

Gemma

[Gemma](#)

[Model scale](#)

[Got Gemma](#)

[Use Gemma](#)

[Have a conversation](#)

[End the conversation](#)

[References](#)

Demo environment

Development boards: Jetson nano

SD (TF) card: 64G

It is recommended to run the 4B and below parameter models

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

Model scale

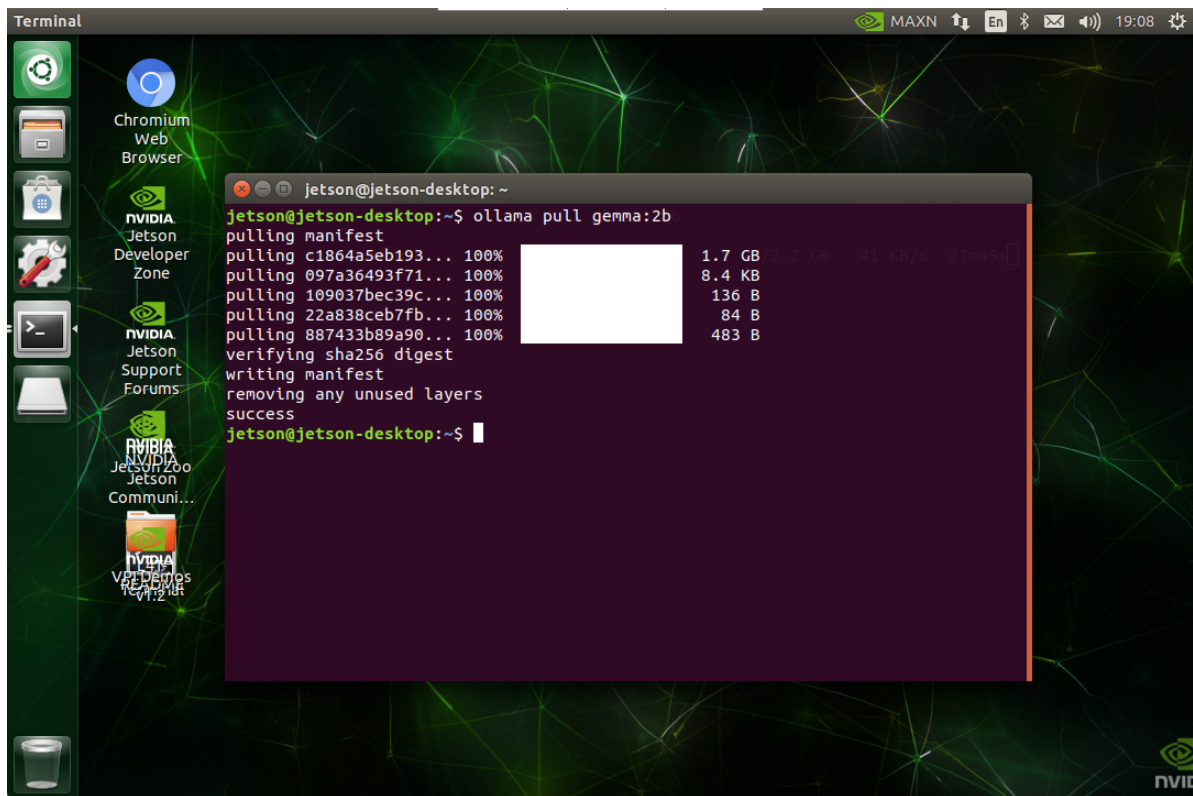
Model	Parameter
Gemma	2B
Gemma	7B

Gemma: Gemma model test using 7B parameters.

