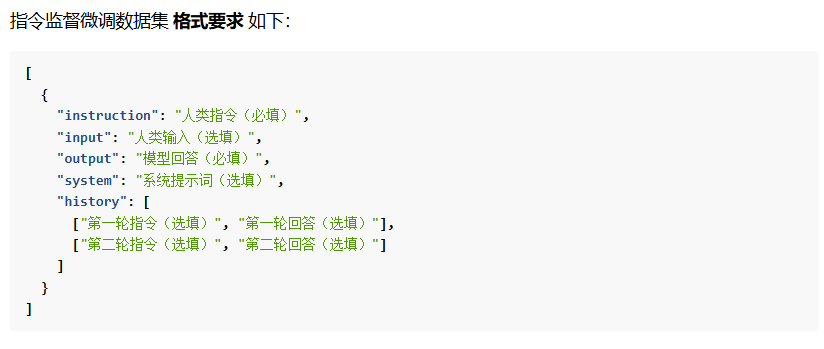
周报10.20

用于微调的问答对数据集的构建非常重要，是定制本地模型中最核心的环节，非常耗时。

1.数据集格式

根据LLamaFactory指定的两种数据集格式，[数据处理 - LLaMA Factory](https://llamafactory.readthedocs.io/zh-cn/latest/getting_started/data_preparation.html)，选择较易处理的Alpaca格式，将数据表转换为Alpaca格式的Json数据，用于模型微调。



2.使用现有的项目

Bonito是一个由Brown大学BATS研究实验室开发的开源项目,旨在解决：将未标注的文本转换为特定任务的指令微调数据集。

其本身是一个微调过的Mistral 7B模型，专门用于生成问答对。

2.1项目安装

从git手动下载量化版本项目，可在本地电脑运行[GitHub - BatsResearch/bonito: A lightweight library for generating synthetic instruction tuning datasets for your data without GPT.](https://github.com/BatsResearch/bonito)，

打开项目，为防止环境冗余混乱，新建一个conda环境，并使用命令：

conda create -n bonito python=3.9

conda activate bonito

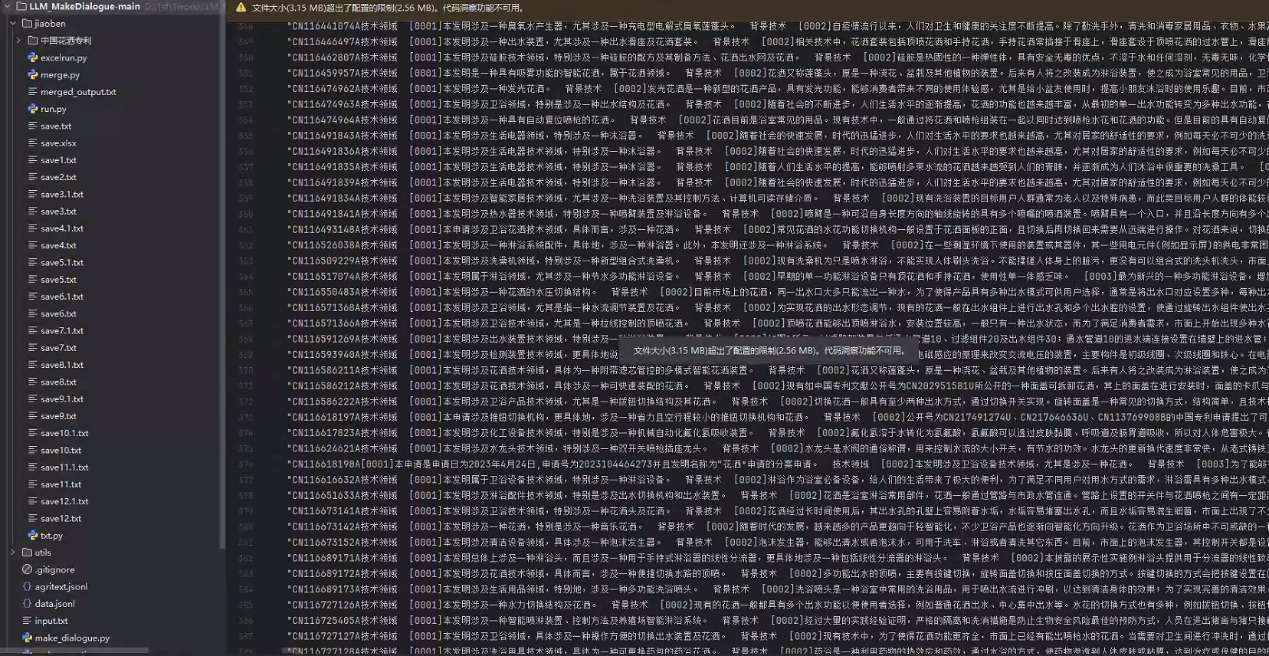
pip install -e .

