## Langchain开发手册

Macbook笔记本，位置在：~/pwork/gpt

公司服务器，172.17.103.152，huangyinfei账户

代码位置：~/pwork/gpt

模型位置：~/model

数据位置：~/data

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. Langchain智能体
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. 前端界面
7. GLM-130B INT8模型
8. 模型微调
9. 模型度量