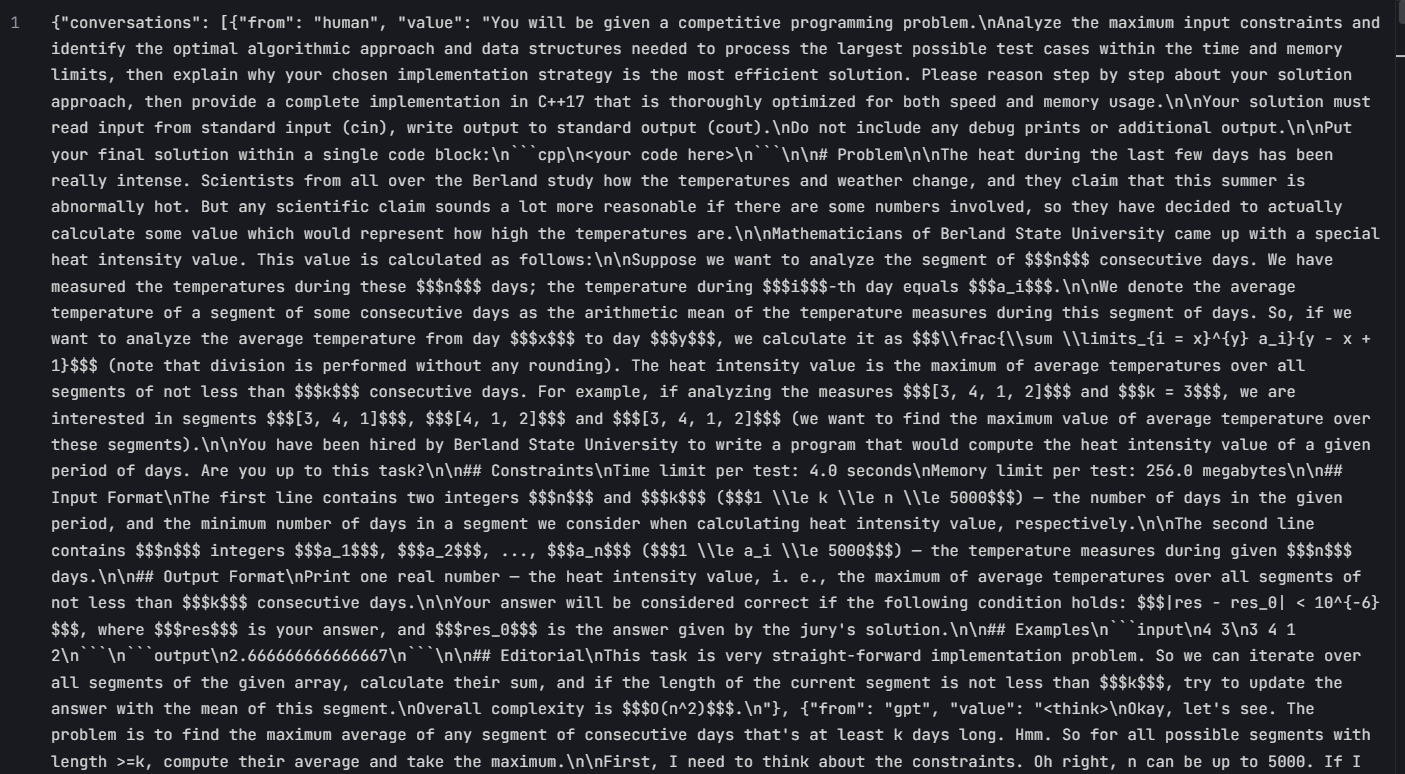
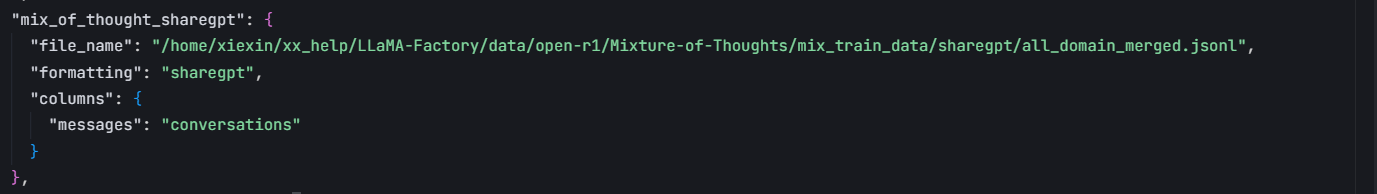
（1）Instruct就是alpaca格式，我描述的有差异