需要vllm来安装并运行模型，安装vllm时遇到较大问题，在本地和虚拟机都进行了尝试，最终耗时3天左右项目仍无法运行，遂放弃，采用其他方法。

3.使用现有大模型处理

3.1处理表格

当前主流的做法是：先使用python脚本将原excel文件进行处理成txt文件，并去除无效冗余的内容，否则大模型对文本的解析会有很多歧义和影响。



2.生成问答对

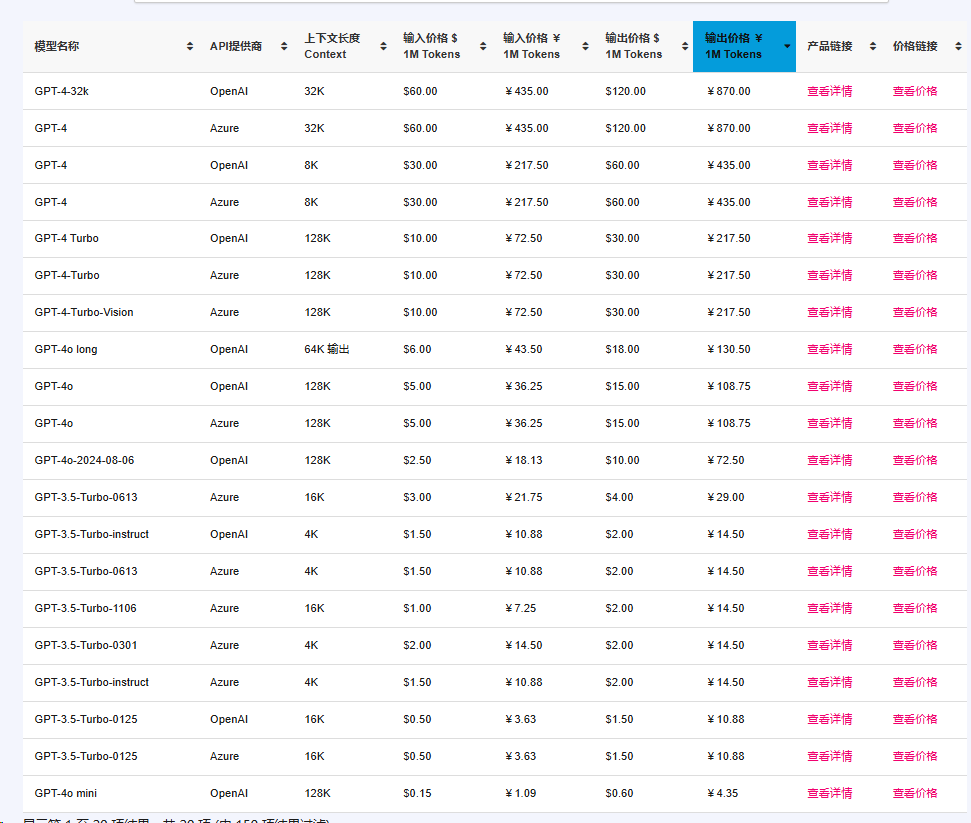
再使用CHATGPT，对txt文件进行解析并生成问答对，同时需要对prompt提出非常细致的要求，以产生可用的问答对。

2.1

需要购买大量APIKEY，来调用GPT生成。采用中转而非直连API，节省成本。

一开始调用GPT3.5turbo生成了一些问答对，效果很不理想

再调用了GPT4，效果非常好，遂采用GPT4开始生成，但生成了100对左右余额就用完了，查询了一下GPT4的API价格，非常非常高



使用多个GPT版本进行效果比对，最终选择GPT-4o-mini，产生的问答对效果还可以。

2.2脚本调用

生成一个问答对需要5-10s左右，中途若网络波动就会断开，也非常耗时。目前生成了3000对问答对，用于模型微调。

3.模型效果

现有问题：回答还是较简单，且专利号不正确，并没有对上

下周工作：生成全部的问答对，并将专利号对上。