## Langchain开发手册

实战一（2～7）

Macbook笔记本，位置在：~/pwork/gpt

前端代码位置：~/vwork/lawtool

公司GPT服务器，172.17.103.152，huangyinfei账户

后端代码位置：~/pwork/gpt

模型位置：~/model

数据位置：~/data

实战二（8～10）

DS主机，代码位置：~/pwork/langchain

《LangChain技术解密：构建大模型应用的全景指南(2024)》

DS主机，代码位置：~/pwork/langchain/agent

《大模型应用开发：动手做AI Agent(2024)》

实战三（11）

DS主机，代码位置：~/pwork/glm

《ChatGLM3大模型本地化部署、应用开发与微调(2024)》

1. 环境检测

查看GPU卡

nvidia-smi

查看CUDA版本

nvcc --version

1. GLM-4-9B模型

编写：infer.py

本地大模型：/home/huangyinfei/model/glm-4-9b-chat

调用model.generate接口访问

询问：从清华园到颐和园怎么走

1. Bge向量模型

编写：bge.py

本地向量模型：/home/huangyinfei/model/bge-large-zh-v1.5

检索本地语料

bge\_retriever.get\_relevant\_documents("熊猫")

1. Langchain问答

编写：lchain.py

本地大模型：/home/huangyinfei/model/glm-4-9b-chat

本地向量模型：/home/huangyinfei/model/bge-large-zh-v1.5

本地文档：test.txt（120问答语料）

服务器上有两张GPU卡，可以指定cuda:1将模型部署到第二张卡上

调用model.chat接口不成功，改为调用model.generate接口

询问：再查查数据相比裁判文书网是否更多更全

1. FastAPI封装chat服务

编写：infer\_service.py

编写：~/pwork/fastapi/main.py

VPN：连接360pai-办公室

GPT服务器启动服务：

uvicorn main:app --host 0.0.0.0 --port 8000 --reload

Macbook笔记本浏览器访问：

http://172.17.103.152:8000/query/?q=你好&c=我是飞行家，

1. 前端界面

创建项目：~/vwork/lawtool

cnpm create vite@latest

cnpm install vue-router@next -S

cnpm install element-plus -S

cnpm install axios

vite.config.js文件：

添加@ alias，以及对应import

添加server-proxy，8080端口，访问后端query接口

main.js添加：

app.use(router)

app.use(ElementPlus)

App.vue，整个替换掉

创建router/index.js，添加/chat路由

创建views/Chat.vue

textarea显示会话历史，el-input输入问题，el-button提交、重置

通过axios调用后端query接口，注意要用/query/

启动前端：npm run dev

前后端连通ok

访问浏览器：http://localhost:8080/#/chat

1. 法律问答语料

百度知道问答数据：lawzhidao\_filter.csv，3.6万条

罪名问答数据：CrimeKgAssitant，qa\_corpus.json，20万条

资管法规数据：LawRefBook

1. 模型服务

编写：chat.py

指定API网关和API\_KEY，即可通过openai接口调用大模型

ChatOpenAI接口可通过chat, predict, predict\_messages, generate方法进行问答

编写：chatcomp.py

OpenAI接口可通过chat.completions.create方法进行问答

1. 链式调用

编写：simplechain.py

提示词+大模型，链式调用：title => story => advertisement

编写：sequchain.py

提示词+大模型，链式调用：{ 'title', 'story\_type', 'name'} => {'title', 'story'} => {'title', 'story', 'adv'}

1. 智能体

编写：ReAct.py

通过create\_react\_agent集成DeepSeek大模型和SerpAPI搜索工具，即可增加联网检索能力

编写：ReAct2.py

通过create\_react\_agent集成DeepSeek大模型、SerpAPI搜索工具、llm-math数学工具，即可增加联网检索与数字计算能力

编写：Plan\_n\_Execute.py

通过PlanAndExecute智能体，先规划若干步骤，再按照步骤逐一执行，通过@tool标注自定义工具

1. GLM封装

模型下载到本地

apt install git-lfs

git lfs install

git clone https://huggingface.co/THUDM/chatglm3-6b-128k

pip install protobuf transformers==4.30.2 cpm\_kernels torch>=2.0 gradio mdtex2html sentencepiece accelerate

编写：4.1.2.py

用fastapi封装glm问答接口

编写：llmglm.py（4.1.3.py）

定义LLM子类，要有\_call接口。之后即可直接作为大模型导入LangChain

from llmglm import ChatGLM

llm = ChatGLM()

编写：\_4.1.4.py

用LangChain集成大模型、提示词模板、链式调用