## Langchain开发手册

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

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. 智能问答服务
2. 裁判文书语料
3. 裁判文书智能体
4. 律师文书语料
5. 律师文书智能体
6. GLM-130B INT8模型
7. 模型微调
8. 模型度量