Struct Fields	***const char* model_path:*** the path to the RKLLM model file; ***int32_t num_npu_core:*** the number of NPU cores used during model inference; The "rk3576" platform has configurable range [1, 2]; while the "rk3588" is [1, 3]; ***bool use_gpu:*** whether to use GPU for prefill acceleration, default option is false; ***int32_t max_context_len:*** the maximum context length during inference; ***int32_t max_new_tokens:*** the maximum number of generated tokens in inferencing; ***int32_t top_k:*** top-k sampling is a text generation method that selects the next token only from the top-k tokens with the highest probabilities predicted by the model. This method helps reduce the risk of generating low-probability or meaningless tokens. A higher top-k value (e.g., 100) will consider more token choices, resulting in more diverse text generation, while a lower value (e.g., 10) will focus on the most probable tokens, generating more conservative text. The default value is 40; ***float top_p:*** top-p sampling, also known as nucleus sampling, is another text generation method that selects the next token from a group of tokens with cumulative probabilities of at least p. This method balances diversity and quality by considering the probabilities of tokens and the number of sampled tokens. A higher top-p value (e.g., 0.95) results in more diverse text generation, while a lower value (e.g., 0.5) generates more focused and conservative text. The default value is 0.9; ***float temperature:*** a hyperparameter that controls the randomness of generated text by adjusting the probability distribution of output tokens. A higher temperature (e.g., 1.5) makes the output more random and creative. When the temperature is high, the model considers more options with lower probabilities when selecting the next token, resulting in more diverse and unexpected outputs. A lower temperature (e.g., 0.5) makes the output more focused and conservative. Lower temperatures mean that the model is more likely to choose high-probability tokens, resulting in more consistent and predictable outputs. In the extreme case of a temperature of 0, the model always