Got Gemma

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

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



Use Gemma

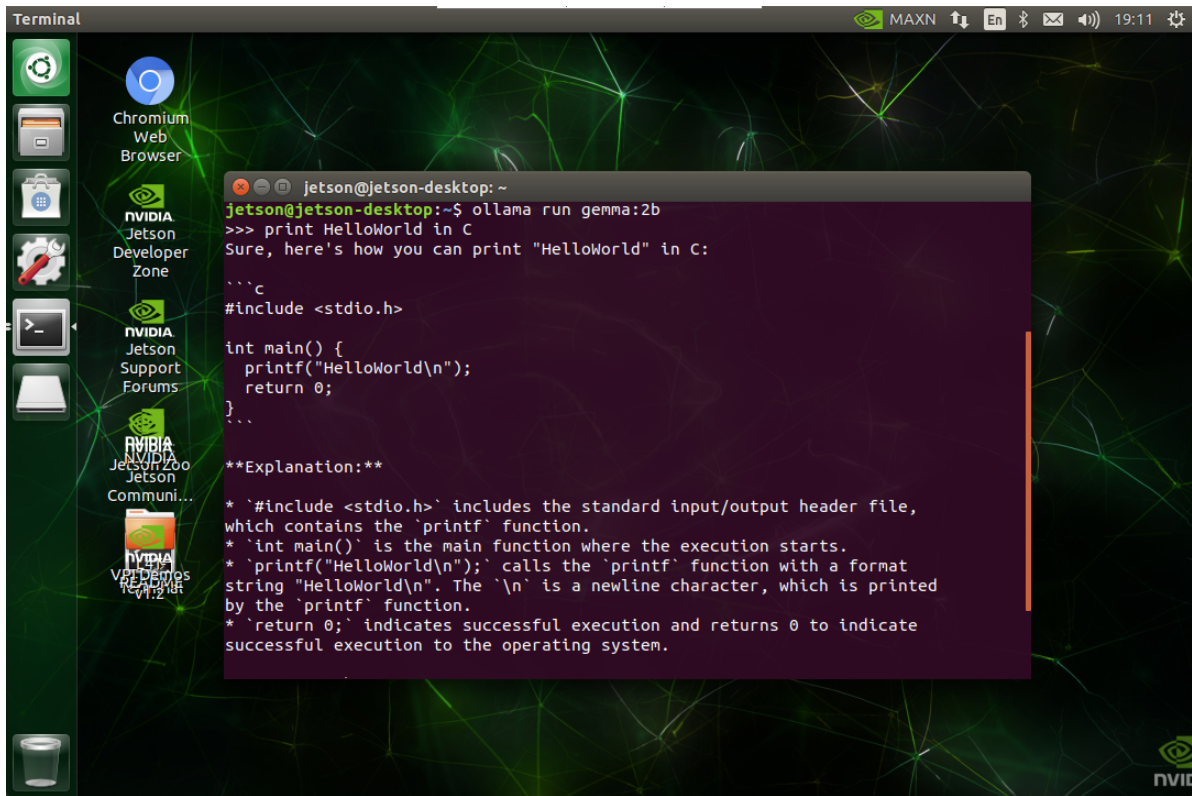
If the system does not have a running model, the system will automatically pull the Gemma 2B model and run it.