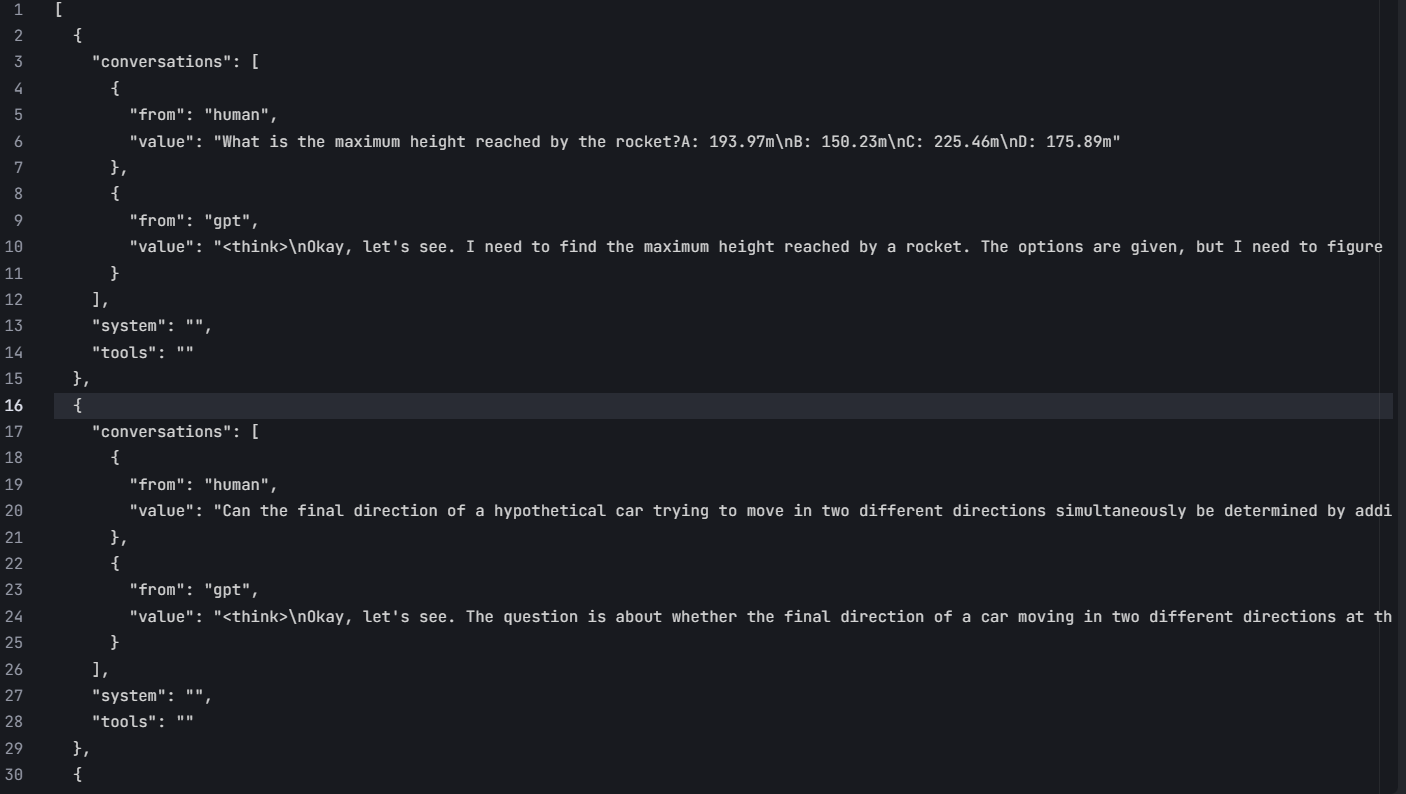
1. 正确的jsonl格式是指标准的jsonl数据，按行存储对象。原生的Llamafactory是能够支持处理json和jsonl的数据格式的，json和jsonl文件的区别在于如果开启了streaming使用流式处理，要使用jsonl格式的数据才能够正确的加载。



