Code Llama

> **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	×	×

Code Llama is an open source large language model (LLM) designed by Meta AI specifically for understanding and generating code.

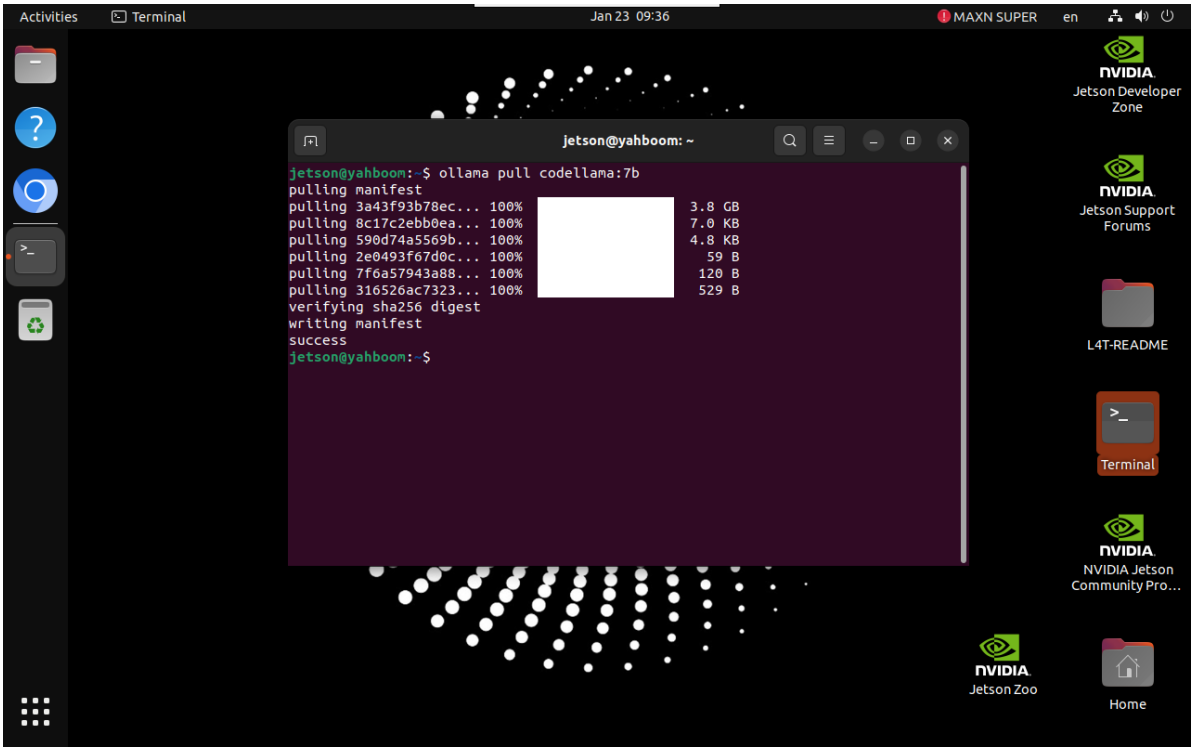
1. Model scale

Model	Parameters
Code Llama	7B
Code Llama	13B
Code Llama	34B
Code Llama	70B

2. Pull Code Llama

Using the pull command will automatically pull the model from the Ollama model library:

```
ollama pull codellama:7b
```



