

Gemma

- Gemma
 - Model size
- Pull Gemma
- Use Gemma
 - Run Gemma
 - Start a conversation
 - End the conversation
- References

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	Ollama	Open WebUI
Jetson Orin NX 16GB	√	√
Jetson Orin NX 8GB	√ (Need to run the small parameter version)	√ (Need to run the small parameter version)
Jetson Orin Nano 8GB	√ (Need to run the small parameter version)	√ (Need to run the small parameter version)
Jetson Orin Nano 4GB	√ (Need to run the small parameter version)	√ (Need to run the small parameter version)

Gemma is a new open model developed by Google and its DeepMind team.

Model size

Model	Parameters
Gemma	2B
Gemma	7B

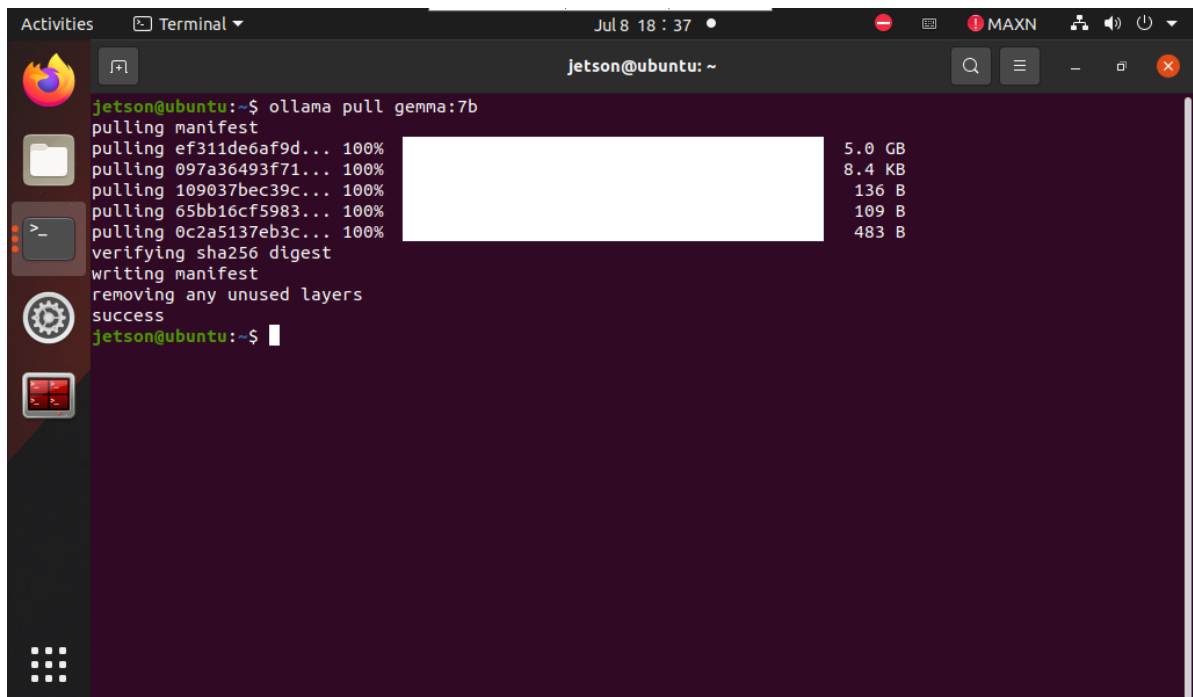
Pull Gemma

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

```
ollama pull gemma:7b
```

Model with small parameters: motherboards with 8G or less memory can run this

```
ollama pull gemma:2b
```



```
jetson@ubuntu:~$ ollama pull gemma:7b
pulling manifest
pulling ef311de6af9d... 100% 5.0 GB
pulling 097a36493f71... 100% 8.4 KB
pulling 109037bec39c... 100% 136 B
pulling 65bb16cf5983... 100% 109 B
pulling 0c2a5137eb3c... 100% 483 B
verifying sha256 digest
writing manifest
removing any unused layers
success
jetson@ubuntu:~$
```

Use Gemma

Run Gemma

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

```
ollama run gemma:7b
```

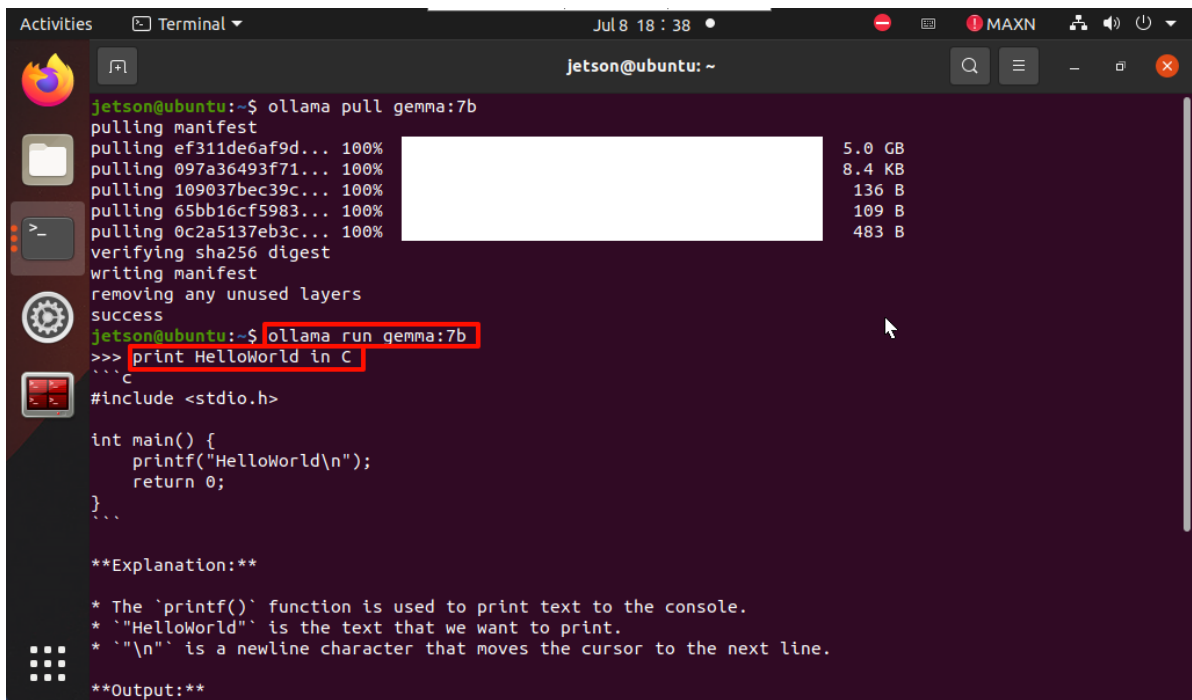
Model with small parameters: motherboards with 8G or less memory can run this

```
ollama run gemma:2b
```

