

Returns	<p>0: Indicates the success of the chat template.</p> <p>-1: Indicates failure to set the chat template.</p>
---------	--

3.3 Board-side Inference Example

The directory, `examples/DeepSeek-R1-Distill-Qwen-1.5B_Demo`, includes the C++ project for the inference on the board-side, and which is synchronously updated with this documentation. Furthermore, compilation scripts are provided for users to facilitate the compilation of the project and the completion of the board-side inference of the RKLLM model.

3.3.1 Complete Code of Example Project

The complete C++ code example for inference calls is located in the `example/DeepSeek-R1-Distill-Qwen-1.5B_Demo/deploy/src` directory of the toolchain. The `llm_demo.cpp` file serves as an example of large language model inference. These examples include the full process, including model initialization, inference, handling outputs with callback functions, and releasing model resources. Users can refer to the relevant code to implement custom functionality.

3.3.2 Instructions for Example Project

Under the directory of `examples/DeepSeek-R1-Distill-Qwen-1.5B_Demo/deploy`, not only does it contain sample code for invoking the RKLLM model inference, but also includes compilation scripts `build-android.sh` and `build-linux.sh`. This section will provide a brief explanation of how to use the sample code, taking compiling executable files for Linux systems as an example using the `build-linux.sh` script:

Firstly, users need to prepare cross-compilation tools on their own, noting that the recommended version of the compilation tools in section 2.3 is 10.2 or above. Subsequently, before the compilation process, users should replace the path to the cross-compilation tools in `build-linux.sh` themselves: