Prof. Aviv Zohar

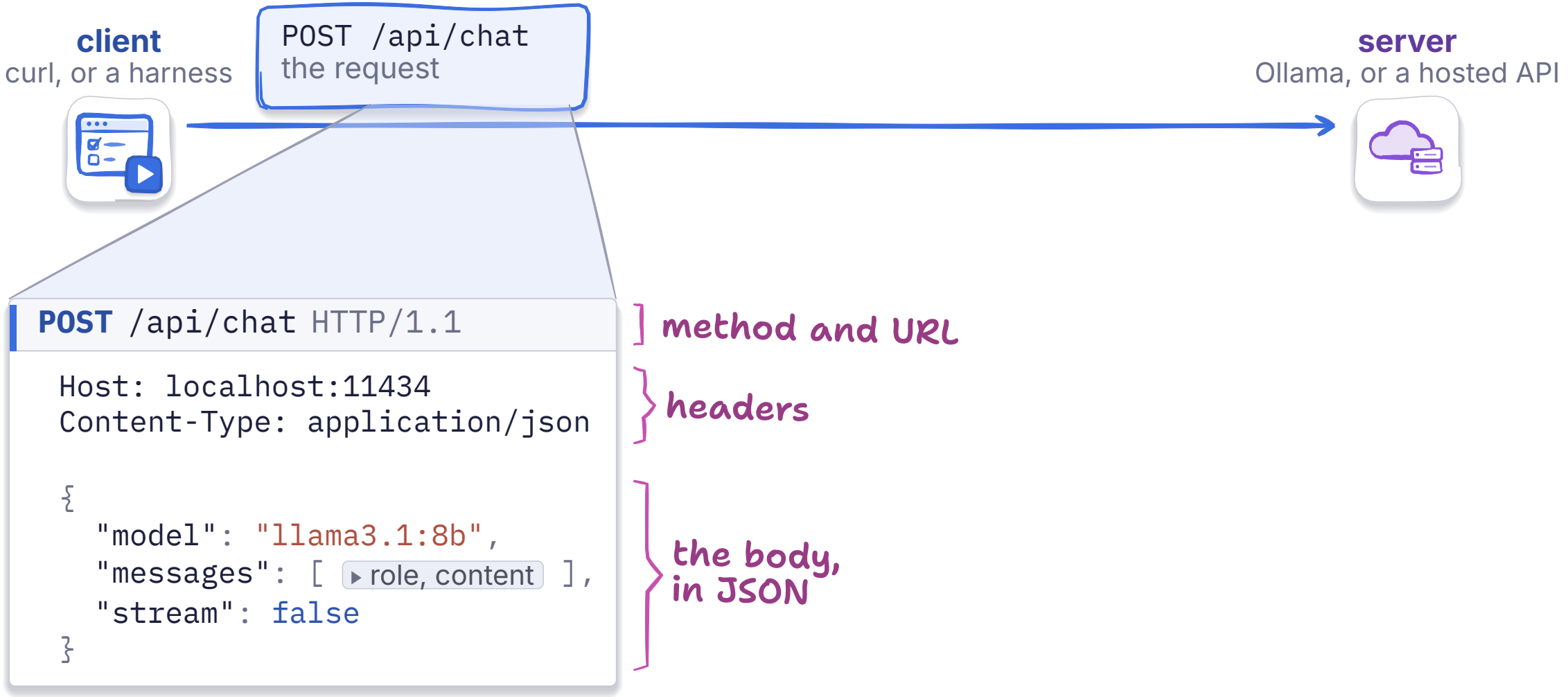
Semester A, 2026–7

Crafting an API Request

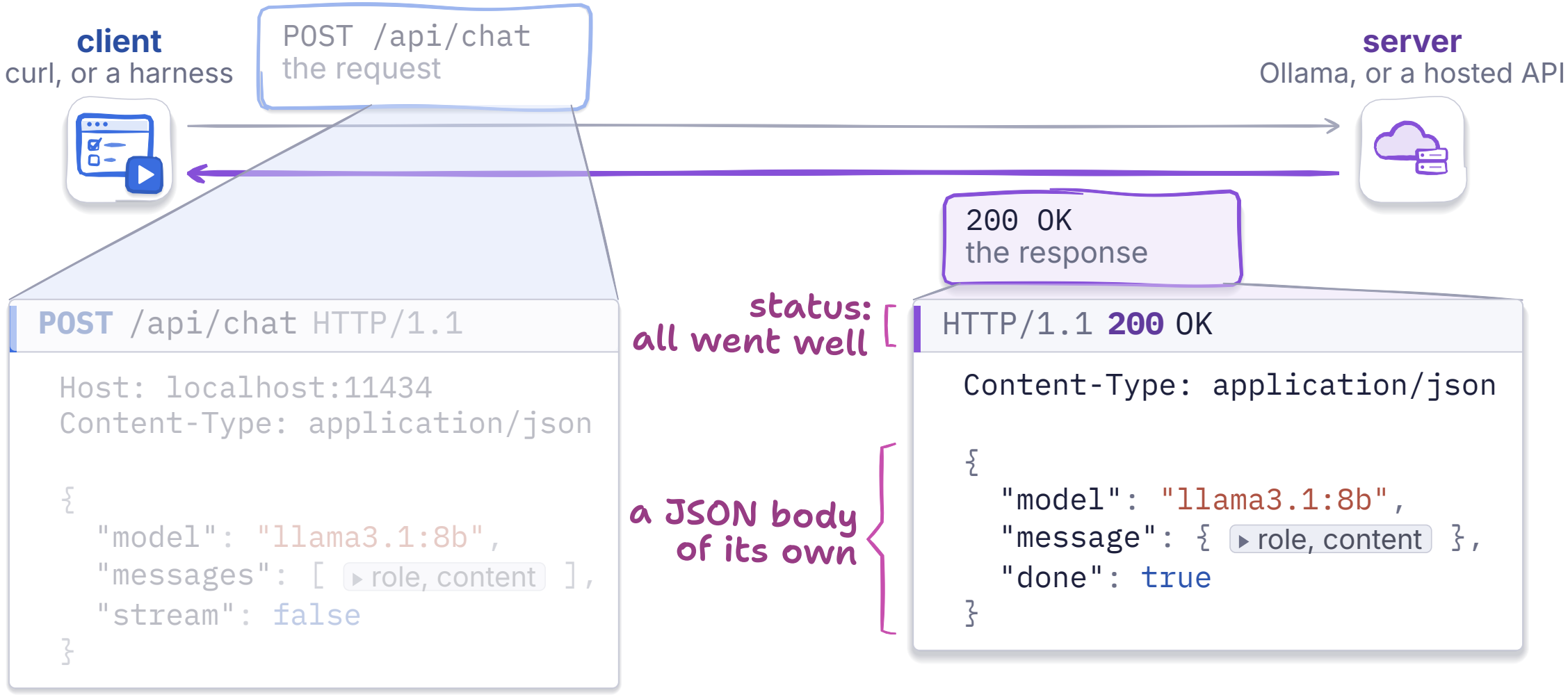
One HTTP request by hand, and the same shape at a hosted API



An API call is a request and its response



An API call is a request and its response



An API call is a request and its response

client
curl, or a harness

server
Ollama, or a hosted API

`POST /api/chat`
the request

A REST API: a program reached at URLs that take and return JSON over HTTP.





`200 OK`
the response

POST `/api/chat` HTTP/1.1

```
Host: localhost:11434
Content-Type: application/json

{
  "model": "llama3.1:8b",
  "messages": [ ▸ role, content ],
  "stream": false
}
```

HTTP/1.1 **200** OK

```
Content-Type: application/json

{
  "model": "llama3.1:8b",
  "message": { ▸ role, content },
  "done": true
}
```

One request by hand, with curl

shell

```
~ $ curl localhost:11434/api/chat -d @turn1.json
```

-d sends the file as the body of a POST

turn1.json

```
{
  "model": "llama3.1:8b",
  "stream": false,
  "messages": [
    { "role": "user", "content":
        "What is the capital of France? One word." }
  ]
}
```

one JSON object comes back, not a stream

One request by hand, with curl

shell

```
~ $ curl localhost:11434/api/chat -d @turn1.json
{"model":"llama3.1:8b","created_at":"2026-09-06T17:07…
```

turn1.json

```
{
  "model": "llama3.1:8b",
  "stream": false,
  "messages": [
    { "role": "user", "content":
        "What is the capital of France? One word." }
  ]
}
```

HTTP/1.1 **200** OK

```
{
  "model": "llama3.1:8b",
  "message": {
    "role": "assistant",
    "content": "Paris."
  },
  "done": true,
  "done_reason": "stop",
  ▸ 7 more: times and counts
}
```

how it ended: more near the end

captured 2026-09-06 · Ollama 0.30.6

The second call knows nothing of the first

POST /api/chat — call 1

```
"model": "llama3.1:8b", "stream": false,
"messages": [
  { "role": "user", "content":
      "What is the capital of France? One word." }
]
```

reply Paris.

Your program now holds the question and the reply.

The second call knows nothing of the first

POST /api/chat — call 1

```
"model": "llama3.1:8b", "stream": false,
"messages": [
  { "role": "user", "content":
      "What is the capital of France? One word." }
]
```

reply Paris.

POST /api/chat — call 2

```
"model": "llama3.1:8b", "stream": false,
"messages": [
```

no earlier messages

```
  { "role": "user", "content":
      "And of Italy?" }
]
```

reply You're referring to the Italian wine. The answer is Prosecco, ...

The second call knows nothing of the first

POST /api/chat — call 1

```
"model": "llama3.1:8b", "stream": false,
"messages": [
  { "role": "user", "content":
      "What is the capital of France? One word." }
]
```

reply Paris.

POST /api/chat — call 2

```
"model": "llama3.1:8b", "stream": false,
"messages": [
```

no earlier messages

```
  { "role": "user", "content":
      "And of Italy?" }
]
```

reply You're referring to the Italian wine. The answer is Prosecco, ...

The server holds no conversation state. The harness resends the whole history with every call.

The history travels in the request

POST /api/chat — call 1 · 1 message · 20 tokens in

```
"messages": [
  { "role": "user", "content":
      "What is the capital of France? One word." }
]
```



captured 2026-09-06 · Ollama 0.30.6 · token counts: prompt_eval_count

The history travels in the request

POST /api/chat — call 2 · 3 messages · 36 tokens in

```
"messages": [
  { "role": "user", "content":
      "What is the capital of France? One word." },
  { "role": "assistant", "content":
      "Paris." },
  { "role": "user", "content":
      "And of Italy?" }
]
```

reply Paris.

reply Rome.

captured 2026-09-06 · Ollama 0.30.6 · token counts: prompt_eval_count

The history travels in the request

POST /api/chat — call 3 · 5 messages · 57 tokens in

```
"messages": [
  { "role": "user", "content":
      "What is the capital of France? One word." },
  { "role": "assistant", "content":
      "Paris." },
  { "role": "user", "content":
      "And of Italy?" },
  { "role": "assistant", "content":
      "Rome." },
  { "role": "user", "content":
      "Which of the two is further north?" }
]
```

reply Paris.

reply Rome.

reply Paris is further north than Rome.

Each call sends the whole transcript, so the request grows as the conversation does.

captured 2026-09-06 · Ollama 0.30.6 · token counts: prompt_eval_count

A hosted API takes the same shape

POST /api/chat localhost:11434

```
{
  "model": "llama3.1:8b",
  "messages": [
    { "role": "user", "content":
        "…" }
  ]
}
```

A model name, and a list of messages, each with a role and content.

captured shape, contents elided

A hosted API takes the same shape

POST /api/chat localhost:11434

```
{
  "model": "llama3.1:8b",
  "messages": [
    { "role": "user", "content":
        "…" }
  ]
}
```

Not byte-for-byte the same.
Both send the model messages to continue.

POST https://api.anthropic.com/v1/messages

```
x-api-key: $ANTHROPIC_API_KEY
anthropic-version: 2023-06-01

{
  "model": "claude-opus-5",
  "max_tokens": 1024,
  "system": "You are a coding agent. …",
  "messages": [
    { "role": "user", "content": "…" }
  ],
  "tools": [ ▸ tool definitions ]
}
```

the output limit

hosted request illustrative

A hosted API takes the same shape

```
POST /api/chat                                  localhost:11434
{
  "model": "llama3.1:8b",
  "messages": [
    { "role": "user", "content":
        "…" }
  ]
}
```

```
POST https://api.anthropic.com/v1/messages
x-api-key: $ANTHROPIC_API_KEY
anthropic-version: 2023-06-01

{
  "model": "claude-opus-5",
  "max_tokens": 1024,
  "system": "You are a coding agent. …",
  "messages": [                                  in the context
    { "role": "user", "content": "…" }
  ],
  "tools": [ ▸ tool definitions ]
}
```



It builds this request: instructions, history and tools, in the provider's format.

hosted request illustrative

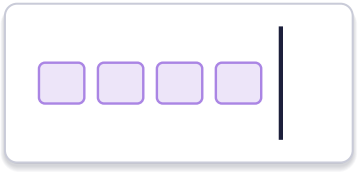
A few fields control sampling and length



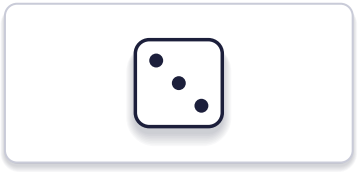
temperature
changes the sampling distribution



top-p, top-k
restrict which tokens are considered



max tokens
limits the length of the output



seed
controls the random sequence,
where supported

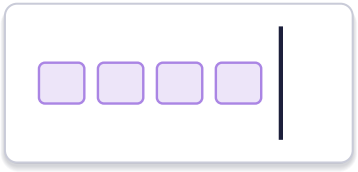
A few fields control sampling and length



