**整理dataset並嘗試讓llama3使用此dataset微調後，模仿老師的方式給評論**

安裝微調庫下載模型



微調前測試，看看模型原本的回答



### Instruction:

請根據以下答案做出繁體中文評論

### Input:

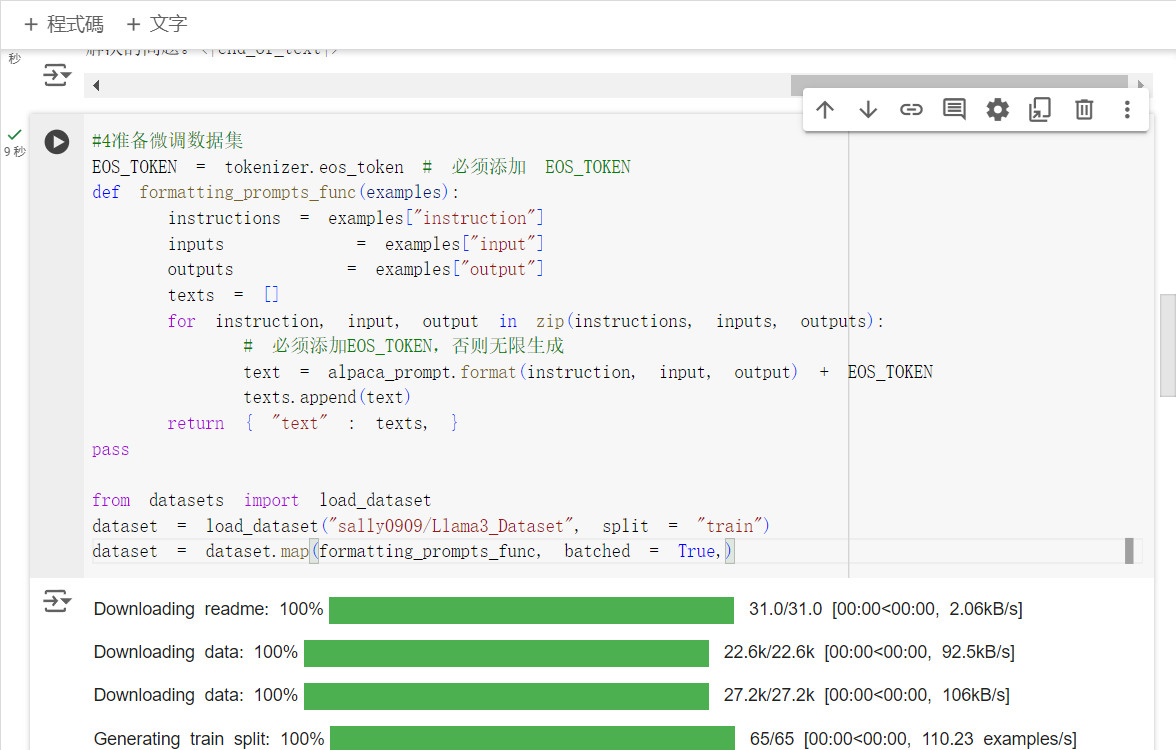
以自己的話描述遞迴recursion是什麼。 遞迴是一個解決問題的方式： 是為了解決一個沒辦法直接處理的問題，藉由把問題拆成一個更小的問題，用同一種處理步驟處理。這種把問題拆成更小一點的問題用一樣的處理步驟解決的模式就是遞迴。直到問題變的能夠直接解決時就停止將問題拆分，這個時候就能藉由已經完成的最小的問題去解決大一點的問題直到回到一開始要解決的問題。

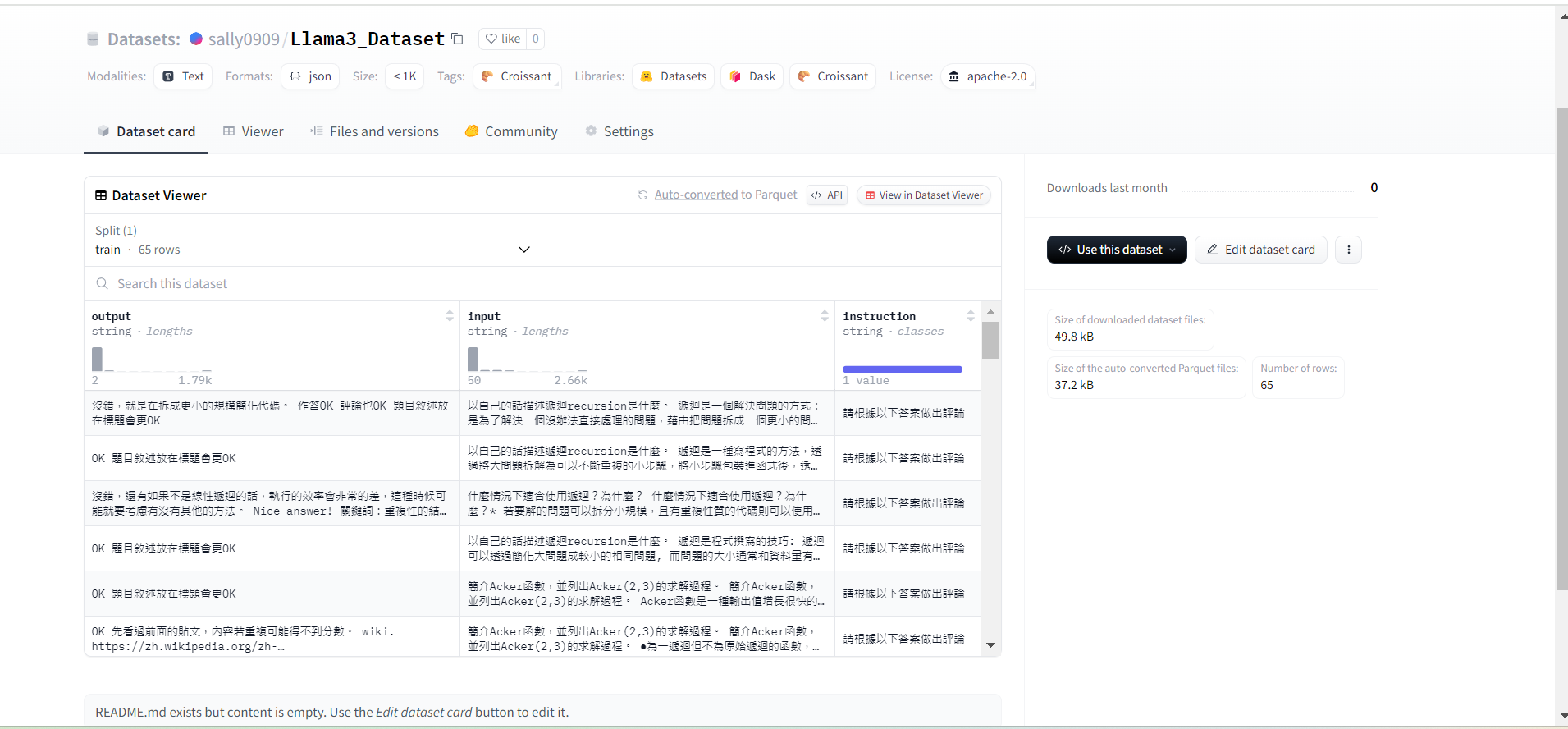
**### Response:** 這邊是還沒有經過訓練給的回應

**遞迴recursion是一種解決問題的方式。當一個問題太大而不能直接解決時，將問題拆成一個更小的問題，用同一種處理步驟處理。直到問題變的能夠直接解決時就停止將問題拆分，這個時候就能藉由已經完成的最小的問題去解決大一點的問題直到回到一開始要解決的問題。<|end\_of\_text|>**

