工作总结9.29

1.LLM（大语言模型）

先写了一个单层的transformer块，输出这一批次文本数据中，所有单词的概率权重。

图形用户界面, 文本

描述已自动生成

文本

描述已自动生成基于transformer构建了一个本地小模型，给予提示词，预测输出后面的词语，将数据分为4个批次，每个批次的矩阵设置为64\*256，训练1000次，得到如下结果：

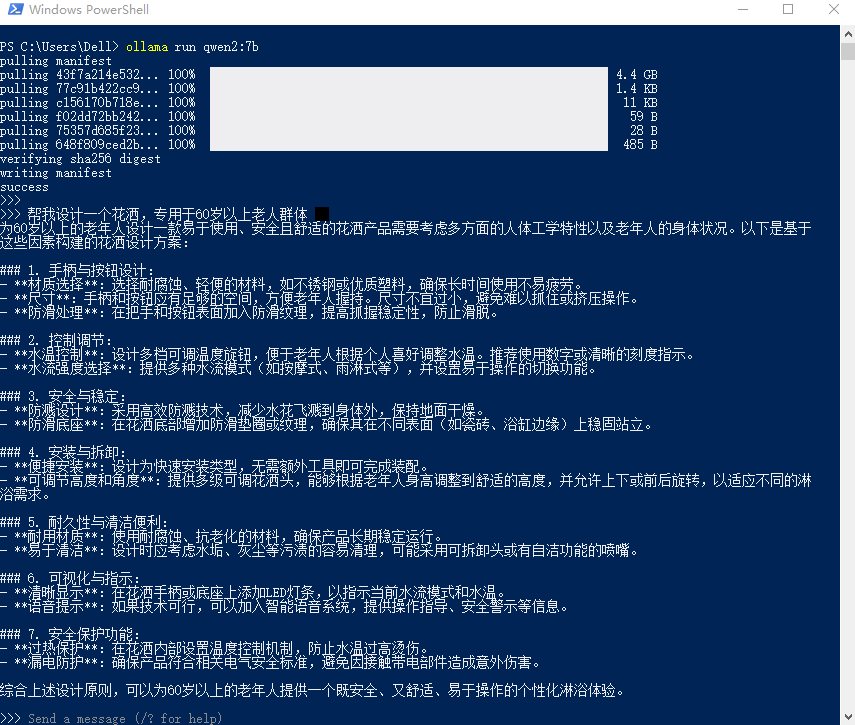
The boy the best solution that ambiguity and valued and about their solutions. Put yourself first impression that outweigh explaining how they say of you areiates their one purchase builds trust and credibility. By investing time these issues from a clear understanding of the customer feels, you can alleviate both you know summary and genuinely caring to engage with the solution.

Lastly, we can be used to address the customer's specific needs and tailor the customer. Pay close attention to this chapter that our words with their physical will enable them to build rapport

无法使用

2.使用开源LLM

选用24年6月阿里发布的QWEN 2模型作为底座，安装Ollama进行本地部署：



3.使用一个开源的前端界面open web-ui，方便对大模型以及Ollama操作，此ui需要使用Docker部署：

图形用户界面, 应用程序

描述已自动生成

图形用户界面, 文本, 应用程序

描述已自动生成在相应端口启动ui，调用Ollama并载入QWEN 2大模型：

下周工作：

LLaMA-Factory微调工具和LangChain微调框架

微调方法：Lora、P-tuning等