DeepComputing

DC-ROMA RISC-V Mainboard II AI Model User Guide (Ubuntu AI Image)

Document Properties	
Document version	V1.0
Latest release date	Oct 28, 2025
Applicable product model	DC-ROMA RISC-V Mainboard II for Framework Laptop 13
Operating system version	Ubuntu 24.04
Document target audience	Users holding the DC-ROMA RISC-V Mainboard II series products
Document overview	This document applies to users of the factory-installed image system or those who download the image from: http://120.92.155.32:8082/artifactory/virtOS/fml13v03-eswin/15019-ubuntu-24.04-desktop-grub-sdcard-AI.zip

Catalog

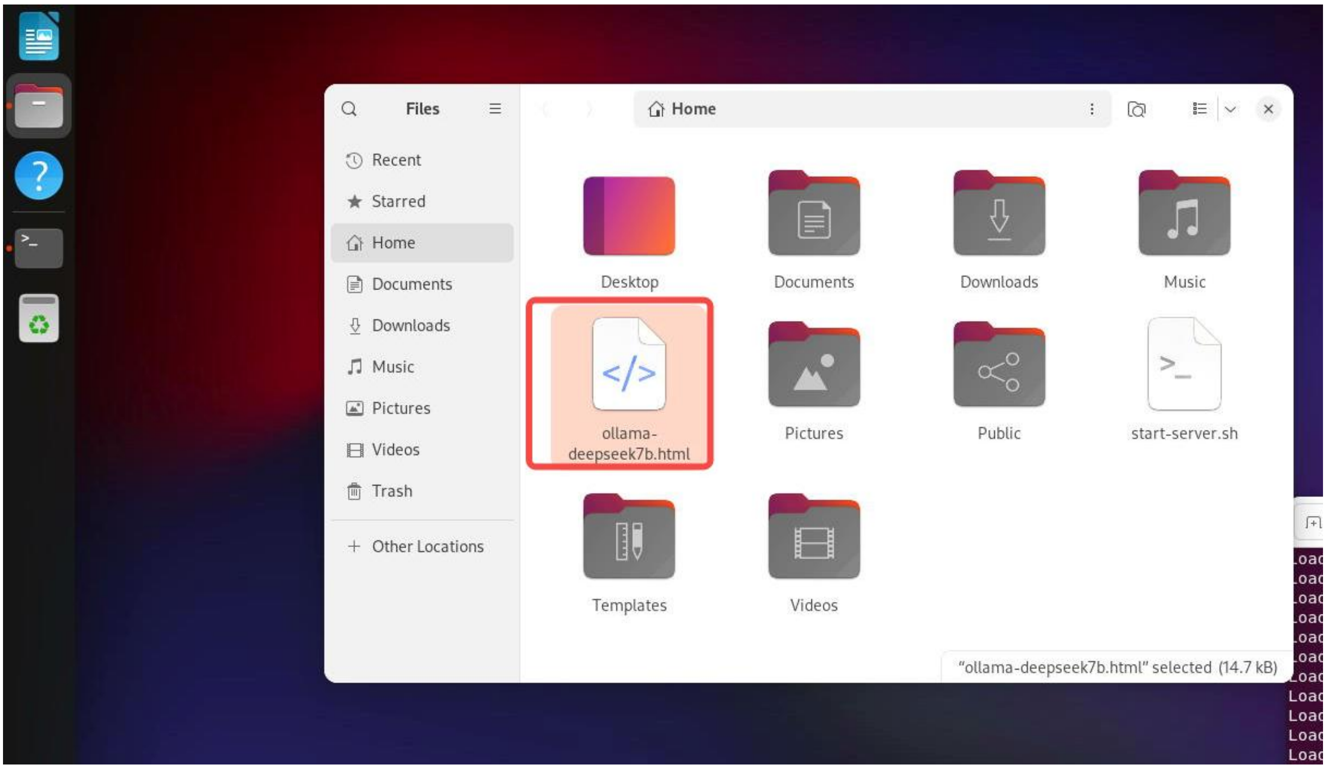
Ollama Containerized Deepseek-R1-7B Usage Guide

The AI PC comes pre-installed with a containerized local DeepSeek 7B model via Ollama and includes a startup script. Simply run the script to use the model. For an enhanced experience, a visual frontend is included.

1、Open Terminal and run:

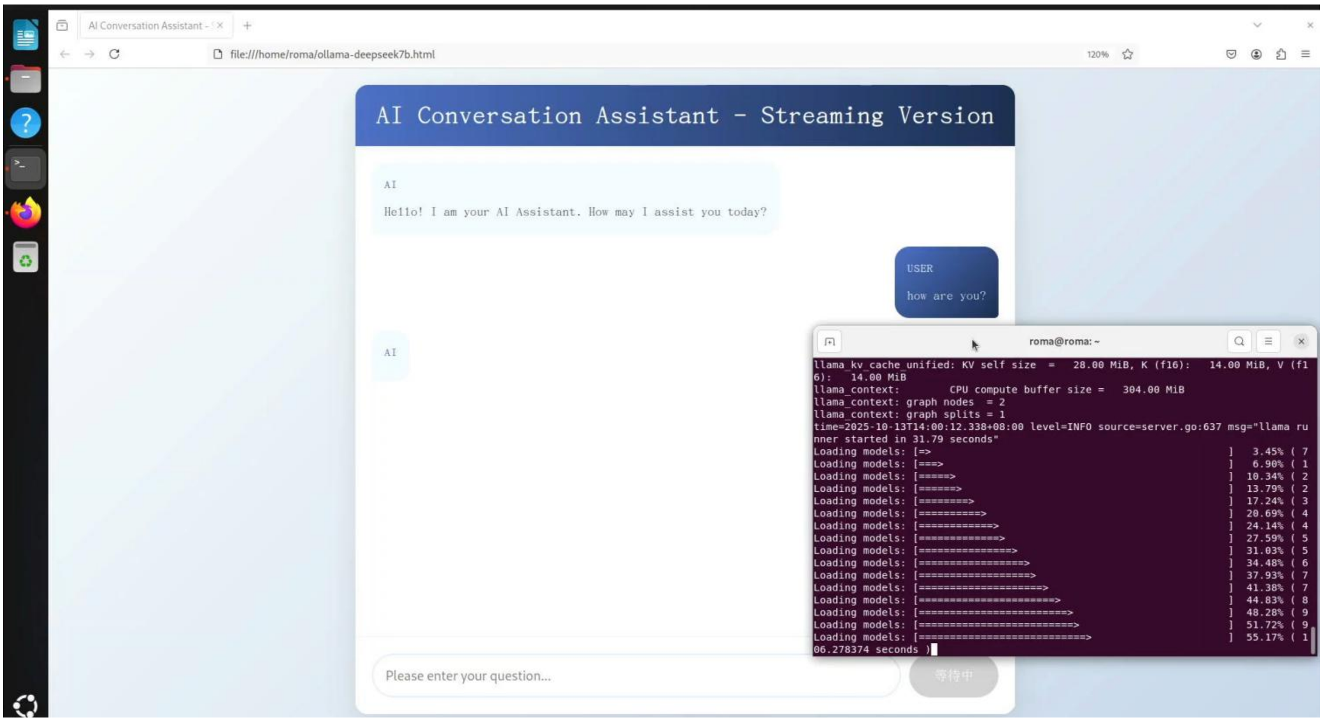
```
sudo  ./start-server.sh
```

2、Double-click <u>ollama-deepseek7b.html</u> to open the visual interface.



"ollama-deepseek7b.html" selected (14.7 kB)

3、Ask questions in the dialog box. Note: Initial query loads the model (~3 minutes).



AI Conversation Assistant -
file:///home/roma/ollama-deepseek7b.html
120%

AI Conversation Assistant - Streaming Version

AI

Hello! I am your AI Assistant. How may I assist you today?

USER

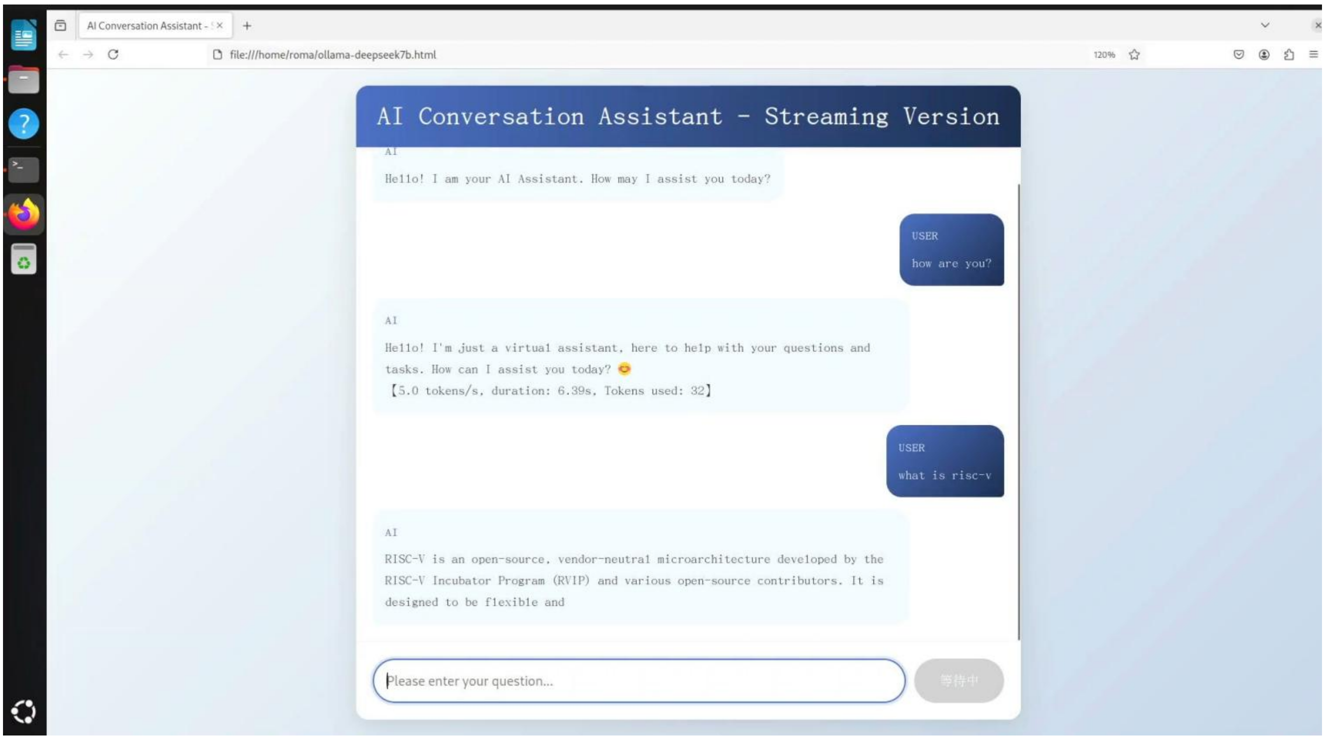
how are you?

AI

```
llama_kv_cache_unified: KV self size  =   28.00 MiB, K (f16):   14.00 MiB, V (f1
6):   14.00 MiB
llama_context:        CPU compute buffer size =   304.00 MiB
llama_context: graph nodes  = 2
llama_context: graph splits = 1
time=2025-10-13T14:00:12.338+08:00 level=INFO source=server.go:637 msg="llama ru
nner started in 31.79 seconds"
Loading models: [=>                                                    ]   3.45% ( 7
Loading models: [===>                                                  ]   6.90% ( 1
Loading models: [=====>                                                ]  10.34% ( 2
Loading models: [======>                                               ]  13.79% ( 2
Loading models: [========>                                             ]  17.24% ( 3
Loading models: [==========>                                           ]  20.69% ( 4
Loading models: [============>                                         ]  24.14% ( 4
Loading models: [=============>                                        ]  27.59% ( 5
Loading models: [===============>                                      ]  31.03% ( 5
Loading models: [=================>                                    ]  34.48% ( 6
Loading models: [==================>                                   ]  37.93% ( 7
Loading models: [====================>                                 ]  41.38% ( 7
Loading models: [======================>                               ]  44.83% ( 8
Loading models: [========================>                             ]  48.28% ( 9
Loading models: [=========================>                            ]  51.72% ( 9
Loading models: [===========================>                          ]  55.17% ( 1
06.278374 seconds )
```

Please enter your question...

