Fuctiom	rkllm_ abort
Introduction	Used to interrupt the RKLLM model inference process.
Parameters	***LLMHandle handle:*** the target handle registered during model initialization;
Returns	**0** indicates successful interruption of the model; **-1** indicates a failure to interrupt the model.

The example code is as follows:

```
// llmHandle is the the target handle registered
rkllm_abort(llmHandle);
```

3.2.7 Release Model

After completing all model inference calls, users need to call the rkllm_destroy() function to destroy the RKLLM model and release the CPU, GPU, and NPU computing resources allocated, for use by other processes or models. The specific function definition is as follows:

Table 3-23 Interface Specification for the rkllm_destory Function

Fuctiom	rkllm_ destroy
Introduction	Used to destroy the RKLLM model and release all computing resources.
Parameters	***LLMHandle handle:*** the target handle registered during model initialization;
Returns	**0** indicates successful destruction and release of the RKLLM model; **-1** indicates a failure in releasing the model.

The example code is as follows:

```
// llmHandle is the the target handle registered
rkllm_destroy(llmHandle);
```

3.2.8 Load LoRA Model

RKLLM supports running LoRA models simultaneously with the base model during inference. Before invoking the rkllm_run interface, you can load a LoRA model via the rkllm_load_lora interface. RKLLM allows loading multiple LoRA models; each call to rkllm_load_lora loads one LoRA model. The specific function definition is as follows: