

Qwen2 model

Qwen2 model

[Model scale](#)

[Performance performance](#)

[Got Qwen2](#)

[Use Qwen2](#)

[Run Qwen2](#)

[Dialogue](#)

[End conversation](#)

[Reference material](#)

Demonstration environment

Development Board : Raspberry Pi 5B

SD(TF)card: 64G (Above 16G, the larger the capacity, the more models can be experienced)

Raspberry Pi 5B (16G RAM): Run 14B and below parameter models

Raspberry Pi 5B (8G RAM): Run 8B and below parameter models

Raspberry Pi 5B (4G RAM): Run 3B and below parameter models

Raspberry Pi 5B (2G RAM): Run 0.5B and below parameter models

Model scale

| Model | 参数 |
|-------|------|
| Qwen2 | 0.5B |
| Qwen2 | 1.5B |
| Qwen2 | 7B |
| Qwen2 | 72B |

Raspberry Pi 5B (8G RAM): Test Qwen2 model with parameters of 7B and below.

Raspberry Pi 5B (4G RAM): Test Qwen2 model with parameters of 1.5B and below.

Performance performance

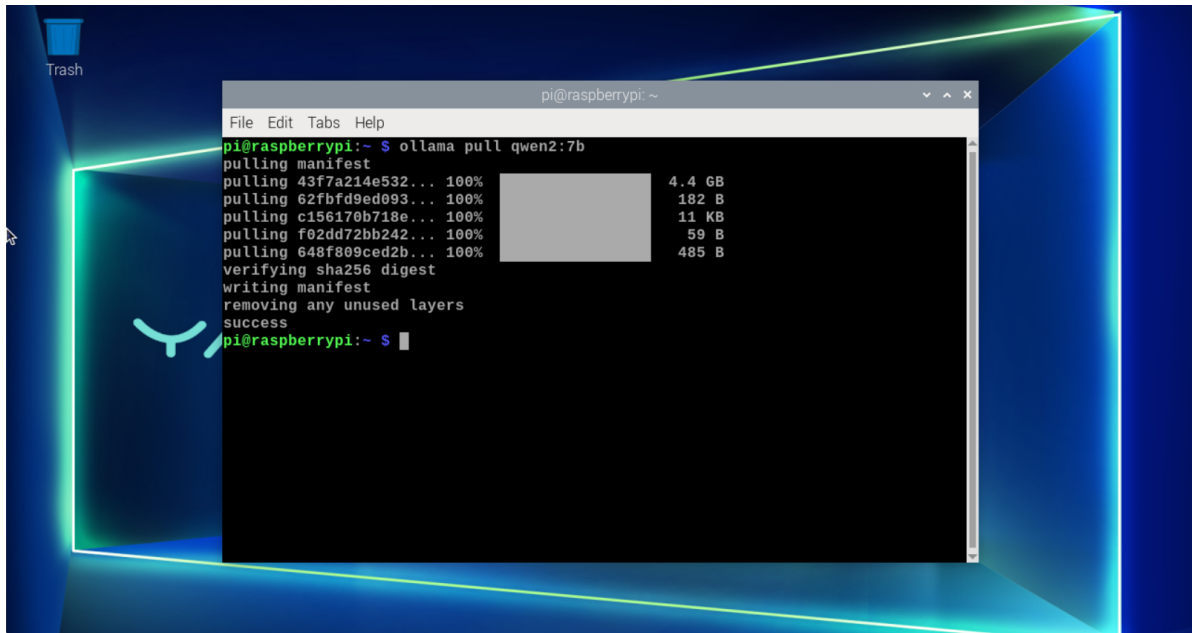
| | Qwen2-7B Instruct | Llama3-8B Instruct | GLM4-9B Chat |
|------------|----------------------|-----------------------|-----------------|
| AlignBench | 7.21 | 6.20 | 7.01 |
| MT-Bench | 8.41 | 8.05 | 8.35 |
| MMLU | 70.5 | 68.4 | 72.4 |
| GSM8K | 82.3 | 79.6 | 79.6 |
| MATH | 49.6 | 30.0 | 50.6 |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

Got Qwen2

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

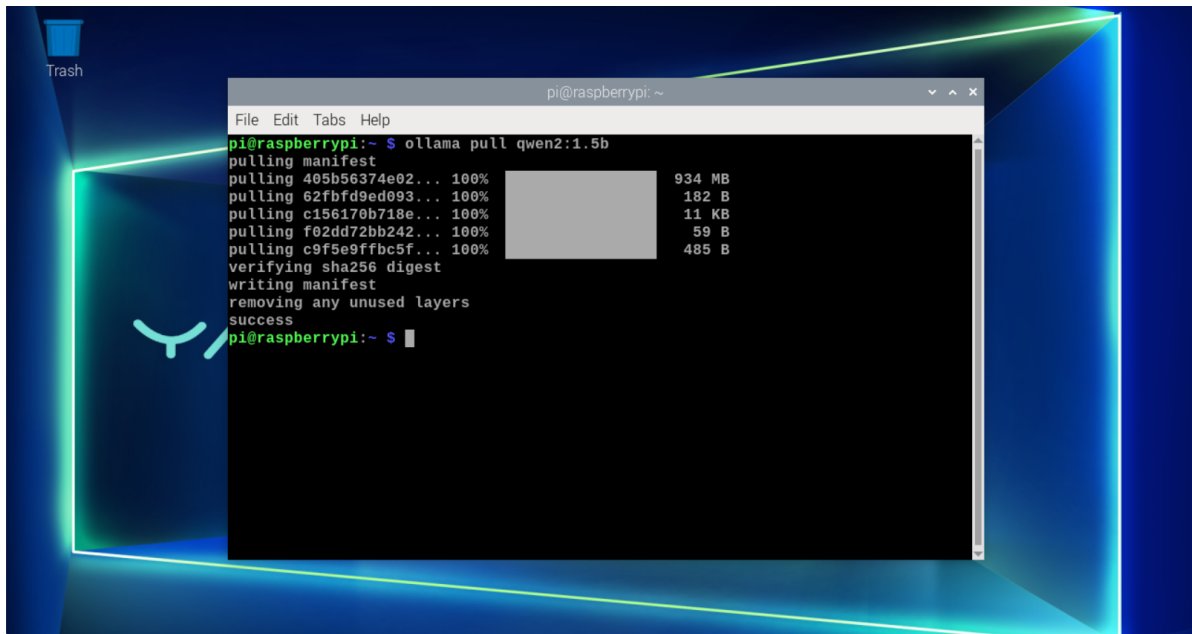
Raspberry Pi 5B (8G RAM)

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



Raspberry Pi 5B (4G RAM)

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



Use Qwen2

Run Qwen2

Raspberry Pi 5B (8G RAM)

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
```

Raspberry Pi 5B (4G RAM)

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
```

Dialogue

Raspberry Pi 5B (8G RAM)

```
Tell me something about large language models.
```

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

```
pi@raspberrypi: ~  
File Edit Tabs Help  
pi@raspberrypi:~$ ollama run qwen2:7b  
>>> Tell me something about large language models.  
Large language models are sophisticated artificial intelligence systems designed to generate human-like text by learning patterns and structures from vast amounts of textual data. These models have been trained on massive datasets, sometimes containing billions or even trillions of parameters, making them highly complex and capable of producing a wide range of outputs, including essays, stories, poetry, and code.  
  
The training process involves feeding the model large volumes of text, allowing it to learn patterns in language use, semantics, syntax, and context. This enables the model to generate responses that are coherent, relevant, and sometimes even creative.  
  
One key feature of large language models is their ability to handle a variety of input prompts or questions, which they then use to produce output that continues the text in a plausible manner. The more data the model has been trained on, the better it can generalize across different types of content and topics.  
  
These models have numerous applications, including but not limited to:  
  
1. Automated Writing: They are used for generating text, which can be useful for creating content for websites, blogs, or even marketing materials.  
2. Virtual Customer Service: Large language models can simulate human-like conversations with customers, providing personalized responses and potentially improving customer service experiences.  
3. Translation Services: By understanding the nuances of various languages, these models can translate text between languages accurately and efficiently.  
4. Educational Tools: They can be used to create adaptive learning platforms that tailor educational content based on individual student needs.  
5. Research and Development: In fields like medicine, science, and technology, large language models assist in summarizing research papers, generating hypotheses, or even helping with data analysis.  
  
However, they also come with challenges and limitations. Large language models can generate biased or inappropriate content if the training data contains such biases. Additionally, their outputs are not always factually accurate, especially when dealing with new topics that were not part of their training dataset. Furthermore, the computational resources required to train these models are significant and may pose ethical concerns regarding energy consumption.
```

Raspberry Pi 5B (4G RAM)

Tell me something about large language models.