4、 Sample Execution



AI Conversation Assistant -
file:///home/roma/ollama-deepseek7b.html
120%

AI Conversation Assistant - Streaming Version

AI

Hello! I am your AI Assistant. How may I assist you today?

USER

how are you?

AI

Hello! I'm just a virtual assistant, here to help with your questions and tasks. How can I assist you today? 😊
【5.0 tokens/s, duration: 6.39s, Tokens used: 32】

USER

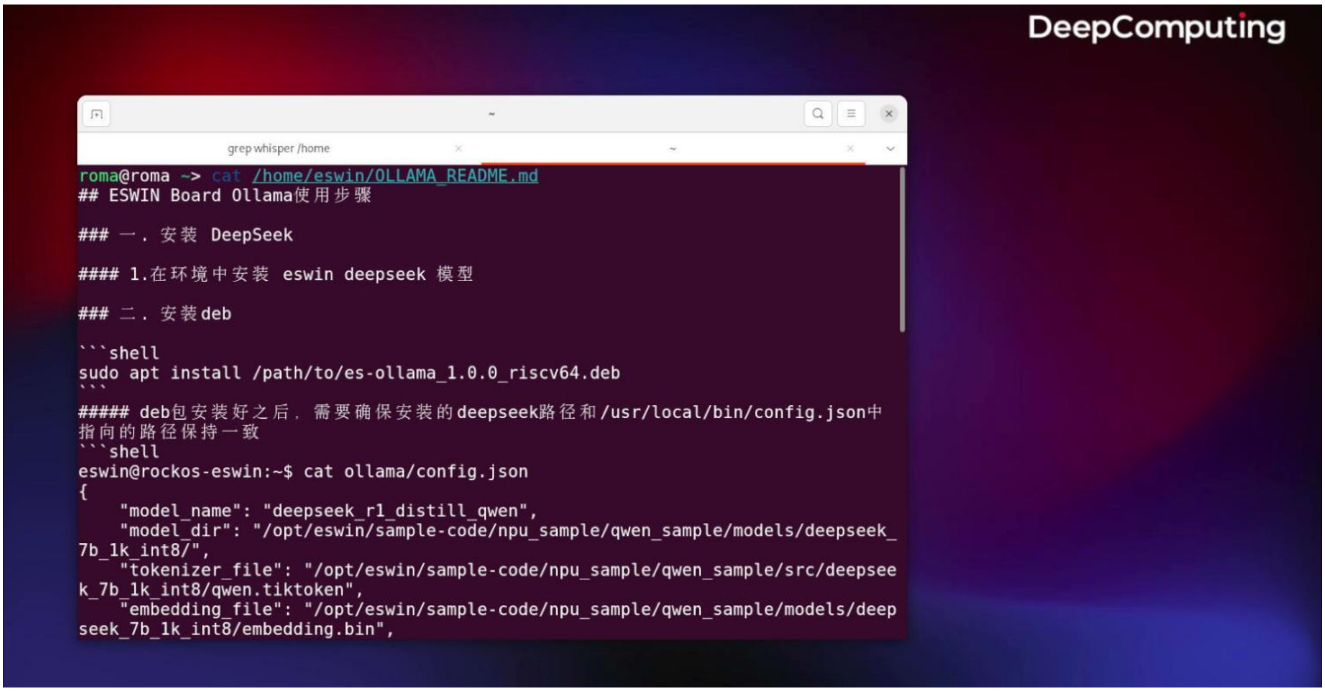
what is risc-v

AI

RISC-V is an open-source, vendor-neutral microarchitecture developed by the RISC-V Incubator Program (RVIP) and various open-source contributors. It is designed to be flexible and

Please enter your question...

5、You can view the usage steps in the /home/eswin/OLLAMA_README.mddirectory.