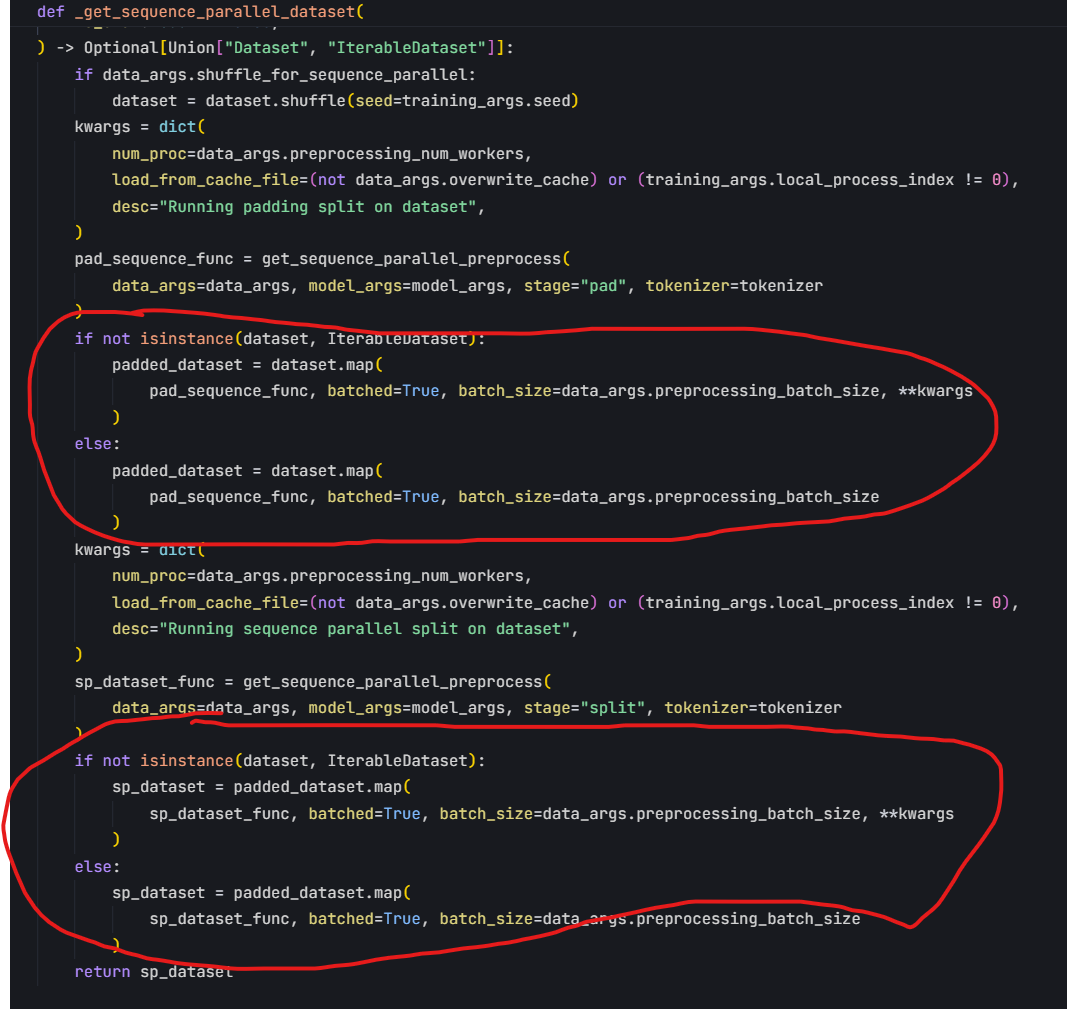
从json文件变成jsonl文件，只需要在datainfo中只需要把file\_name改变即可。



错误的格式是指我之前用的是json数据的格式，用多行来存储一个json对象，只是文件命名的后缀是jsonl。Json对象中system和tools可有可无，如果要添加这两个属性，在data\_info中columns中添加这两个属性即可。



1. 改loader只是interableDataset加载数据上的问题，360-llamafactory没考虑到开启streaming设置时使用interableDataset的加载的问题。会对streaming开启时加载的数据有影响。我依据终端报错信息，更改了LLaMA-Factory/src/llamafactory/data/loader.py文件中的函数\_get\_sequence\_parallel\_dataset，其他src目录下的内容没有做任何改动



1. 如果是使用了packing参数，cutoff\_len要在原基础上加1，是360-llamafactory的代码逻辑，不加1会有对应的终端报错提示，这个很容易发现
2. 验证了分别使用Instruct和sharegpt格式的4K数据进行模型训练时的显存占用，完全相同。是在处理完数据，模型刚进入训练时检查的显存占用
3. 上传的新文件ParquetConvertInstruct.py用于转换mix-of-thought为alpaca或者sharegpt文件，MergeJson用于合并多个json（jsonl）为一个json(jsonl)文件。