Client

chat bot

ChatIO

prompt_for_input

prompt_for_output

stream_output

conversation

conv_id : optional [Any]

conv_id : string

message : List

model name :

offset : Int.

roles : List.

sep : str.

controller

dispatch_method.

heart_beat_thread

worker_info.

cache flow worker

block_size

context_len : int

controller_addr.

is_server_running : bool

seq_counter.

model.

Model worker

context_len

controller_addr.

device

generate_stream_func

model

model_name.

tokenizer

worker_addr

worker_id