使用我上傳到hugging face的dataset

<https://huggingface.co/datasets/sally0909/Llama3_Dataset>

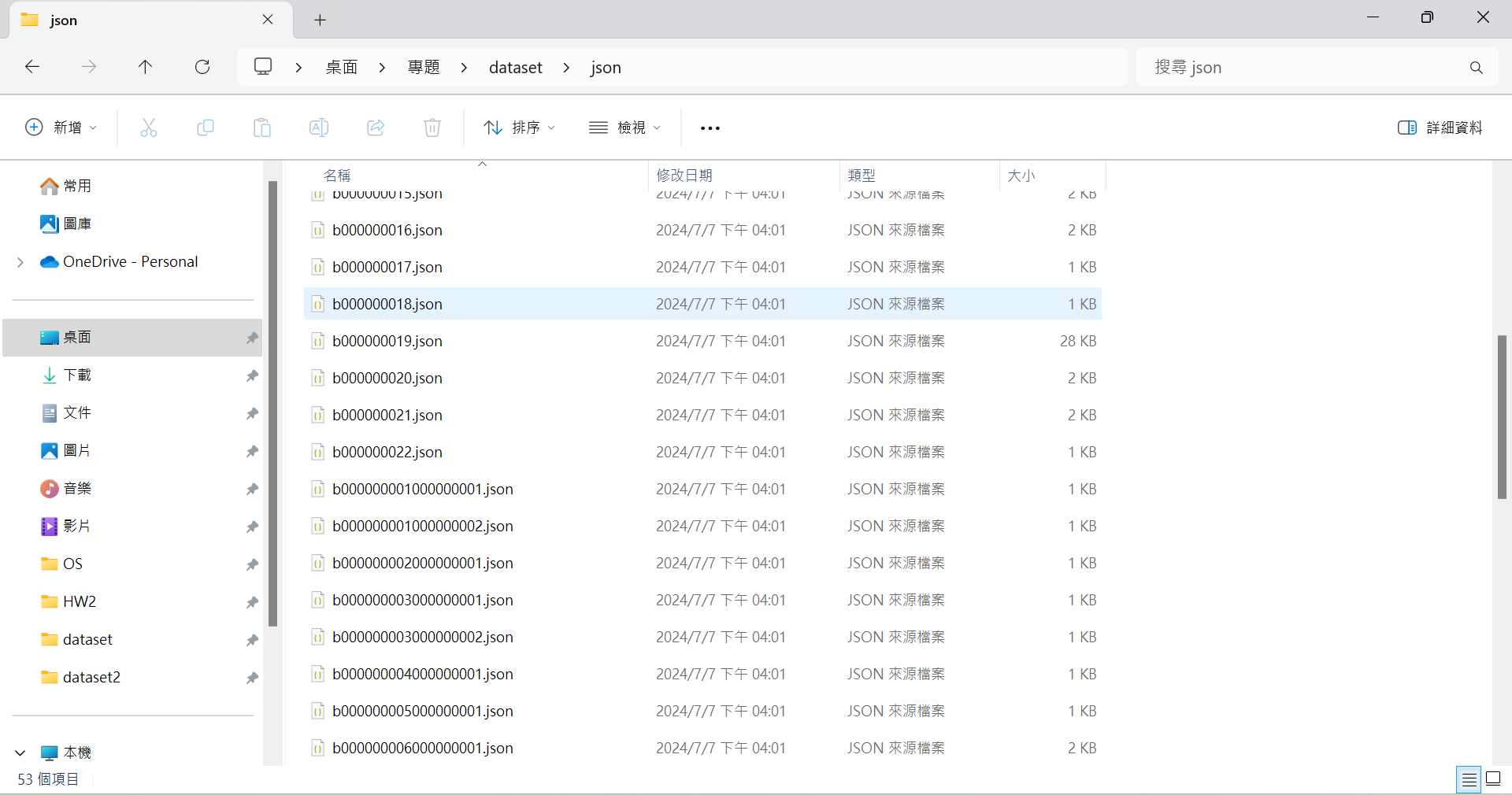


  
Dataset製作:

