

export RKLLM LOG LEVEL=2

3.4 Implementation of Board-side Server

After using RKLLM-Toolkit to convert the model and obtain the RKLLM model, users can deploy server-side services on Linux development boards. This involves setting up a server on a Linux device and exposing network interfaces to everyone in the local area network. Subsequently, the RKLLM model can be accessed by other users via the specified address, thus facilitating efficient and concise interaction. This section will introduce two different server deployment implementations.

- 1) RLM-Server-Flask based on Flask: Users can achieve API access between the client and server using request requests. In the provided RKLLM-Server-Flask example, the send-receive structure is specially set to be the same as the OpenAI-API interface, facilitating quick replacement for users on existing development bases.
- 2) RKLLM-Server-Gradio based on Gradio: By reference to the provided example, users can rapidly construct a web server for visual interaction. Furthermore, the example illustrates the utilisation of the Gradio API interface, which enables users to undertake secondary development.

Both examples of server implementations mentioned above are located in the examples/rkllm_server_demo directory. This directory contains specific code for both implementations, one-click deployment scripts, and API interface calling examples. Users can choose different examples for reference and further development. The directory structure is as follows:

```
examples/rkllm server demo
-rkllm server
                   # Board-side Deployment Required Files
  — lib
                         # RKLLM Runtime
   - flask server.py
                            # RKLLM-Server-Flask Example
   - gradio server.py
                           # RKLLM-Server-Gradio Example
-build rkllm server gradio.sh # One-click Deployment Script -Gradio
chat api flask.py
                        # API Interface Example -Flask
chat api gradio.py
                         # API Interface Example -Gradio
Readme.md
```

3.4.1 Deployment Example of RKLLM-Server-Flask

In the deployment example of RKLLM-Server-Flask, the main focus is on using the Flask