# 周报 2025.12.8-2025.12.14

## 陈嘉乔 2022060909001

## 本周的工作

1. 调用 deepseek API 在本地成功复现了sgpt
   学习shell-gpt项目的组成结构等

```
(shellagent) PS K:\PROJECTS\shell_gpt> sgpt --shell "list all files in the current path by the order of its create time"
Get-ChildItem | Sort-Object CreationTime
[E]xecute, [M]odify, [D]escribe, [A]bort: e


    目录: K:\PROJECTS\shell_gpt


Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
d-----        2025/11/21     15:38                .devcontainer
d-----        2025/11/21     15:38                .github
-a----        2025/11/21     15:38           2564 CONTRIBUTING.md
-a----        2025/11/21     15:38            304 Dockerfile
-a----        2025/11/21     15:38           1093 LICENSE
-a----        2025/11/21     15:38          27454 README.md
-a----        2025/11/21     15:38           2760 pyproject.toml
d-----        2025/11/21     15:38                scripts
d-----        2025/11/28     14:09                sgpt
d-----        2025/11/21     15:38                tests
```

2. 下载Qwen2.5:7b模型，实现了一个基于Flask的HTTP服务器，用于托管Qwen-7B语言模型。
   主要工作包括：
   i. 加载Qwen2.5:7B-instruct模型的tokenizer和模型，使用4-bit量化（BitsAndBytesConfig）以优化内存使用。实测本地能跑（RTX3070Ti-laptop）
   ii. 配置模型为推理模式（eval），支持设备映射和内存限制。
   iii. 提供/generate POST端点，接收用户提示（prompt），生成文本响应。使用模型的generate方法，设置温度、top_p等参数控制生成质量。
   iv. 服务器运行在本地127.0.0.1:8000端口。

```
Loading checkpoint shards: 100%|████████████████████████████████████████| 4/4 [00:19<00:00,  4.94s/it]
Model server is ready.
 * Serving Flask app 'model_server'
 * Debug mode: off
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
 * Running on http://127.0.0.1:8000
```

3. 实现了一个命令行客户端，作为shell助手，使用LangChain框架与上述服务器交互。
   主要工作包括：
   i. 定义QwenHTTP类，继承LangChain的LLM接口，通过HTTP请求调用服务器的生成端点。
   ii. 加载外部提示模板（从prompts/shell_assistant.prompt），构建对话链，包括记忆（ConversationBufferMemory）和提示模板。
   iii. 运行CLI循环，接收用户输入，生成shell命令建议，并维护对话历史。
   iv. 集成RunnablePassthrough和链式调用，实现自然语言到安全shell命令的转换。

```
(shellagent) PS K:\PROJECTS\shell_agent> & K:/pros/anaconda/envs/shellagent/python.exe k:/PROJECTS/shell_agent/shell_agent_client.py
--------------------------------
k:\PROJECTS\shell_agent\shell_agent_client.py:52: LangChainDeprecationWarning: Please see the migration guide at: https://python.langchain.com/docs/versions/migrating_memory/
  memory = ConversationBufferMemory(memory_key="history", return_messages=False,)
User: hi
Assistant: Hello! How can I help you today with a shell command?
--------------------------------
User:
Assistant: Sure, I can help with that. Please provide the shell command you'd like to convert or execute.
--------------------------------
User: 查看当前目录下大小超过 100MB 的文件
Assistant: 要查看当前目录下大小超过100MB的文件，可以使用以下命令：

```sh
find . -type f -size +100M
```

这个命令会在当前目录及其子目录中查找所有大于100MB的文件。
```

## 下周的计划

1. 目前的 memory 非结构化、非持久化、不可回溯，改为使用SQLite，构建数据库存储
2. 构建(自然语言需求, 标准命令, 错误案例)的向量数据库，学习rag相关知识