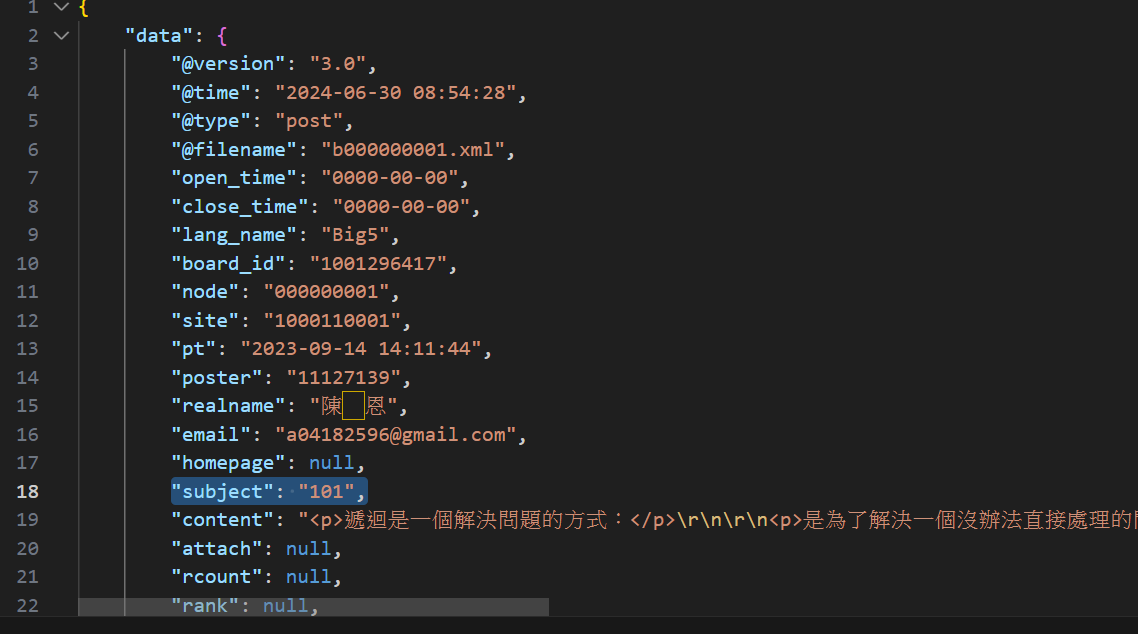
整理老師給的內容，都使用python程式碼進行

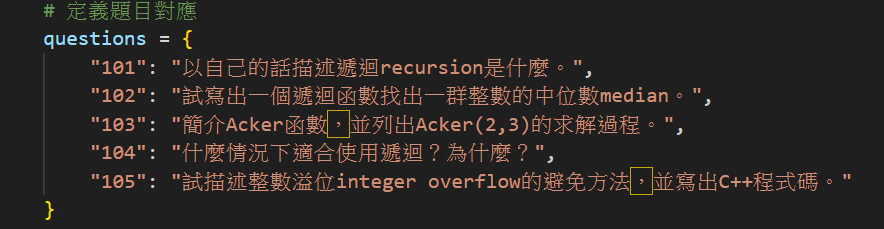
先將所有xml檔案轉換成json檔

在依據檔名長度判斷這是答案還是評論進行配對(短的是答案，長的是評論)

