Ollama

> **Demo Environment**

Development Boards: RDK X5, Jetson Orin, Raspberry Pi 5

Ollama is an open-source tool designed to simplify the deployment and operation of large language models, allowing users to use high-quality language models in their local environment.

1. Large Language Models (LLMs)

Large Language Models (LLMs) are a class of advanced text generation systems based on artificial intelligence. Their main characteristic is their ability to learn and understand human language through large-scale training data and generate natural and fluent text.

2. Ollama Installation

Note: It is recommended to use a pre-configured environment image from Yabo Smart. Personal environment configuration takes a long time, and installation may encounter issues depending on the system version.

This tutorial demonstrates installing Ollama using a script on an RDK X5 motherboard. (**The large language model image provided by Yabo Intelligence is already configured; no further installation is required.**)

Script Installation

```
sudo apt install curl -y
sudo curl -fsSL https://ollama.com/install.sh | sh
```

The entire installation process takes a relatively long time; please be patient!

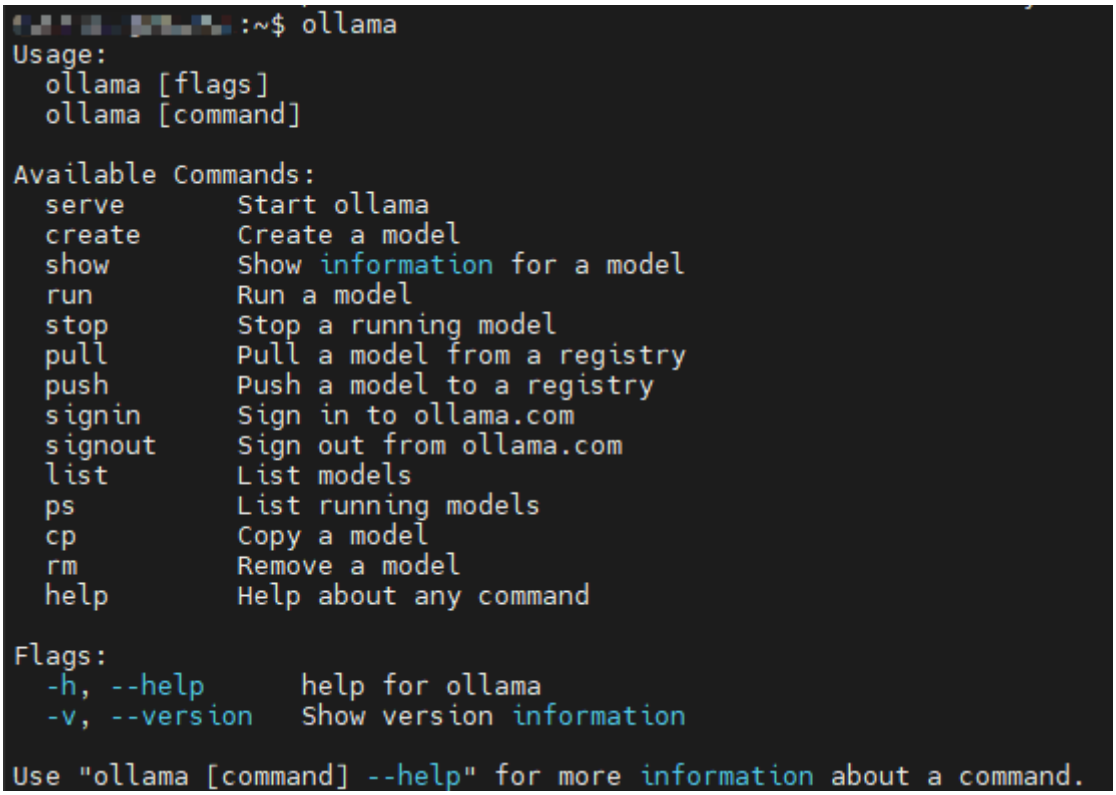
Note: Ollam download failures are normal. Try multiple times, or use a proxy to download. You'll need to search online for instructions on using a proxy.

Once the installation is complete, it should display:

```
:~$ sudo curl -fsSL https://ollama.com/install.sh | sh
>>> Cleaning up old version at /usr/local/lib/ollama
>>> Installing ollama to /usr/local
>>> Downloading Linux arm64 bundle
############################################################################ 100.0%
>>> Creating ollama user...
>>> Adding ollama user to render group...
>>> Adding ollama user to video group...
>>> Adding current user to ollama group...
>>> Creating ollama systemd service...
>>> Enabling and starting ollama service...
Created symlink /etc/systemd/system/default.target.wants/ollama.service → /etc/systemd/system/ollama.service.
pcilib: Cannot open /proc/bus/pci
lspci: Cannot find any working access method.
pcilib: Cannot open /proc/bus/pci
lspci: Cannot find any working access method.
>>> The Ollama API is now available at 127.0.0.1:11434.
>>> Install complete. Run "ollama" from the command line.
WARNING: No NVIDIA/AMD GPU detected. Ollama will run in CPU-only mode.
```

3. Using Ollam

Type `ollama` in the terminal to see the following prompts:



Command	Function
ollama serve	Start Ollam
ollama create	Create a model from a model file
ollama show	Display model information
ollama stop	Stop a model
ollama pull	Pull a model from the registry
ollama push	Push a model to the registry
ollama list	List models
ollama ps	List running models
ollama cp	Copy a model
ollama rm	Delete a model
ollama help	Get help information for any command

4. Uninstalling Ollama

- Delete the service

```
sudo systemctl stop ollama
sudo systemctl disable ollama
sudo rm /etc/systemd/system/ollama.service
```

- Delete a file

```
sudo rm $(which ollama)
```

- Delete a model and service user and group

```
sudo rm -r /usr/share/ollama
sudo userdel ollama
sudo groupdel ollama
```

References

> **Ollama**

Official Website: https://ollama.com/

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