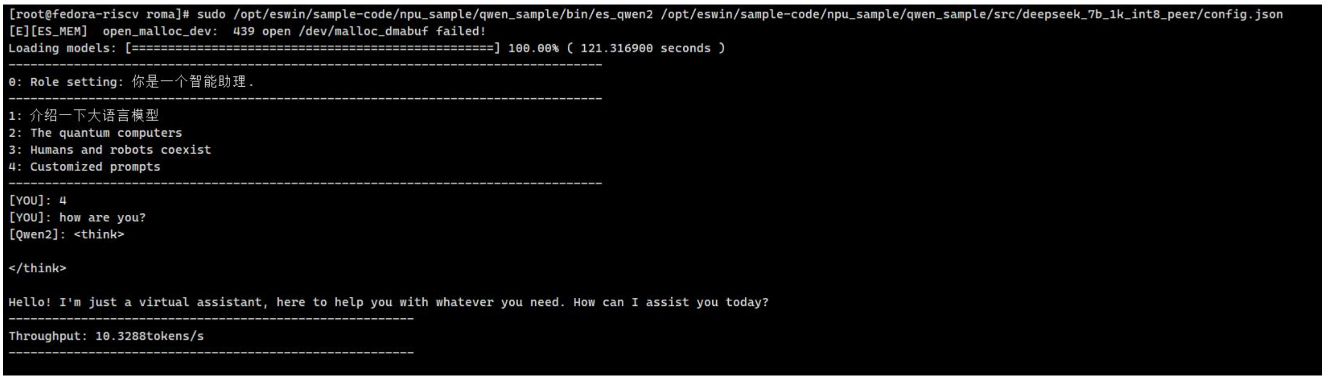
~~~
roma@roma ~> cat /home/eswin/OLLAMA_README.md
## ESWIN Board Ollama使用步骤

### 一. 安装 DeepSeek

#### 1.在环境中安装 eswin deepseek 模型

### 二. 安装deb

```shell
sudo apt install /path/to/es-ollama_1.0.0_riscv64.deb
```
##### deb包安装好之后，需要确保安装的deepseek路径和/usr/local/bin/config.json中
指向的路径保持一致
```shell
eswin@rockos-eswin:~$ cat ollama/config.json
{
    "model_name": "deepseek_r1_distill_qwen",
    "model_dir": "/opt/eswin/sample-code/npu_sample/qwen_sample/models/deepseek_
7b_1k_int8/",
    "tokenizer_file": "/opt/eswin/sample-code/npu_sample/qwen_sample/src/deepsee
k_7b_1k_int8/qwen.tiktoken",
    "embedding_file": "/opt/eswin/sample-code/npu_sample/qwen_sample/models/deep
seek_7b_1k_int8/embedding.bin",
~~~

Terminal Execution: Deepseek-7B with Dual-Gen NPU Acceleration

1、 Launch Model via Terminal:

```
sudo /opt/eswin/sample-code/npu_sample/qwen_sample/bin/es_qwen2
/opt/eswin/sample-
code/npu_sample/qwen_sample/src/deepseek_7b_1k_int8_peer/config.json
```

2、 Select the interaction pattern according to the prompt

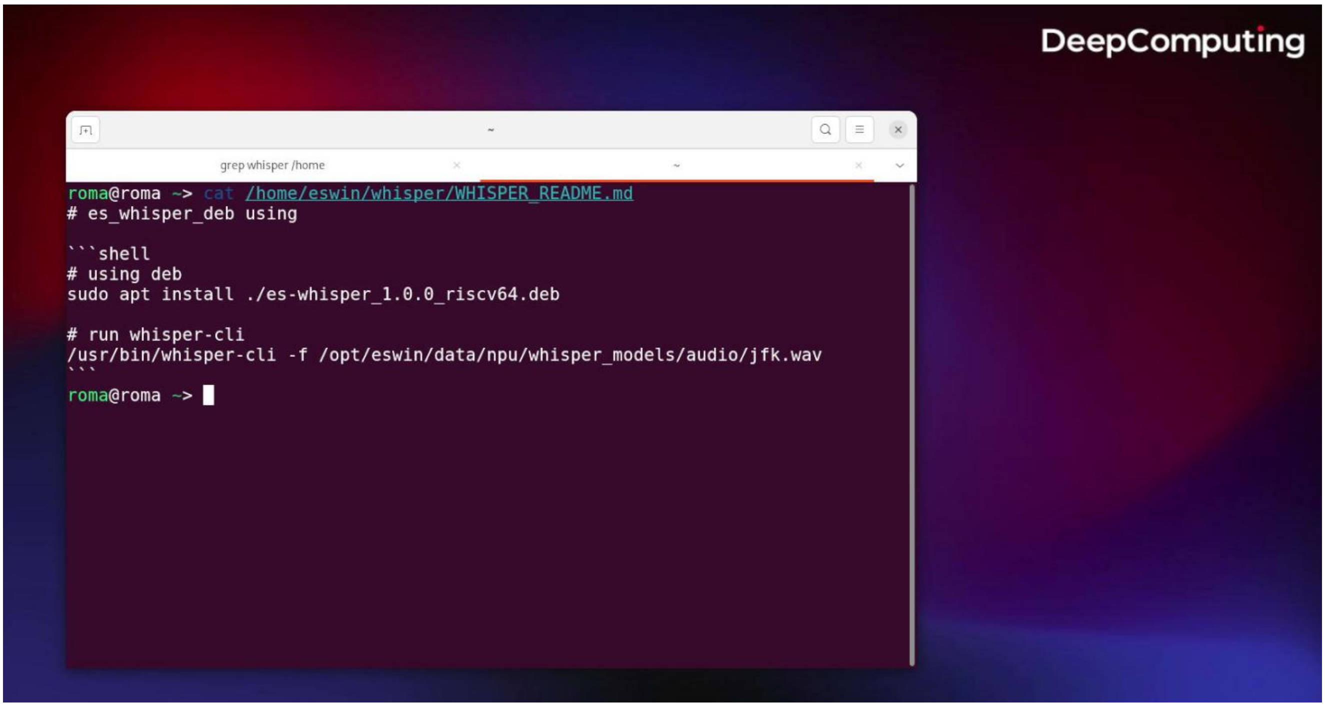


```
[root@fedora-riscv roma]# sudo /opt/eswin/sample-code/npu_sample/qwen_sample/bin/es_qwen2 /opt/eswin/sample-code/npu_sample/qwen_sample/src/deepseek_7b_1k_int8_peer/config.json
[E][ES_MEM]  open_malloc_dev:  439 open /dev/malloc_dmabuf failed!
Loading models: [==================================================] 100.00% ( 121.316900 seconds )
--------------------------------------------------------------------------------
0: Role setting: 你是一个智能助理.
--------------------------------------------------------------------------------
1: 介绍一下大语言模型
2: The quantum computers
3: Humans and robots coexist
4: Customized prompts
--------------------------------------------------------------------------------
[YOU]: 4
[YOU]: how are you?
[Qwen2]: <think>

