```
# Set the path of the cross-compiler GCC_COMPILER_PATH=~/gcc-arm-10.2-2020.11-x86_64-aarch64-none-linux-gnu/bin/aarch64-none-linux-gnu
```

Subsequently, users can use build-linux.sh to initiate the compilation process. Upon completion of the compilation, users will obtain the corresponding llm_demo program, which will installed in the install/demo_linux_aarch64/llm_demo directory directory. Following this, the executable file, library folder, and the RKLLM model (previously converted and quantized using the RKLLM-Toolkit tool) should be pushed to the board-side.

```
adb push install/demo_Linux_aarch64  /data
adb push /PC/path/to/your/rkllm/model /data/demo_Linux_aarch64
```

After completing the above steps, users can enter the terminal interface of the board using adb, and navigate to the corresponding /data/demo_linux_aarch64 directory. Then, the inference of RKLLM model on the board can be invoked using the following command:

```
adb shell
cd /data/demo_linux_aarch64
export LD_LIBRARY_PATH=./lib
./llm_demo /path/to/your/rkllm/model 1024 2048
```

With the above operations, users can enter the example inference interface, interact with the board-side model for inference, and obtain real-time inference results from the RKLLM model.

3.3.3 Monitor inference performance and Log Viewing

To monitor the inference performance of RKLLM on the board like the above figure, you can use the command:

```
export RKLLM_LOG_LEVEL=1
```

```
I rkllm: ------------------------------------------------------------------------------------
I rkllm:  Stage         Total Time (ms)  Tokens    Time per Token (ms)      Tokens per Second
I rkllm: ------------------------------------------------------------------------------------
I rkllm:  Init          1430.74          /         /                        /
I rkllm:  Prefill       98.41            10        9.84                     101.61
I rkllm:  Generate      641.28           9         71.25                    14.03
I rkllm: ------------------------------------------------------------------------------------
I rkllm:  Memory Usage (GB)
I rkllm:  2.04
I rkllm: ------------------------------------------------------------------------------------
```

Figure 3-1 RKLLM inference performance logs on the hardware platform

This will display the number of tokens processed and the inference time for both the Prefill and Generate stages after each inference, as shown in the figure below. This information will help you evaluate the performance by providing detailed logging of how long each stage of the inference process takes.If you need to view more detailed logs, such as the tokens after encoding the prompt, you can use the following command: