

# 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	<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

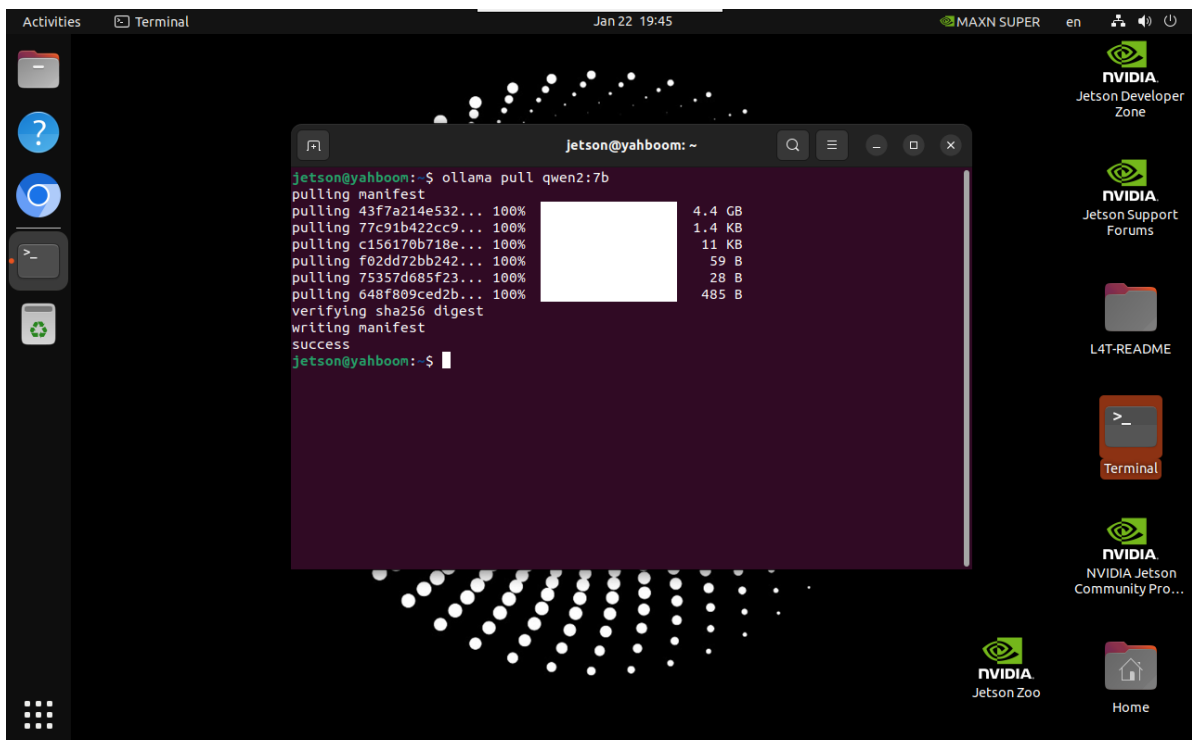
## 3. Pull Qwen2

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
```



## 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
```

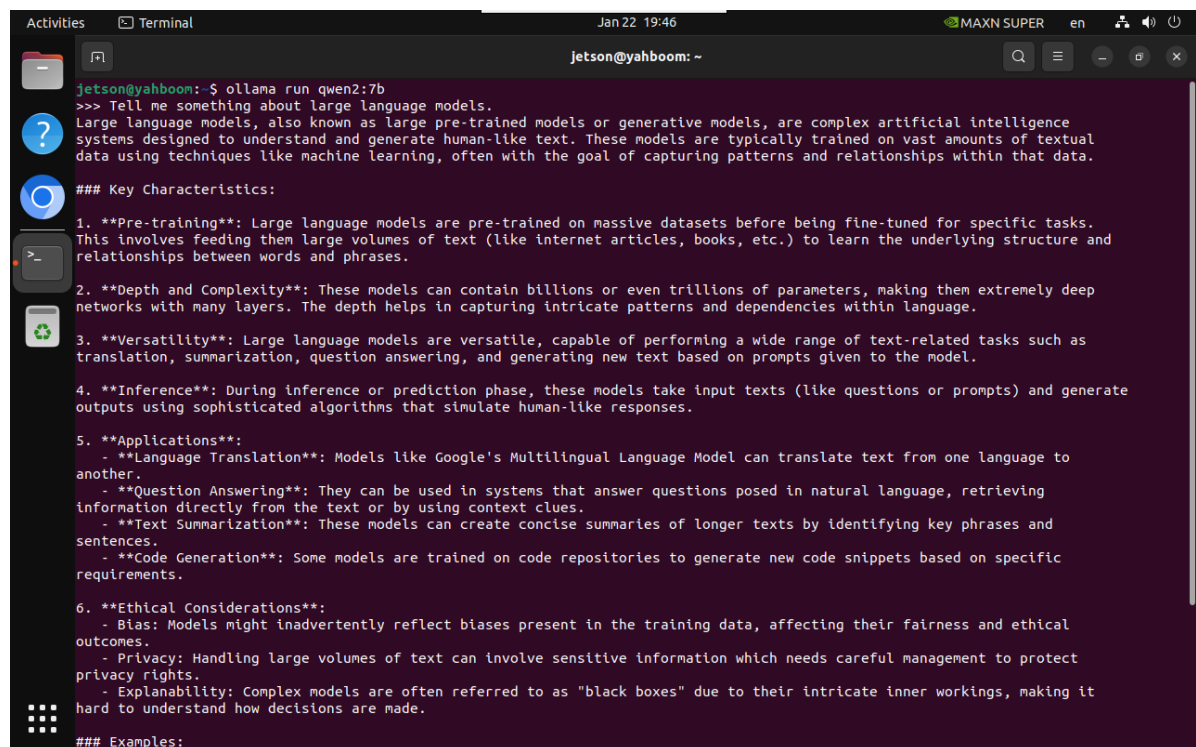
Small parameter version model: motherboards with 8G memory or less can run this

```
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!



