Table 3-24 Interface Specification for the rkllm_load_lora Function

Fuctiom	rkllm_load_lora
Introduction	Used to load LoRA model for the base model.
Parameters	***LLMHandle handle:*** The target handle registered during model initialization. See section 3.2.4 on initializing the model. ***RKLLMLoraAdapter lora_adapter*:*** Parameter for loading the LoRA model.
Returns	**0** indicates the LoRA model was successfully loaded. **-1** indicates the model loading failed.

Table 3-25 Explanation of RKLLMLoraAdapter Structure

Definition	RKLLMLoraAdapter
Introduction	Used to configure parameters when loading a LoRA model.
Struct Fields	***const char* lora_adapter_path:*** The path to the LoRA model to be loaded. ***const char* lora_adapter_name:*** The name of the LoRA model to be loaded, defined by the user, used to select the specified LoRA during inference. ***float scale:*** The degree to which the LoRA model adjusts the base model parameters during inference.

Here is an example Code for Loading LoRA:

```
RKLLMLoraAdapter lora_adapter;
memset(&lora_adapter, 0, sizeof(RKLLMLoraAdapter));
lora_adapter.lora_adapter_path = "lora.rkllm";
lora_adapter.lora_adapter_name = "lora_name";
lora_adapter.scale = 1.0;
ret = rkllm_load_lora(llmHandle, &lora_adapter);
if (ret != 0) {
    printf("\nload lora failed\n");
}
```

3.2.9 Load Prompt Cache

During the model inference process, the Prefill stage typically consumes a significant amount of computational resources and time, especially when the Prompt is long. To accelerate this process,