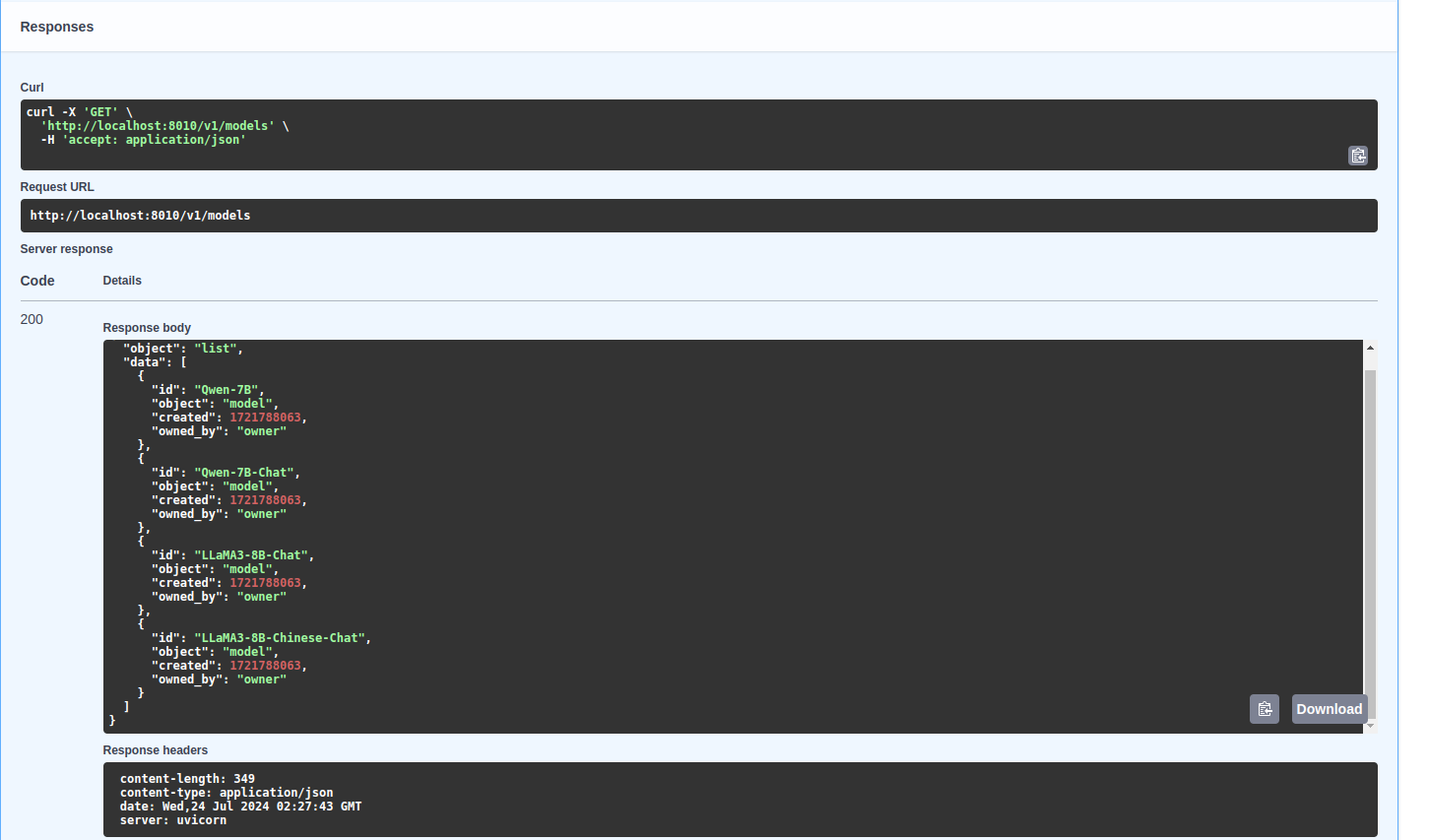
获取可用模型接口

curl -X 'GET' '<http://localhost:8010/v1/models>' -H 'accept: application/json'

无参数

{ "object": "list", "data": [ { "id": "Qwen-7B", "object": "model", "created": 1721788063, "owned\_by": "owner" }, { "id": "Qwen-7B-Chat", "object": "model", "created": 1721788063, "owned\_by": "owner" }, { "id": "LLaMA3-8B-Chat", "object": "model", "created": 1721788063, "owned\_by": "owner" }, { "id": "LLaMA3-8B-Chinese-Chat", "object": "model", "created": 1721788063, "owned\_by": "owner" } ] }



# 模型接口

curl -X 'POST' \

'http://localhost:8010/v1/chat/completions' \

-H 'accept: application/json' \

-H 'Content-Type: application/json' \

-d '{

"model": "LLaMA3-8B-Chat",

"messages": [

{

"role": "user",

"content": "hello"

}

],

"do\_sample": true,

"temperature": 0.95,

"top\_p": 0.7,

"n": 1,

"max\_tokens": 4096,

"stream": true

}'

模型没有加载的情况

{ "id": "chatcmpl-0b311fdee324454db07d46e4670ce89e", "object": "chat.completion", "created": 1721800810, "model": "demo\_model\_name", "choices": [ { "index": 0, "message": { "role": "assistant", "content": "model not load", "tool\_calls": null }, "finish\_reason": "stop" } ], "usage": { "prompt\_tokens": 10, "completion\_tokens": 20, "total\_tokens": 30 } }

模型加载的情况返回流输出

# 模型卸载

curl -X 'POST' \

'http://localhost:8010/v1/unload\_model' \

-H 'accept: application/json' \

-H 'Content-Type: application/json' \

-d '{

"id": "LLaMA3-8B-Chat",

"object": "model",

"created": 0,

"owned\_by": "owner"

}'

返回

{ "id": "LLaMA3-8B-Chat", "object": "model", "created": 1721806213, "owned\_by": "owner" }

# 加载模型

curl -X 'POST' \

'http://localhost:8010/v1/load\_model' \

-H 'accept: application/json' \

-H 'Content-Type: application/json' \

-d '{

"id": "LLaMA3-8B-Chat",

"object": "model",

"created": 0,

"owned\_by": "owner"

}'

返回值

{

"id": "LLaMA3-8B-Chat",

"object": "model",

"created": 1721806614,

"owned\_by": "owner"

}