	more sensitive. The default value is 0.1; ***bool skip_special_token:*** whether to skip special tokens and not output them, such as the end-of-sequence token <EOS>. ***bool is_async:*** whether to use asynchronous mode. ***const char img_start*:*** option to set the start marker for multimodal input image encoding, which needs to be configured in multimodal input mode. ***const char img_end*:*** option to set the end marker for multimodal input image encoding, which needs to be configured in multimodal input mode. ***const char img_content*:*** option to set the content marker for multimodal input image encoding, which needs to be configured in multimodal input mode. ***RKLLMExtendParam extend_param:*** the special parameters for controlling inference. ***n_keep:*** The number of cache entries to retain at the beginning when clearing the KV cache. In multi-turn conversations, the n_keep value must be at least as long as the system_prompt length.

Table 3-11 Explanation of RKLLMExtendParam Structure

Definition	RKLLMExtendParam
Introduction	The special parameters for controlling inference.
Struct Fields	***int32_t base_domain_id:*** controls from which domain the RKLLM model starts initialization, default is 0. ***int8_t embed_flash:*** Controls whether to store the model's vocabulary in flash memory to save memory. Set to 0 to disable and 1 to enable. ***int8_t enabled_cpus_num:*** Sets the number of CPUs to use for inference. The range varies depending on the chip model. For RK3588/3576, the range is 1-8; for RK3562, the range is 1-4, with the default set to 4.