

NAME

cmchat – interactive chat with a local language model

SYNOPSIS

cmchat [*OPTION*]... *MODEL* [*PROMPT*]

DESCRIPTION

cmchat loads *MODEL* and reads questions at a prompt with line editing, answering each as a turn of one conversation. The answer is streamed as it is generated. Tokens between the chat template's thinking markers are shown dimmed. After each answer a line reports tokens generated, throughput, time to the first token, prompt tokens read and reused, why generation stopped, and context use.

PROMPT is the first question. *MODEL* is any model **cmgenerate**(1) loads; the chat template is detected from the vocabulary.

The KV cache is kept between turns, so each turn processes only its new tokens.

OPTIONS**Generation**

-c, --context *N*

Context length to allocate. Default: 4096.

--window *N*, **--sinks** *N*

Sliding-window attention; see **cmgenerate**(1).

-n, --steps *N*

Maximum tokens per answer. Default: until the end of the turn, bounded by the remaining context.

-t, -k, -p, -s

Temperature, top-k, top-p, seed; as in **cmgenerate**(1). The sampler state persists across turns, so a repeated question can get a different answer.

--system *TEXT*

System prompt.

--raw

No chat template; continue the input as text.

--speculate *N*, **--no-mtp**, **--mtp-from** *PATH*

Speculative decoding; as in **cmgenerate**(1).

--effort *LEVEL*

Reasoning effort; as in **cmgenerate**(1).

Placement

--gpu, **--gpu-backend**, **--gpu-device**, **--gpu-as**, **--cpu-as**, **--kv-format**, **--lora**, **--lora-as**: as in **cmgenerate**(1).

COMMANDS

A line starting with / is a command; // sends a line starting with /.

/help List the commands.

/new Start the conversation over, keeping the model.

/system *TEXT*

Set the system prompt and start over.

/temperature, **/top-k**, **/top-p**, **/seed**, **/steps**

Set the sampling option; **/steps 0** means until the end of the turn. **/settings** shows the current values.

/context

Context use.

/tokens *TEXT*

Tokenisation of *TEXT*.

/quit Exit.

LINE EDITING

Arrow keys, Ctrl-R history search (this session only) and the other readline bindings work. A line ending in \ continues on the next line. Ctrl-C discards the current input; Ctrl-D at an empty prompt exits. When the context is full, **/new** starts over.

SEE ALSO

cmgenerate(1), **cmagent(1)**, **cmserve(1)**