Start a conversation

```
print HelloWorld in C
```

The time to reply to the question depends on the hardware configuration, please be patient!



```
jetson@ubuntu:~$ ollama pull gemma:7b
pulling manifest
pulling ef311de6af9d... 100%
pulling 097a36493f71... 100%
pulling 109037bec39c... 100%
pulling 65bb16cf5983... 100%
pulling 0c2a5137eb3c... 100%
verifying sha256 digest
writing manifest
removing any unused layers
success
jetson@ubuntu:~$ ollama run gemma:7b
>>> print HelloWorld in C
...
#include <stdio.h>

int main() {
    printf("HelloWorld\n");
    return 0;
}

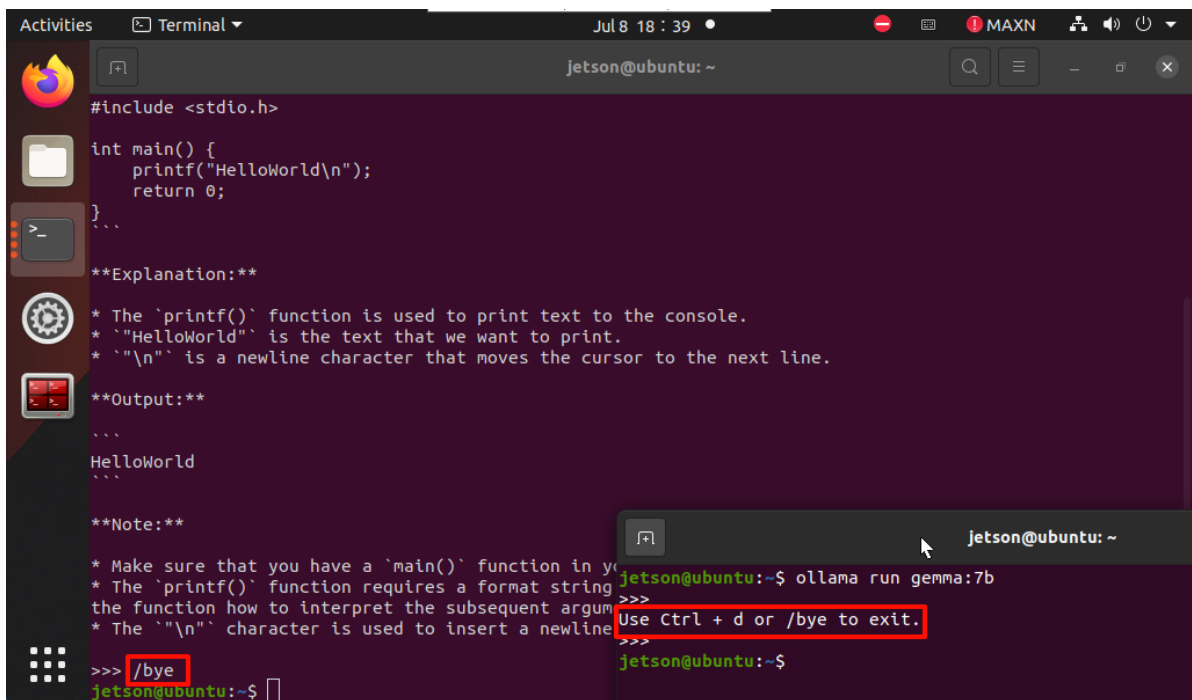
**Explanation:**

* The `printf()` function is used to print text to the console.
* `HelloWorld` is the text that we want to print.
* `"\n"` is a newline character that moves the cursor to the next line.

**Output:**
...
HelloWorld
...
```

End the conversation

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



```
#include <stdio.h>

int main() {
    printf("HelloWorld\n");
    return 0;
}

**Explanation:**

* The `printf()` function is used to print text to the console.
* `HelloWorld` is the text that we want to print.
* `"\n"` is a newline character that moves the cursor to the next line.

**Output:**
...
HelloWorld
...

**Note:**

* Make sure that you have a `main()` function in your code.
* The `printf()` function requires a format string to tell the function how to interpret the subsequent arguments.
* The `"\n"` character is used to insert a newline character.

>>> /bye
jetson@ubuntu:~$
```

References

Ollama

Official website: <https://ollama.com/>

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

Gemma

GitHub: <https://github.com/google-deepmind/gemma>

Ollama corresponding model: <https://ollama.com/library/gemma>

