TinyLlama

> **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	√	√

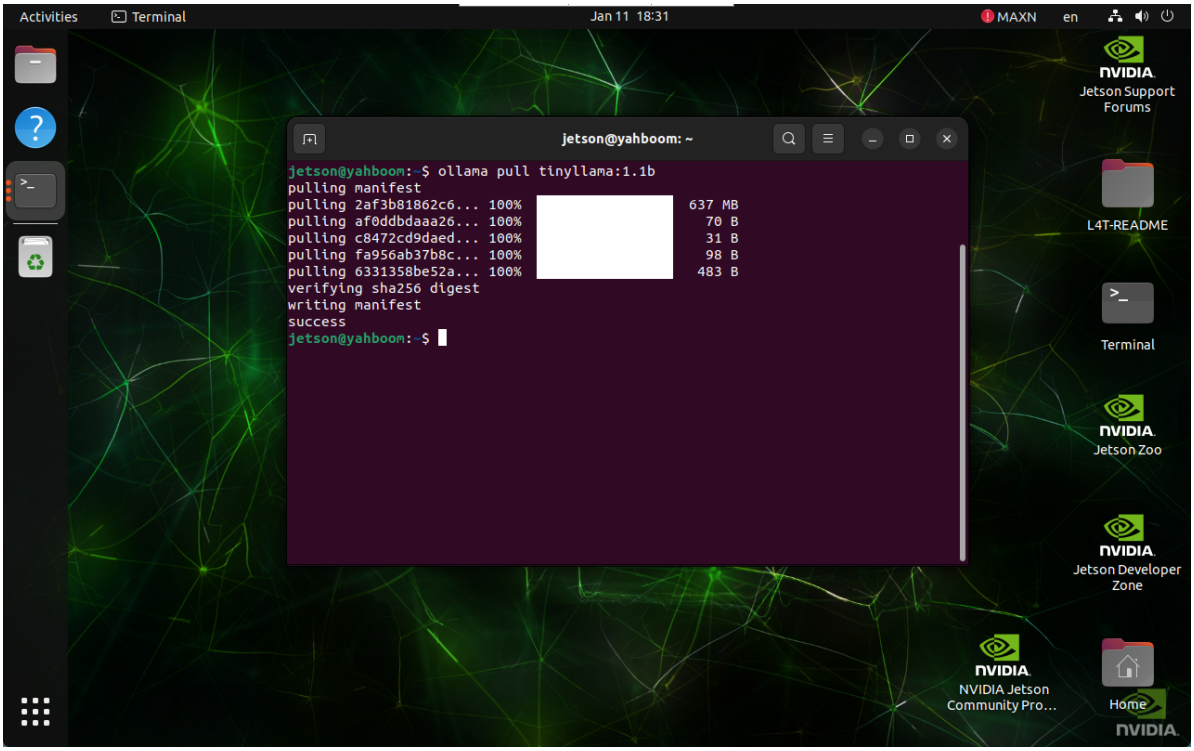
1. Model scale

Model	Parameters
TinyLlama	1.1B

2. Pull TinyLlama

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

```
ollama pull tinyllama:1.1b
```



