基本文件结构

AIserver文件夹下存放了所有的程序代码，其结构如下

AIserver/

├── data/（rag文件）

│ └── txt/

│ └── xlsx/

│ └── character.xlsx

├── EmoLLMRAGTXT/

│ └── vector\_db/(储存向量文件)

└── model/

│ └── embedding\_model/

│ └── bge-small-zh-v1.5.pkl

│ └── rerank\_model/(服务器中无法使用)

└── src/

│ └── config/

│ └── config.py（认证信息和变量信息）

│ └── until/

│ └── llm.py（模型调用函数）

│ ├── chat\_session.py（构建对话）

│ ├── chatbot.py（构建对话机器人）

│ ├── data\_processing\_for\_server.py（rag流程）

├── Server.py（服务器）

程序运行流程大致为：

1. data\_processing\_for\_server.py文件加载向量化模型，读取character.xlsx文件并对问题进行向量化，生成向量文件储存到EmoLLMRAGTXT等待调用。
2. 前端发送请求给服务器，请求中包括用户id、用户输入和调用模型信息，传输给相应接口
3. Server文件根据接口调用chatbot
4. Chatbot向chat\_session和data\_processing\_for\_server发送请求
5. data\_processing\_for\_server通过用户请求调用向量模型，返回相似回答，输出给chat\_session构建prompt
6. chat\_session调用llm.py把构建好的prompt发送给llm进行数据交换，返回结果
7. 结果通过server传给前端