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

```
pi@raspberrypi: ~  
File Edit Tabs Help  
pi@raspberrypi:~$ 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  
pi@raspberrypi:~$ ollama run qwen2:1.5b  
>>> Tell me something about large language models.  
Large language models, also known as generative language models, are artificial intelligence systems that can generate text in response to prompts or instructions. They are a type of machine learning model and have been developed for a wide range of tasks, including but not limited to translation, summarization, chatbot generation, and generating creative content.  
  
These models typically use deep neural networks and other advanced algorithms to learn from large amounts of text data. The size of the models is measured in terms of parameters or size of the model's hidden layers, with larger models having more parameters that enable them to generate text with greater fluency and diversity.  
  
Large language models have shown remarkable improvement in performance over the years, surpassing even human language abilities in certain areas, such as machine translation. However, they also raise concerns about bias, privacy, and the ethical implications of creating and using large language models for commercial or military purposes.  
  
>>> Send a message (/? for help)
```

End conversation

You can end the conversation by using the shortcut key 'Ctrl+d' or '/bye'.

Raspberry Pi 5B (8G RAM)

```
File Edit Tabs Help
pi@raspberrypi: ~
pi@raspberrypi:~ $ ollama run qwen2:7b
>>>
Use Ctrl + d or /bye to exit.
>>>
pi@raspberrypi:~ $

The training process involves feeding the model with a large amount of text data, including semantics, syntax, and context. This enables the model to generate text that is coherent and even creative.

One key feature of large language models is their ability to generate text that continues the input. The better it can generalize across different contexts, the more useful it is.

These models have numerous applications, including:

1. **Automated Writing**: They are used for generating blog posts, reports, or even marketing materials.
2. **Virtual Customer Service**: Large language models can provide personalized responses and potentially improve customer satisfaction.
3. **Translation Services**: By understanding multiple languages accurately and efficiently.
4. **Educational Tools**: They can be used to create personalized learning materials or assist in individual student needs.
5. **Research and Development**: In fields like drug discovery, they can help in summarizing research papers, generating hypotheses, or analyzing experimental data.

However, they also come with challenges and limitations. For example, they may generate content if the training data contains such biases or misinformation, especially when dealing with new topics that were not part of their training dataset. Furthermore, the computational resources required to train these models are significant and may pose ethical concerns regarding energy consumption.

Overall, large language models represent a powerful tool in the realm of natural language processing (NLP), offering both opportunities for innovation and areas needing careful consideration and management.

>>> /bye
pi@raspberrypi:~ $
```

Raspberry Pi 5B (4G RAM)

```
File Edit Tabs Help
pi@raspberrypi: ~
pi@raspberrypi:~ $ ollama run qwen2:1.5b
>>> Tell me something about large language models.
Large language models, often abbreviated as LLMs or GPT (Generative Pre-trained Transformer), are artificial intelligence systems that have the ability to generate human-like text based on a set of training data and algorithms. These models have been trained on vast amounts of text data, including books, movies, news articles, social media posts, and scientific papers, among others.

Here is some information about large language models:

* **Training**: LLMs are typically trained using deep learning techniques, which involve the use of neural networks to analyze large volumes of text data. This allows them to learn from patterns in the text and generate more sophisticated responses.
* **Processing**: To process natural language, LLMs use complex algorithms that involve understanding the context, grammar, and semantics of the input. This often involves rule-based processing, as well as understanding the meaning behind the words.
* **Applications**: Large language models have a wide range of applications, including text generation, document summarization, keyword extraction, and more.

Some examples of LLMs include GPT-1 and GPT-2, which are based on the Transformer-based pre-trained model (BERT), and GPT-3, which is a more advanced version.

>>> /bye
pi@raspberrypi:~ $
```

Reference material

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>