</think>

Hello! I'm just a virtual assistant, here to help you with whatever you need. How can I assist you today?
--------------------------------------------------------
Throughput: 10.3288tokens/s
--------------------------------------------------------
```

Terminal Execution: Whisper Model

1、Open Terminal and run the command to view the model usage guide:

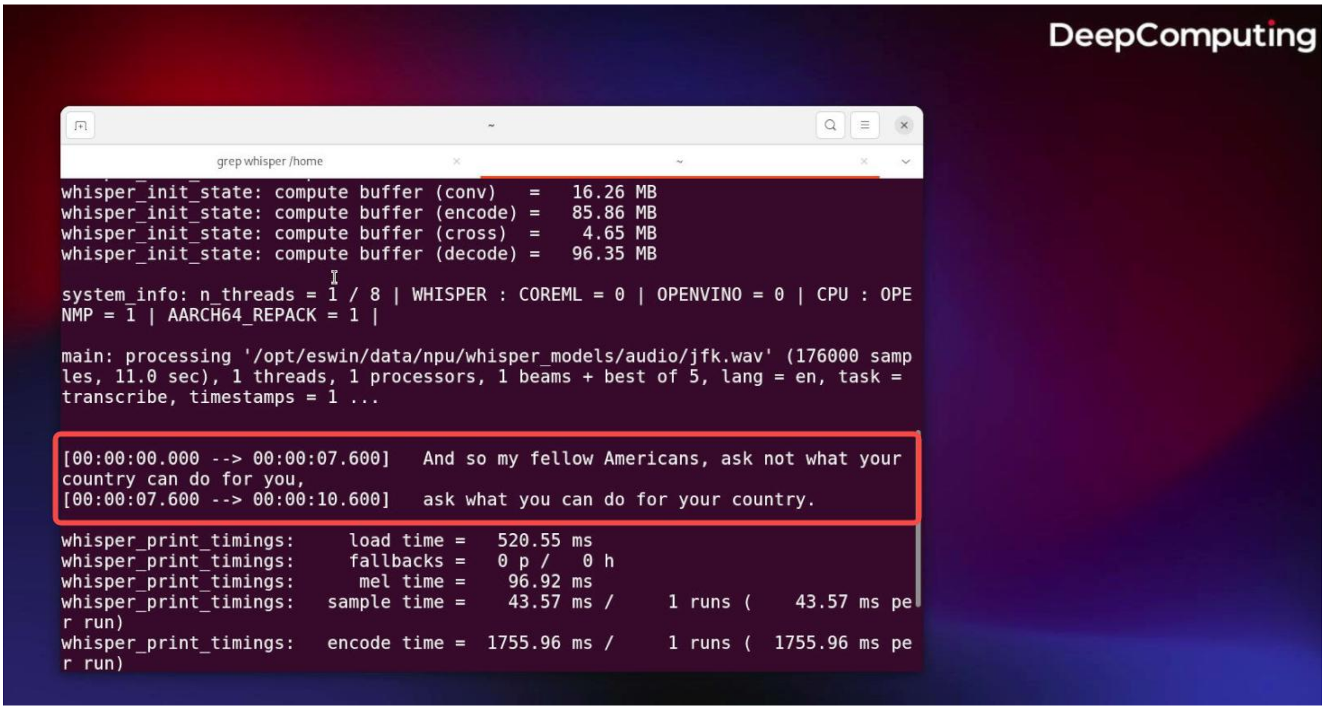
~~~
cat  /home/eswin/whisper/WHISPER_README.md


# es_whisper_deb using
```shell
# using deb
sudo apt install ./es-whisper_1.0.0_riscv64.deb
# run whisper-cli
/usr/bin/whisper-cli -f /opt/eswin/data/npu/whisper_models/audio/jfk.wav
```
~~~



~~~
roma@roma ~> cat /home/eswin/whisper/WHISPER_README.md
# es_whisper_deb using

```shell
# using deb
sudo apt install ./es-whisper_1.0.0_riscv64.deb

# run whisper-cli
/usr/bin/whisper-cli -f /opt/eswin/data/npu/whisper_models/audio/jfk.wav
```
roma@roma ~>
~~~

2、Enter the command to launch the model, which will run a demo converting speech to text.

```
sudo /usr/bin/whisper-cli -f
/opt/eswin/data/npu/whisper_models/audio/jfk.wav
```



```
whisper_init_state: compute buffer (conv)   =   16.26 MB
whisper_init_state: compute buffer (encode) =   85.86 MB
whisper_init_state: compute buffer (cross)  =    4.65 MB
whisper_init_state: compute buffer (decode) =   96.35 MB

system_info: n_threads = 1 / 8 | WHISPER : COREML = 0 | OPENVINO = 0 | CPU : OPE
NMP = 1 | AARCH64_REPACK = 1 |

main: processing '/opt/eswin/data/npu/whisper_models/audio/jfk.wav' (176000 samp
les, 11.0 sec), 1 threads, 1 processors, 1 beams + best of 5, lang = en, task =
transcribe, timestamps = 1 ...


[00:00:00.000 --> 00:00:07.600]   And so my fellow Americans, ask not what your
country can do for you,
[00:00:07.600 --> 00:00:10.600]   ask what you can do for your country.

whisper_print_timings:     load time =   520.55 ms
whisper_print_timings:     fallbacks =   0 p /   0 h
whisper_print_timings:      mel time =    96.92 ms
whisper_print_timings:   sample time =    43.57 ms /     1 runs (    43.57 ms pe
r run)
whisper_print_timings:   encode time =  1755.96 ms /     1 runs (  1755.96 ms pe
r run)
```