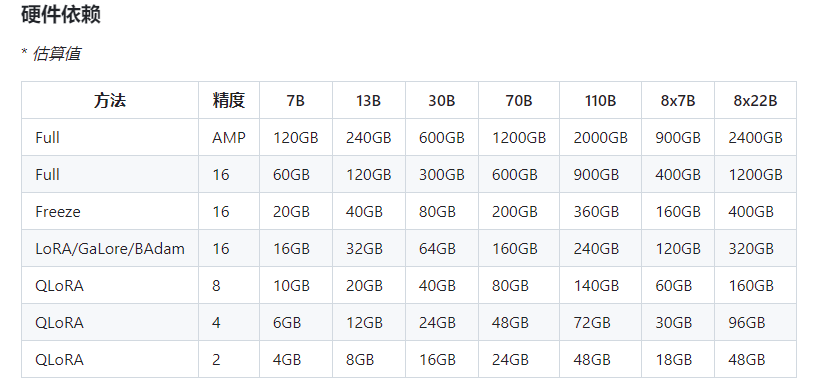
10.13周报

本周进行微调，生成了一个初步的模型。

1.主机准备：

之前Ollama安装的是特殊量化处理过的模型，并不是全量模型，全量模型需要的显存如下图：



电脑显存有12G，采用in8量化版本进行微调是足够使用的，但在最后会出现：量化模型在Lora微调后无法进行合并导出的问题，所以在算力云上租了一个4090D主机，采用qwen2.5-7b-instruct全量对话模型进行微调。

2.微调前准备

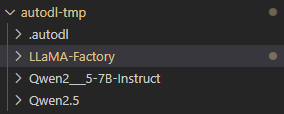
2.1安装qwen2.5-7b

启动主机，在数据盘下载qwen2.5底座，再下载qwen2.5-7b-instruct模型，此时已经可以运行模型：



2.2安装LLaMA-Factory

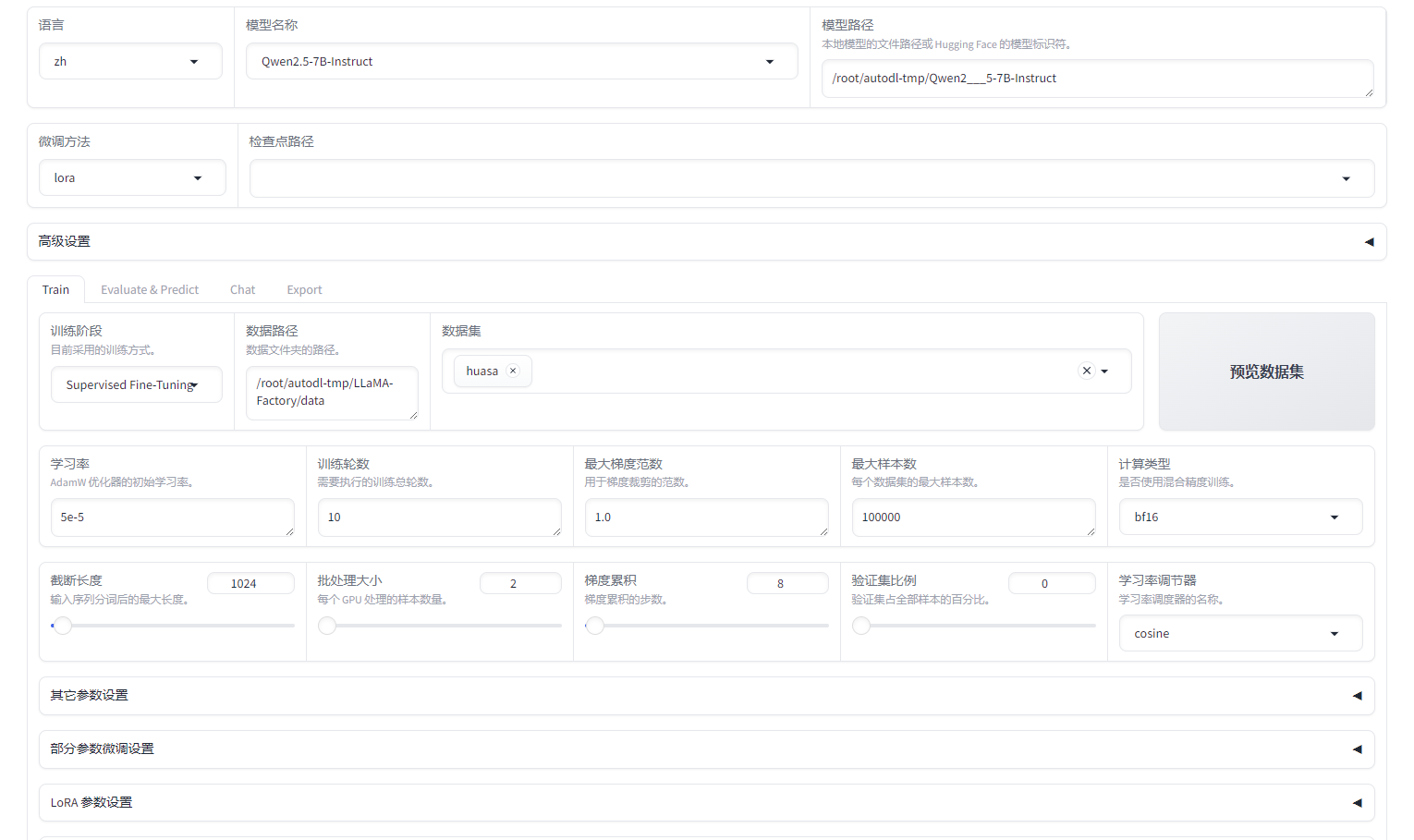
使用git安装开源微调项目 LLaMA-Factory，遇到依赖安装问题，使用指令pip install -r requirements.txt，安装所有指定依赖。



安装完成。

3.微调训练模型

运行LLaMA-Factory/src/webui.py，启动LLaMA-Factory项目



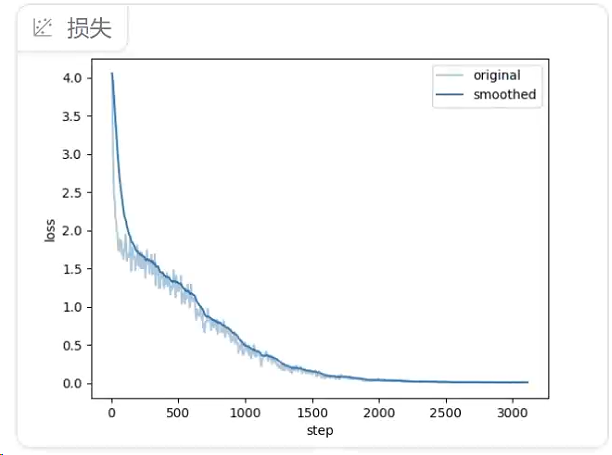
选择模型和数据集，都要使用本地路径。

数据集由excel转换而来。

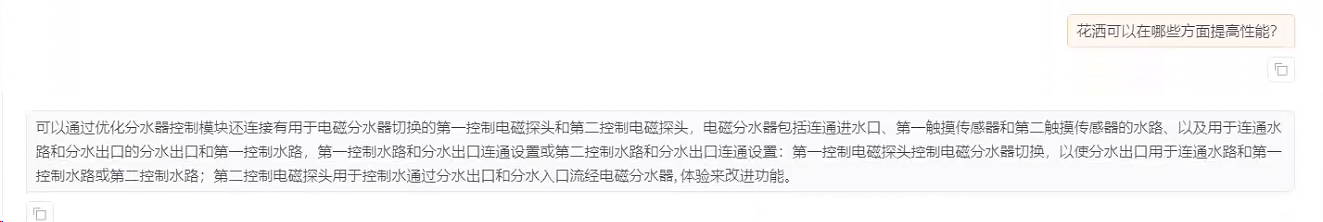
数据在data文件夹，将准备的数据集复制到此处，其中dataset\_info为数据集路径配置，添加huasa.json路径，即可显示数据集文件。

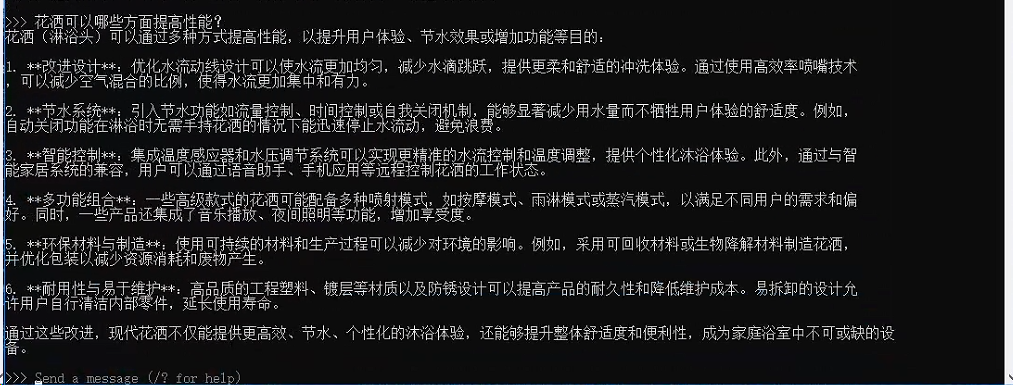
默认Lora训练的参数，设置训练10次，训练过程如图：

训练完成后。得到模型的训练损失曲线图：

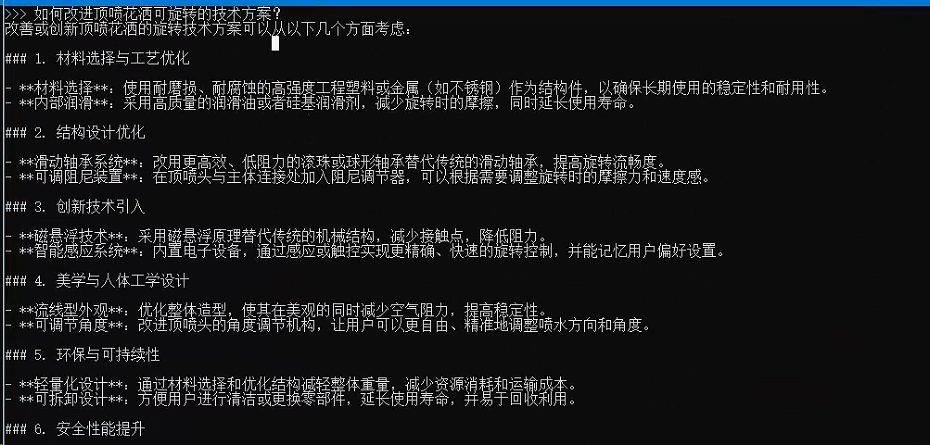


在微调后的模型上，对话测试, 并对比原模型：









下周任务：

研究如何生成更好的数据集，并研究微调的详细参数对模型的影响，再进一步训练模型。