```
Activities Terminal Jan 22 19:46 MAXN SUPER en jetson@yahboom: ~
jetson@yahboom:~$ ollama run qwen2:7b
>>> Tell me something about large language models.
Large language models, also known as large pre-trained models or generative models, are complex artificial intelligence systems designed to understand and generate human-like text. These models are typically trained on vast amounts of textual data using techniques like machine learning, often with the goal of capturing patterns and relationships within that data.

### Key Characteristics:

1. **Pre-training**: Large language models are pre-trained on massive datasets before being fine-tuned for specific tasks. This involves feeding them large volumes of text (like internet articles, books, etc.) to learn the underlying structure and relationships between words and phrases.

2. **Depth and Complexity**: These models can contain billions or even trillions of parameters, making them extremely deep networks with many layers. The depth helps in capturing intricate patterns and dependencies within language.

3. **Versatility**: Large language models are versatile, capable of performing a wide range of text-related tasks such as translation, summarization, question answering, and generating new text based on prompts given to the model.

4. **Inference**: During inference or prediction phase, these models take input texts (like questions or prompts) and generate outputs using sophisticated algorithms that simulate human-like responses.

5. **Applications**:
   - **Language Translation**: Models like Google's Multilingual Language Model can translate text from one language to another.
   - **Question Answering**: They can be used in systems that answer questions posed in natural language, retrieving information directly from the text or by using context clues.
   - **Text Summarization**: These models can create concise summaries of longer texts by identifying key phrases and sentences.
   - **Code Generation**: Some models are trained on code repositories to generate new code snippets based on specific requirements.

6. **Ethical Considerations**:
   - Bias: Models might inadvertently reflect biases present in the training data, affecting their fairness and ethical outcomes.
   - Privacy: Handling large volumes of text can involve sensitive information which needs careful management to protect privacy rights.
   - Explainability: Complex models are often referred to as "black boxes" due to their intricate inner workings, making it hard to understand how decisions are made.

### Examples:
```

## 4.3. End the conversation

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

```
Activities Terminal Jan 22 19:46 MAXN SUPER en
jetson@yahboom: ~
2. **Depth and Complexity**: These models can contain billions or even trillions of parameters, making them extremely deep networks with many layers. The depth helps in capturing intricate patterns and dependencies within language.
3. **Versatility**: Large language models are versatile, capable of performing a wide range of text-related tasks such as translation, summarization, question answering, and generating new text based on prompts given to the model.
4. **Inference**: During inference or prediction phase, these models take input texts (like questions or prompts) and generate outputs using sophisticated algorithms that simulate human-like responses.
5. **Applications**:
  - **Language Translation**: Models like Google's Multilingual Language Model can translate text from one language to another.
  - **Question Answering**: They can be used in systems that answer questions posed in natural language, retrieving information directly from the text or by using context clues.
  - **Text Summarization**: These models can create concise summaries of longer texts by identifying key phrases and sentences.
  - **Code Generation**: Some models are trained on code repositories to generate new code snippets based on specific requirements.
6. **Ethical Considerations**:
  - Bias: Models might inadvertently reflect biases present in the training data, affecting their fairness and ethical outcomes.
  - Privacy: Handling large volumes of text can involve sensitive information which needs careful management to protect privacy rights.
  - Explainability: Complex models are often referred to as "black boxes" due to their intricate inner workings, making it hard to understand how decisions are made.

### Examples:
  - **GPT (Generative Pre-trained Transformers)** series by Google, specifically GPT-2 and GPT-3, which have been widely discussed for their ability to generate coherent text across a variety of topics.
  - **T5 (Text-to-Text Transfer Transformer)** by Google, designed as an encoder-decoder model capable of performing various natural language tasks.

### Conclusion:
  Large language models represent cutting-edge advancements in AI technology, pushing the boundaries of what can be achieved with machine learning in terms of text understanding and generation. They are rapidly evolving fields that have significant implications for both practical applications like online services and research into human cognition and language processing.

>>> /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>