3. Use TinyLlama

3.1. Run TinyLlama

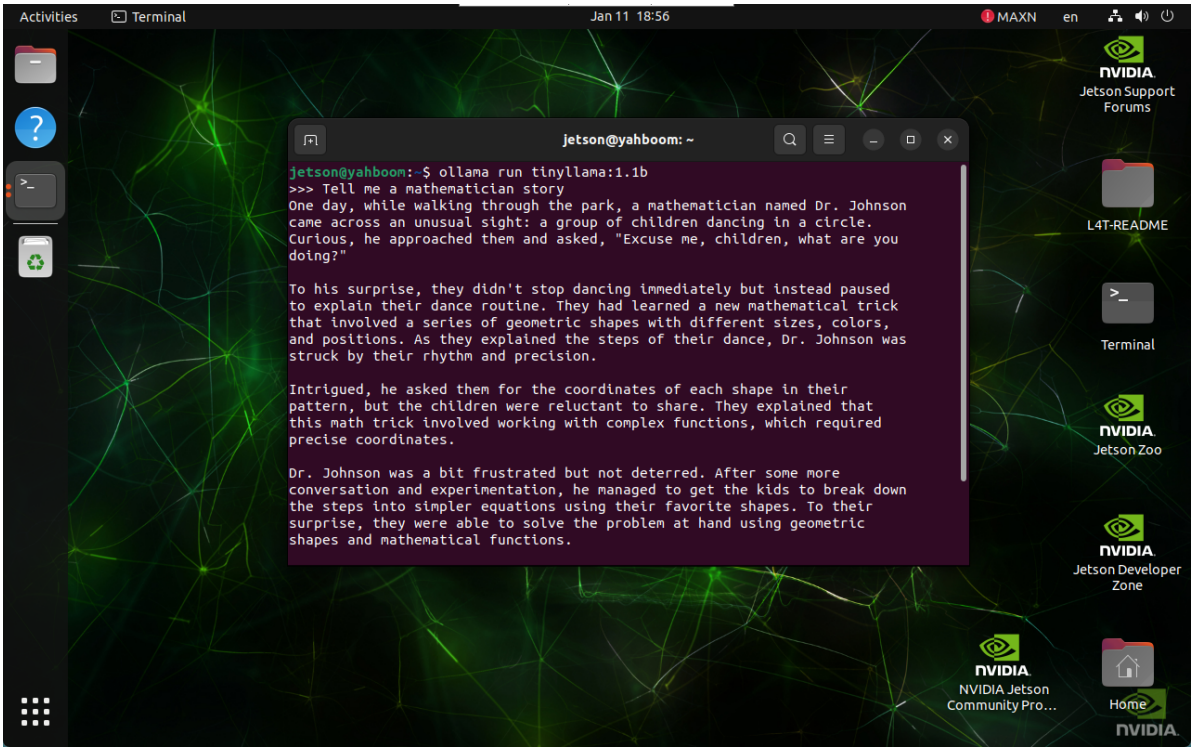
If the system does not have a running model, the system will automatically pull the TinyLlama 1.1B model and run it:

```
ollama run tinyllama:1.1b
```

3.2. Have a conversation

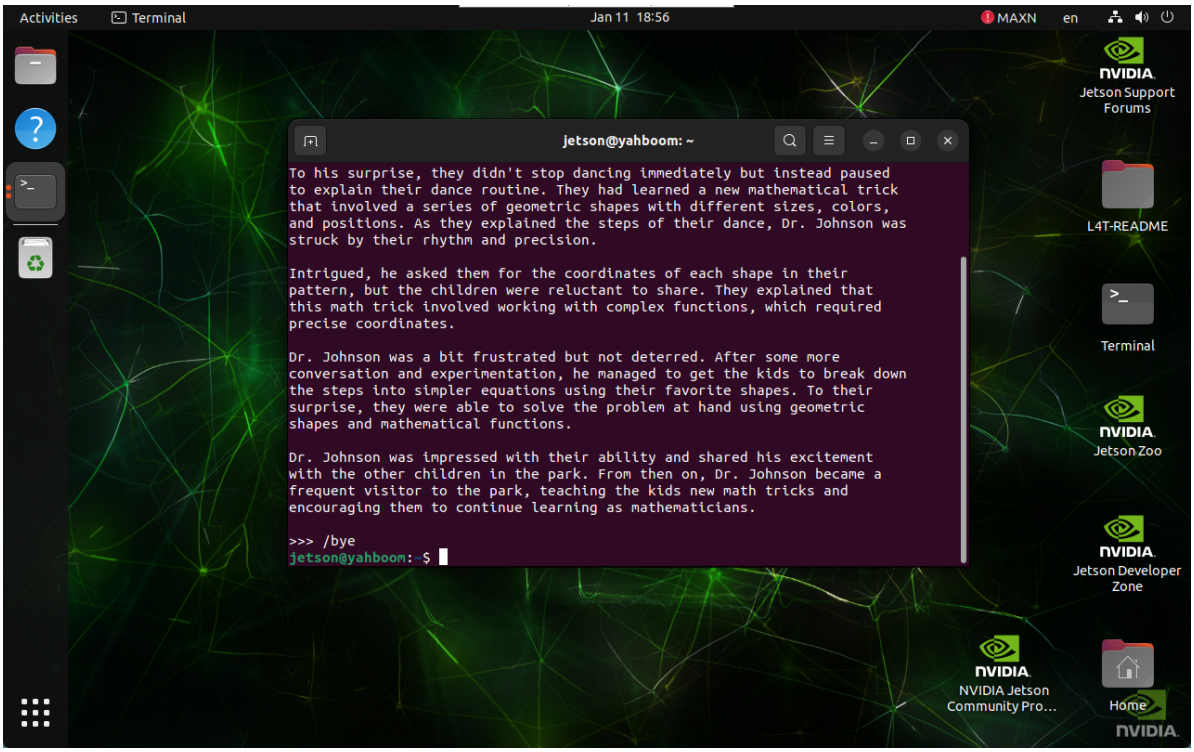
```
Tell me a mathematician story
```

The time to reply to the question is related to the hardware configuration, please 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

> **TinyLlama**

GitHub: https://github.com/jzhang38/TinyLlama

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