3. Use Code Llama

3.1. Run Code Llama

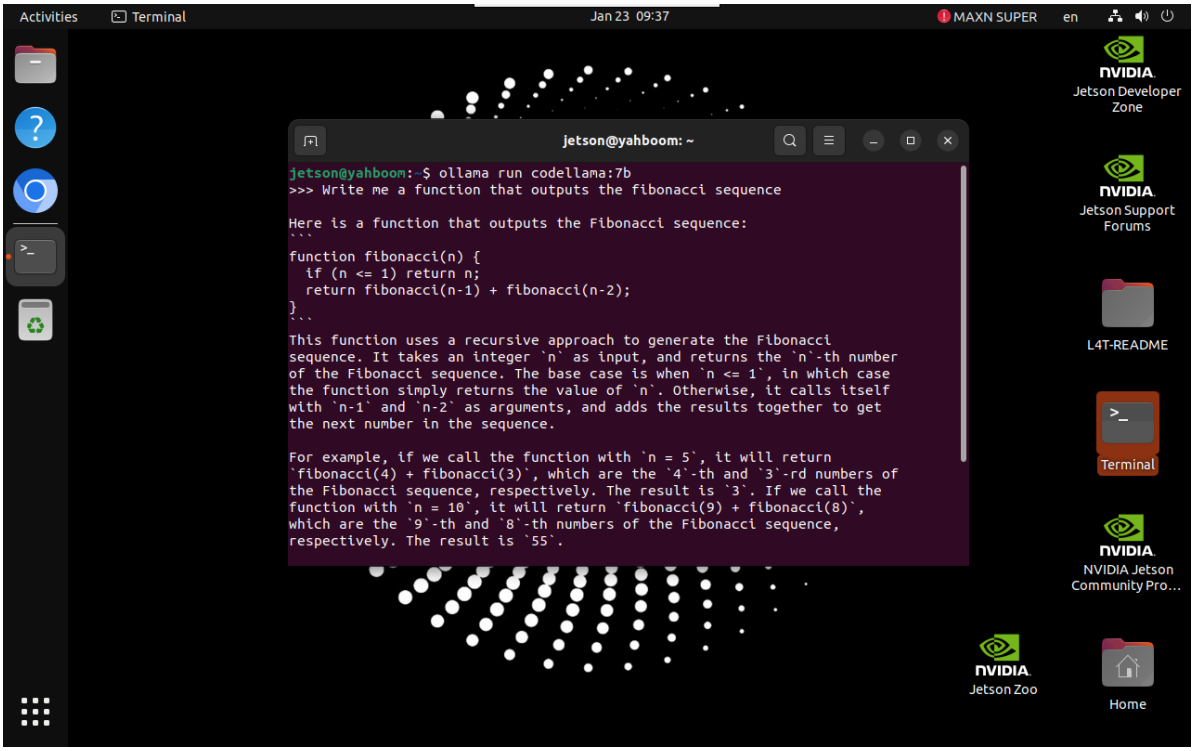
If the system does not have a running model, the system will automatically pull the Code Llama 7B model and run it:

```
ollama run codellama:7b
```

3.2. Have a conversation

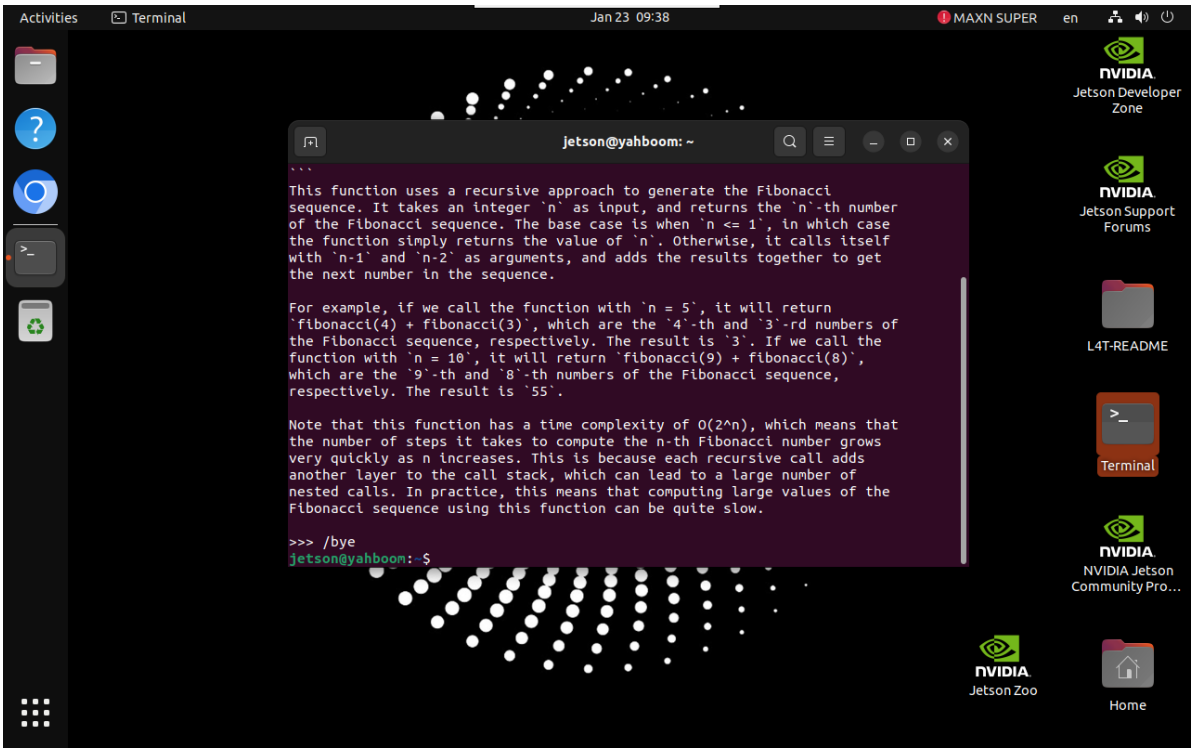
```
Write me a function that outputs the fibonacci sequence
```

The time to reply to the question is related to the hardware configuration, so be patient!



3.3. End the conversation

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



References

Ollama

Official website: https://ollama.com/

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

Code Llama

Ollama corresponding model: https://ollama.com/library/codellama

GitHub: https://github.com/meta-llama/codellama