research aims to improve model efficiency while maintaining or enhancing their
performance, exploring methods like pruning large models, using smaller variants for specific tasks, and developing new
architectures that can handle large-scale datasets more effectively.

In summary, large language models represent a significant advancement in the field of artificial intelligence, offering
powerful tools for natural language processing with diverse applications across multiple industries.

>>> /bye
jetson@yahboom: $
```

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>