specify file paths when the user needs to load a Lora model or use the prompt feature.

The following is a simple example of how to use the one-click deployment script `build_rkllm_server_flask.sh`:

```
./build_rkllm_server_flask.sh
    --workshop /path/to/workshop
    --model_path /path/to/model.rkllm
    --platform rk3588
```

After executing the above command, the one-click deployment script will perform the following steps:

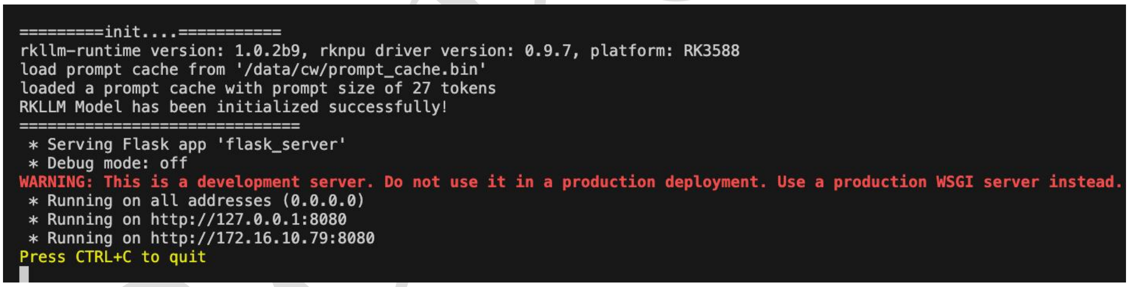
1) Check the Linux environment on the board.

2) Automatically install the Flask library if not already installed.

3) Push the necessary files under rkllm_server_demo/rkllm_server to the board.

4) Index the RKLLM model in the preset working directory for RKLLM-Server-Flask.

Once you see the message "RKLLM Model has been initialized successfully!" in the terminal, it indicates that the RKLLM-Server-Flask example has been successfully launched.



```
=========init....===========
rkllm-runtime version: 1.0.2b9, rknpu driver version: 0.9.7, platform: RK3588
load prompt cache from '/data/cw/prompt_cache.bin'
loaded a prompt cache with prompt size of 27 tokens
RKLLM Model has been initialized successfully!
=============================
 * Serving Flask app 'flask_server'
 * Debug mode: off
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
 * Running on all addresses (0.0.0.0)
 * Running on http://127.0.0.1:8080
 * Running on http://172.16.10.79:8080
Press CTRL+C to quit
```

Figure 3-2 Successful deployment of RKLLM-Server-Flask in terminal

By referring to the specific code logic in build_rkllm_server_flask.sh, users can understand the detailed deployment process of the RKLLM-Server-Flask example. This enables users to customize the deployment implementation of their server more flexibly. It is important to emphasize that in step 3 of the one-click deployment script, the script automatically synchronizes the current version of RKLLM Runtime to rkllm_server/lib/librkllmrt.so. This ensures that flask_server.py calls the current version of librkllmrt.so during runtime.

3.4.1.2 Server-side: Introductions for RKLLM-Server-Flask Example

In this section, we will outline and introduce the implementation approach of the deployment