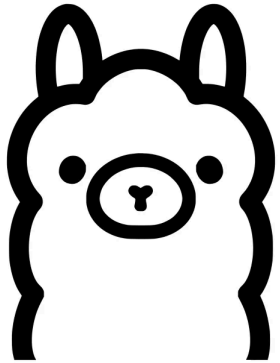

Ollama

本地快速上手大型语言模型

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

📚 LangChain - 链式调用

操作步骤

- 使用 `LangChain` 链式调用语法使用 `Ollama` 模型
- 运行确认

操作演示

main.py

```python
import time
from langchain_ollama import ChatOllama
from langchain.prompts import ChatPromptTemplate
from langchain.schema.output_parser import StrOutputParser  # 字符串输出解析器


def evalEndTime(start_time):
    end_time = time.time()  # 获取结束时间
    execution_time = "(程序运行时间：%.2f 秒)" % (
        end_time - start_time
    )  # 计算程序运行时间
    return execution_time


print("=" * 100)
start_time = time.time()  # 获取开始时间

llm = ChatOllama(
    model="llama3.1:8b",
    # model="gemma2:2b",
    # model="gemma2:9b",
    temperature=0.5,
)
print(">", llm)

# 提示词模版
messages = [
    ("system", "你是一位{career}专家，你经常辅导你的学生。"),
    ("human", "我想学习{language}，给我几个建议好吗？"),
]
prompt_template = ChatPromptTemplate.from_messages(messages)

# 划重点：使用Chain调用模型
chain = prompt_template | llm | StrOutputParser()

# 一次返回
# result = chain.invoke({"career": "医学", "language": "按摩"})
# print(result)

# 流式返回
for chunk in chain.stream({"career": "医学", "language": "按摩"}):
    print(chunk, end="", flush=True)

print()
# 打印结束时间
```

下课时间

课件下载： https://github.com/komavideo/lesson_ollama