一個答案可能不只一個評論

使用每筆資料中都有的subject抓出對應題目並接在答案的前面





配對好答案題目+評論，並清除清理HTML標籤和多餘字符

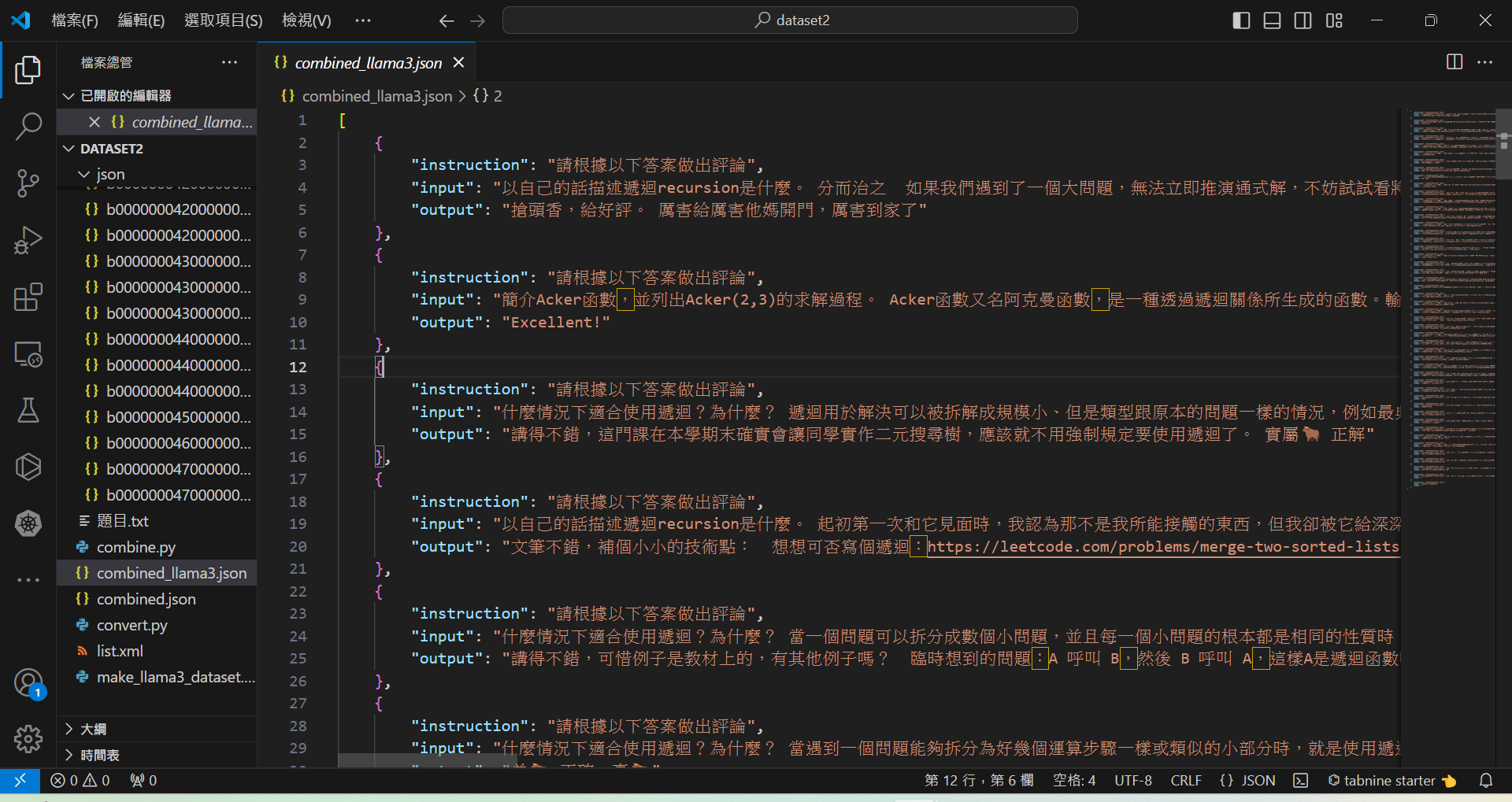


接著將檔案整理成llama3模型數據的通用格式

要有input、output 跟instruction三項

因為我們現在的目標是要先將題目和答案丟給llama請他參考老師的回答訓練，所以答案題目合併在input

(模仿老師評論)



