Introduction	Used for simulate model inference on the PC.
Parameters	***messages:*** The text input, which needs to include the appropriate prompts. ***args:*** Inference configuration parameters, such as sampling parameters like top_k.
Returns	The logits values inferred by the model.

The example code is as follows:

```
args ={
    "max_length":128,
    "top_k":1,
    "temperature":0.8,
    "do_sample":True,
    "repetition_penalty":1.1
    }

mesg = "Human: How’s the weather today?\nAssistant:"
print(llm.chat_model(mesg, args))
```

The above operations cover all steps of model conversion and quantization in the RKLLM-Toolkit. Depending on different requirements and application scenarios, users can choose different configuration options and quantization methods for customised settings, which facilitates subsequent deployment.

3.2 Inference Implementation in Board-side

This chapter introduces the usage of the general API interface functions. Users can refer to the content of this chapter to construct C++ code and implement inference of RKLLM models on the board to obtain inference results. The RKLLM board-side inference implementation is as follows:

1) Define the callback function callback().
2) Define the RKLLM model parameter structure RKLLMParam.
3) Initialize the RKLLM model with rkllm_init().
4) Perform model inference with rkllm_run().
5) Process the real-time inference results returned by the callback function callback().
6) Destroy the RKLLM model and release resources with rkllm_destroy().

In the subsequent parts of this chapter, the document will provide detailed explanations of each step in the process and provide detailed explanations of the functions involved.