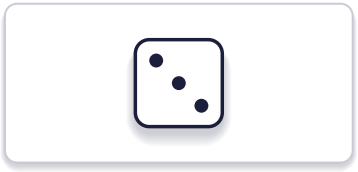
temperature
changes the sampling distribution



top-p, top-k
restrict which tokens are considered



max tokens
limits the length of the output



seed
controls the random sequence, where supported

Even with a seed, other hardware or serving versions can change the result.

POST /api/chat

```
{
  "model": "llama3.1:8b",
  "messages": [ ▸ … ],
  "options": {
    "temperature": 0.7,
    "top_p": 0.9,
    "top_k": 40,
    "num_predict": 256,
    "seed": 42
  }
}
```

Ollama calls the limit `num_predict`. Names and defaults vary by API and model.

option values illustrative

The reply also says why it stopped and what it used

```
HTTP/1.1 200 OK

{
  "model": "llama3.1:8b",
  "message": { ▸ assistant: Paris. },
  "done": true,
  "done_reason": "stop",          ⓐ
  "prompt_eval_count": 20,        ⓑ
  "eval_count": 3,
  ▸ 5 more: created_at, durations
}
```

ⓐ why it stopped

Anthropic's names

`end_turn`	finished
`tool_use`	asks for a tool
`max_tokens`	hit the output limit

Ollama's `"stop"` says it finished.

ⓑ how many tokens

20 in **3** out

A billed API charges by these counts.
Locally, the cost is your own compute.

The reply also says why it stopped and what it used

```
HTTP/1.1 200 OK

{
  "model": "llama3.1:8b",
  "message": { ▸ assistant: Paris. },
  "done": true,
  "done_reason": "stop",
  "prompt_eval_count": 20,
  "eval_count": 3,
  ▸ 5 more: created_at, durations
}
```

ⓐ

ⓑ

ⓑ how many tokens

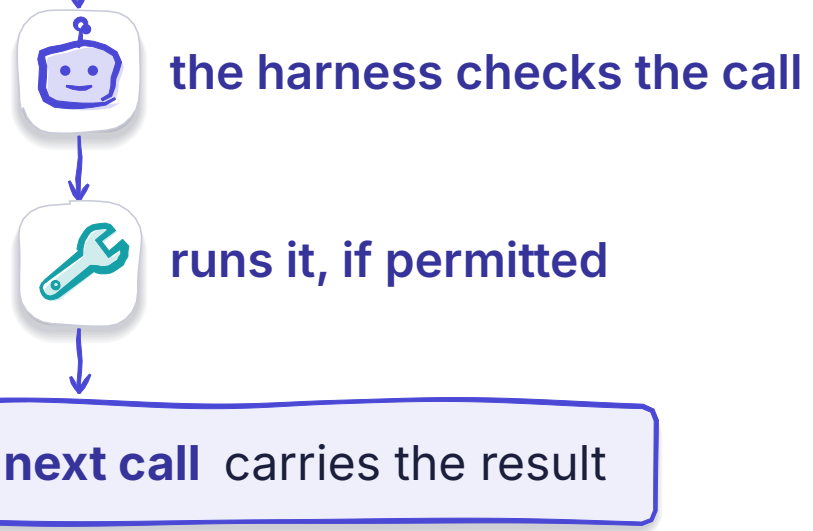
20 in **3** out

A billed API charges by these counts.
Locally, the cost is your own compute.

ⓐ why it stopped

Anthropic's names

`end_turn`	finished
`tool_use`	asks for a tool
`max_tokens`	hit the output limit

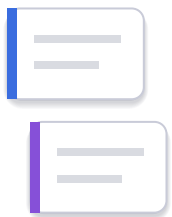


reply captured 2026-09-06 · Ollama 0.30.6

Building a model request

One call

A POST with a JSON body, and a JSON reply.



History in the body

The conversation is the list of messages you send.

Same shape, hosted

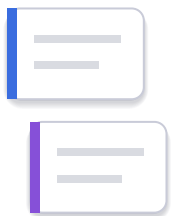
Another endpoint, a key, more fields; the same messages.



Building a model request

One call

A POST with a JSON body, and a JSON reply.



History in the body

The conversation is the list of messages you send.

Same shape, hosted

Another endpoint, a key, more fields; the same messages.



The harness builds this request on every turn of the agent loop.

See you in the next video.