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

Have a conversation

```
print HelloWorld in C
```

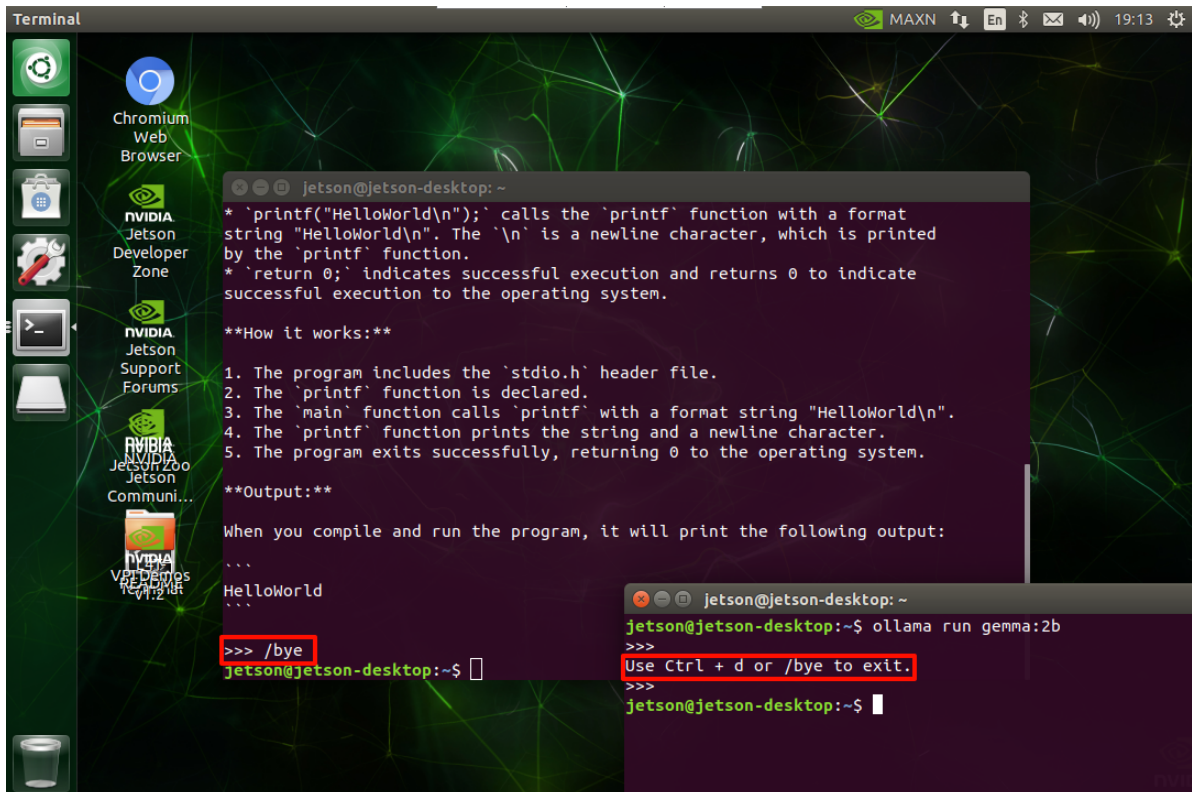
The response time is related to the hardware configuration, please be patient.



```
jetson@jetson-desktop: ~  
jetson@jetson-desktop:~$ ollama run gemma:2b  
>>> print HelloWorld in C  
Sure, here's how you can print "HelloWorld" in C:  
...  
#include <stdio.h>  
  
int main() {  
    printf("HelloWorld\\n");  
    return 0;  
}  
...  
  
**Explanation:**  
  
* '#include <stdio.h>' includes the standard input/output header file,  
which contains the 'printf' function.  
* 'int main()' is the main function where the execution starts.  
* 'printf("HelloWorld\\n");' calls the 'printf' function with a format  
string "HelloWorld\\n". The '\\n' is a newline character, which is printed  
by the 'printf' function.  
* 'return 0;' indicates successful execution and returns 0 to indicate  
successful execution to the operating system.
```

End the conversation

Use the `ctrl+d` shortcut or `/bye` to end the conversation.



```
jetson@jetson-desktop: ~  
jetson@jetson-desktop:~$ ollama run gemma:2b  
>>> print HelloWorld in C  
Sure, here's how you can print "HelloWorld" in C:  
...  
#include <stdio.h>  
  
int main() {  
    printf("HelloWorld\\n");  
    return 0;  
}  
...  
  
**Explanation:**  
  
* '#include <stdio.h>' includes the standard input/output header file,  
which contains the 'printf' function.  
* 'int main()' is the main function where the execution starts.  
* 'printf("HelloWorld\\n");' calls the 'printf' function with a format  
string "HelloWorld\\n". The '\\n' is a newline character, which is printed  
by the 'printf' function.  
* 'return 0;' indicates successful execution and returns 0 to indicate  
successful execution to the operating system.  
  
**How it works:**  
  
1. The program includes the 'stdio.h' header file.  
2. The 'printf' function is declared.  
3. The 'main' function calls 'printf' with a format string "HelloWorld\\n".  
4. The 'printf' function prints the string and a newline character.  
5. The program exits successfully, returning 0 to the operating system.  
  
**Output:**  
  
When you compile and run the program, it will print the following output:  
...  
HelloWorld  
...  
>>> /bye  
jetson@jetson-desktop:~$
```

References

Ollama

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>