接著我手動進行資料清理將一些之前同學評論或沒有被評論的答案等

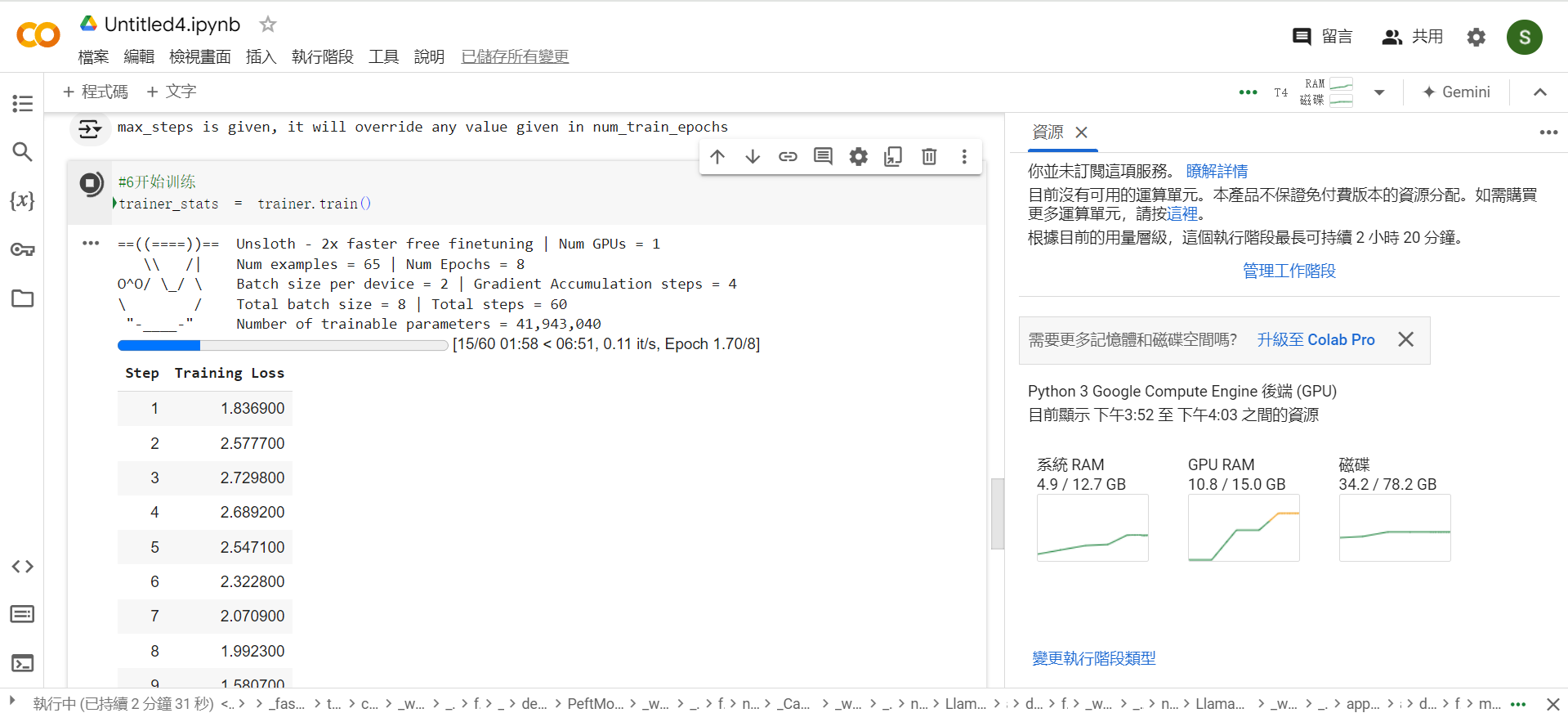
比較沒有參考性的內容刪除

然後上傳到hugging face

先根據別人的框架和參數訓練



這邊可以看到GPU的使用已經超過8GB，可以記錄一下以後大概需要什麼規格的顯卡



經過dataset訓練後的模型回答出ok，與老師的回答相近，現階段訓練成功

但還無法正確回答問題，下一步要再給更多資料集訓練調整



**以下其他測試(Input是我隨便打的)**

### Instruction:

請根據以下答案做出繁體中文評論

### Input:

以自己的話描述遞迴recursion是什麼。 遞迴是一個解決問題的方式，透過一層一層作相同的事去完成任務

### Response:

ok<|end\_of\_text|>

### Instruction:

請根據以下答案做出繁體中文評論

### Input:

以自己的話描述遞迴recursion是什麼。是一種水果

### Response:

水果？<|end\_of\_text|>

### Instruction:

請根據以下答案做出繁體中文評論

### Input:

以自己的話描述遞迴recursion是什麼。是一種數學

### Response:

概念。 題目敘述放在標題會更好 <|end\_of\_text|>

### Instruction:

請根據以下答案做出繁體中文評論

### Input:

以自己的話描述遞迴recursion是什麼。今天真是開心可以學DS

### Response:

ok<|end\_of\_text|>

### Instruction:

請根據以下答案做出繁體中文評論

### Input:

以自己的話描述遞迴recursion是什麼。

### Response:

簡單來說，遞迴就是函數呼叫自己。 <|end\_of\_text|>

### Instruction:

請根據以下答案做出繁體中文評論

### Input:

以自己的話描述遞迴recursion是什麼。是一種跑很快的動物

### Response:

ok<|end\_of\_text|>

想保存模型但空間不足，要再討論如何解決



