Fuctiom	rkllm_clear_kv_cache
Introduction	Used to clear the KV cache.
Parameters	***LLMHandle handle:*** The target handle registered during model initialization (refer to section 3.2.4 Initialization Model). ***keep_system_prompt:*** Whether to retain the system prompt (1 to retain, 0 to clear).
Returns	0: Indicates successful clearing of the KV cache. -1: Indicates failure to clear the KV cache.

3.2.11 Chat Template Settings

During model conversion, RKLLM automatically parses the chat_template field from the Hugging Face model's tokenizer_config.json file. In this field, system_prompt serves as the system prompt to guide model behavior, prompt_prefix acts as the prefix before user input, and prompt_postfix acts as the suffix after user input. During inference, RKLLM automatically applies the parsed prompt template. If modifications are needed, you can use the following function to reconfigure these settings.

The specific function definition is as follows:

Table 3-29 Interface Specification for the rkllm_set_chat_template Function

Fuctiom	rkllm_set_chat_template
Introduction	Used to set the prompt template.
Parameters	***LLMHandle handle:*** The target handle registered during model initialization (refer to section 3.2.4 Initialization Model). ***system_prompt:*** The system prompt guiding the model behavior. ***prompt_prefix:*** The prefix before user input. ***prompt_postfix:*** The suffix after user input.