RKLLM supports loading Prompt Cache from files. By reusing cached content, you can significantly reduce the time spent in the Prefill stage, thereby improving overall inference efficiency.

Before invoking the rkllm_run interface for inference, ensure that the prompt_cache_params parameters are correctly configured. This step allows the model to generate the corresponding Prompt Cache file after inference. When running inference for the first time, the system will automatically generate a Prompt Cache file. This file contains the intermediate results required for the Prefill stage, which can be reused in subsequent tasks. In subsequent inference tasks, you can load previously generated Prompt Cache files by calling the rkllm_load_prompt_cache interface.

The specific function definitions are as follows:

Table 3-26 Interface Specification for the rkllm_load_prompt_cache Function

Fuctiom	rkllm_load_prompt_cache
Introduction	Used to load Prompt Cache file.
Parameters	***LLMHandle handle:*** The target handle registered during model initialization (refer to section 3.2.4 Initialization Model). ***const char* prompt_cache_path:*** The path to the Prompt Cache file to be loaded.
Returns	**0** indicates the Prompt Cache file was successfully loaded. **-1** indicates loading failed.

Table 3-27 Interface Specification for the rkllm_release_prompt_cache Function

Fuctiom	rkllm_release_prompt_cache
Introduction	Used to release the Prompt Cache.
Parameters	***LLMHandle handle:*** The target handle registered during model initialization (refer to section 3.2.4 Initialization Model).
Returns	**0** indicates that the Prompt Cache model was successfully released. -1 indicates that the model release failed.

Note: