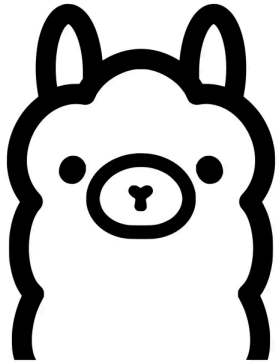


Ollama

本地快速上手大型语言模型

(Llama 3.1, Mistral, Gemma 2, ...)

运行服务模式

操作步骤

- 本地启动 `ollama` 服务
- 运行确认

操作演示

课堂实验

执行脚本

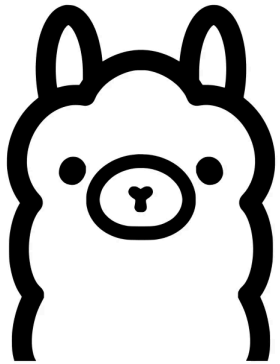
```
# 服务模式
ollama serve --help

# 启动服务
ollama serve

# 服务确认
curl http://localhost:11434/

# JSON解析工具
brew install jq
```

执行脚本

```
# API文档：https://github.com/ollama/ollama/blob/main/docs/api.md
curl http://localhost:11434/api/generate -d '{
  "model": "llama3.1:8b",
  "prompt":"你好",
  "stream": false,
  "options": {
    "temperature": 0.5
  }
}' | jq
# 过滤字段
curl http://localhost:11434/api/generate -d '{
  "model": "llama3.1:8b",
  "prompt":"你好",
  "stream": false,
  "options": {
    "temperature": 0.5
  }
}' \
| jq '.response'
# 过滤json多个字段
# | jq '  | {model,response}'
```

执行脚本

```
# 英文字典
curl http://localhost:11434/api/generate -d '{
  "model": "llama3.1:8b",
  "prompt":"book",
  "stream": false,
  "system": "你是一名语言专家，你的任务是翻译一段英文文本到中文。",
  "options": {
    "temperature": 0
  }
}' \
| jq '.response'
```



下课时间