NAME

cmserve − OpenAI-compatible HTTP server for a local language model

SYNOPSIS

cmserve [*OPTION*]... *MODEL*

DESCRIPTION

cmserve loads *MODEL* and serves it over HTTP with the request and response format of the OpenAI API. **cmgenerate −−serve** runs the same server.

Requests are served one at a time: there is one KV cache. A request whose messages begin with those of the previous request reuses the cached prefix, so a client that resends a growing transcript pays only for the new turn. Requests from clients that take turns do not share a prefix and are processed in full.

There is no TLS and no authentication. A bare port given to **−−listen** binds 127.0.0.1; any other interface must be given as a full socket address.

ENDPOINTS

GET /
: Plain-text list of the endpoints.

GET /health
: **{"status": "ok", "model":** *NAME***}**

GET /v1/models
: The one model, as *NAME*.

POST /v1/chat/completions
: Chat completion from **messages**, a list of **{role, content}** objects rendered with the model's chat template. **content** may be a string or a list of parts, of which the text parts are used. An unknown role is treated as user.

POST /v1/completions
: Text completion of **prompt**, without the chat template.

Both completion endpoints read **temperature**, **top_p**, **top_k** (not in the OpenAI API), **seed**, **max_tokens** or **max_completion_tokens**, and **stream**. With **stream: true** the response is a server-sent event stream of chunks. Other fields are ignored. A missing field takes the value given on the command line.

OPTIONS

Server

−−listen *ADDR*
: Socket address, or a port on 127.0.0.1. Port 0 binds a port chosen by the kernel; the bound address is printed. Default: 127.0.0.1:8080.

−−name *TEXT*
: Model name reported to clients. Default: the file stem of *MODEL*.

Generation

Defaults for requests that do not set them.

−c, **−−context** *N*
: Context length to allocate; shared by all requests. 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).

−−raw
No chat template.

−−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−split**, **−−gpu−as**, **−−cpu−as**, **−−kv−format**, **−−lora**, **−−lora−as**: as in **cmgenerate**(1).

EXAMPLES

```
cmserve model.gguf −−gpu −−listen 8080
curl −s localhost:8080/v1/chat/completions \
    −d '{"messages": [{"role": "user", "content": "Hello"}]}'
OPENAI_BASE_URL=http://127.0.0.1:8080/v1 some−openai−client
```

SEE ALSO

cmgenerate(1), **cmchat**(1), **cmagent**(1)