

Model usage instructions

Model usage instructions

[System image Description](#)

[Parameter Description](#)

[Model Statistics](#)

[Reference material](#)

This tutorial mainly introduces some Raspberry Pi 5 system environments and models that can run on different RAM versions.

System image Description

The offline AI large model development tutorial for Raspberry Pi 5 can use the image already configured by Yahboom: Pi5-AI-Pure.img

System Information

user name: pi

password: yahboom

WiFi hotspot name: Pi_Hot

WiFi hotspot password: 12345678

System environment

- Ollama
 - Llama3
 - Qwen2
 - Phi-3
 - Gemma
 - WizardLM2
 - Code Llama
 - DeepSeek Coder
 - LLaVA
- Docker
- Open WebUI

Due to the limitations of the SD (TF) card, we did not install all the models in the tutorial onto the Pi5AI-Pure.img file. User can delete and download models by yourself.

Parameter Description

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

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

The above conclusion may not be entirely accurate, but it can be used as a reference!

ollam official website operation model description:
Run the 7B parameter model requires at least 8GB of available running memory.
Run the 13B parameter model with at least 16GB of available running memory.
Run the 33B parameter model with at least 32GB of available running memory.

Model Statistics

The following table shows common models based on the Ollama official website model library statistics.

Users can search for the desired models on the Ollama official website by themselves.

In theory, the larger the model parameters, the greater the memory required, and running the model can choose versions with smaller parameters to experience the functionality

| Model | Parameter | Raspberry Pi 5B (4G RAM) | Raspberry Pi 5B (8G RAM) | Running command |
|-----------------|-----------|-----------------------------|-----------------------------|------------------------------|
| Llama 3 | 7B | Not support | Support | ollama run llama3:8b |
| Llama 2 Chinese | 7B | Not support | Support | ollama run llama2-chinese:7b |
| Qwen2 | 0.5B | Support | Support | ollama run qwen2:0.5b |
| Qwen2 | 1.5B | Support | Support | ollama run qwen2:1.5b |

| Model | Parameter | Raspberry Pi 5B (4G RAM) | Raspberry Pi 5B (8G RAM) | Running command |
|-------------------|-----------|-----------------------------|-----------------------------|---------------------------------------|
| Qwen2 | 7B | Not support | Support | ollama run qwen2:7b |
| Phi-3 | 3.8B | Not support | Support | ollama run phi3:3.8b |
| Phi-2 | 2.7B | Support | Support | ollama run phi:2.7b |
| Gemma | 2B | Support | Support | ollama run gemma:2b |
| Gemma | 7B | Not support | Support | ollama run gemma:7b |
| LLaVA | 7B | Not support | Support | ollama run llava:7b |
| WizardLM-2 | 7B | Not support | Support | ollama run wizardlm2:7b |
| DeepSeek Coder | 1.3B | Support | Support | ollama run deepseek- coder:1.3b |
| DeepSeek Coder | 6.7B | Not support | Support | ollama run deepseek- coder:6.7b |
| Yi | 6B | Not support | Support | ollama run yi:6b |

Reference material

Ollama

Website: <https://ollama.com/>

GitHub: <https://github.com/ollama/ollama>