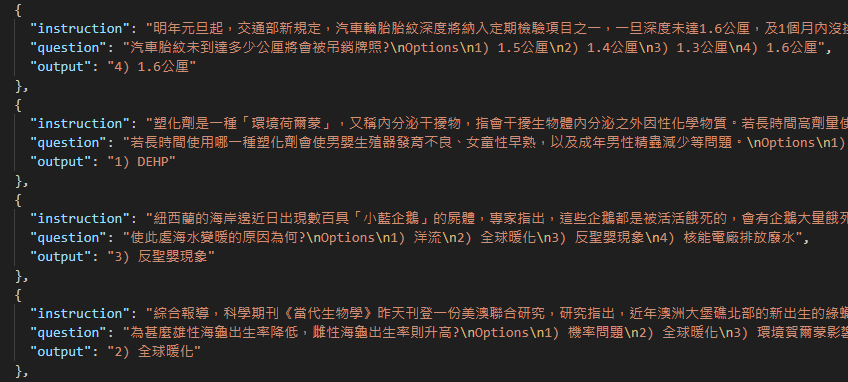
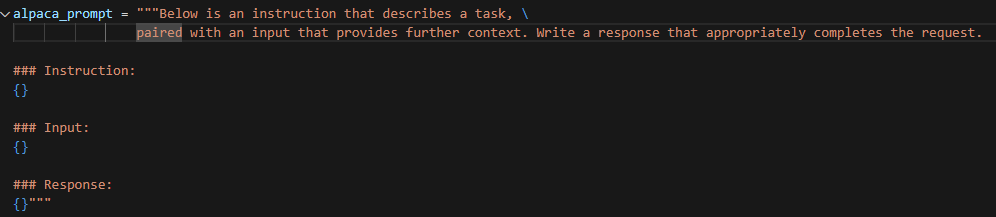
**ZH-TW Reading Comprehension Test for LLMs**

1. **Create DataSets**

因為使用Alpaca\_llama3的ipynb做finetune，因此先將原始數據集整理成json格式方便後續訓練時配合讀取成prompt格式。

DataSets json:



Prompt(可以自行調整):****

1. **Training**

* 參考連結: <https://colab.research.google.com/drive/135ced7oHytdxu3N2DNe1Z0kqjyYIkDXp?usp=sharing#scrollTo=e2pEuRb1r2Vg>

分別使用不同的模型進行訓練，分別有

1. "unsloth/mistral-7b-v0.3-bnb-4bit"
2. “yentinglin/Llama-3-Taiwan-8B-Instruct”
3. “yentinglin/Taiwan-LLM-7B-v2.0-base”

使用usloth library:

可以更高效率的讀取模型；並寫更高效進行模型訓練。

可以在維持不錯生成回答精確度的同時降低記憶體在訓練時的使用量，目前測試能夠限縮到使用 6GB

1. mistral-7b-v0.3-bnb-4bit

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Checkpoint-60 | Checkpoint-500 | Checkpoint-1000 |
|  |  |  |
| Checkpoint-1500 | Checkpoint-2000 |  |

1. Llama-3-Taiwan-8B-Instruct

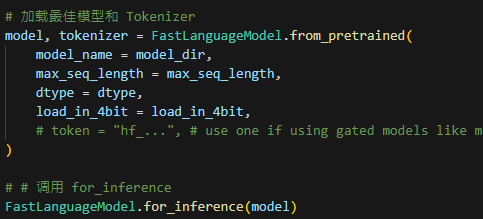
|  |  |  |
| --- | --- | --- |
|  |  |  |
| Checkpoint-60 | Checkpoint-500 | Checkpoint-1000 |
|  |  |  |
| Checkpoint-1500 | Checkpoint-2000 |  |

1. Taiwan-LLM-7B-v2.0-base

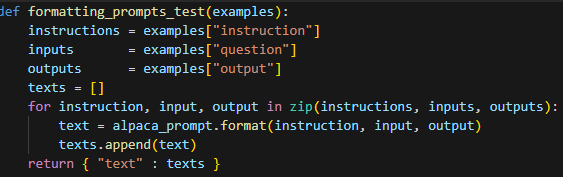
|  |  |
| --- | --- |
|  |  |
| Checkpoint-500 | Checkpoint-1000 |

1. **Inference**

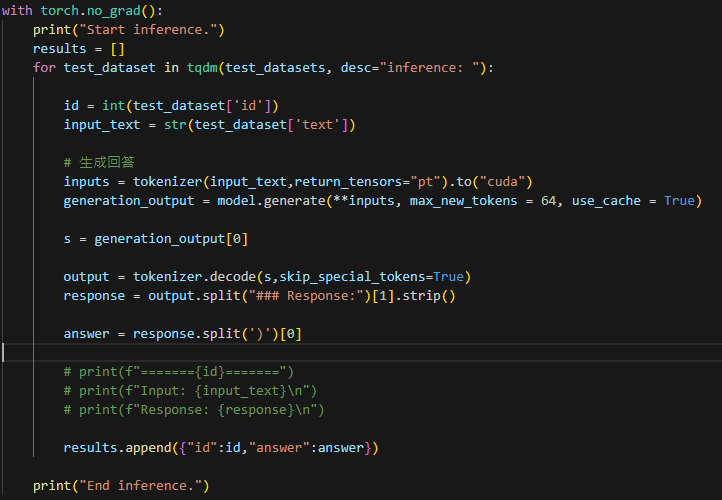
使用usloth載入模型:

****

用與train相同的prompt進行test的讀取。



最後按照平常的格式進行文字生成



P.S

機器學習:

【分類問題】: 需要label正確答案，作為參照。

【文字生成】: 不需要label。需要將所有文字文章、問題、選擇1~4、答案，利用自行設定的prompt串接成一個text，作為input。Encode/decode都使用tokenizer，不過Encode部分可以包還在SFTTrainer、Trainer訓練器做轉換，不須在資料前處理的時候就先手動轉換Encode。避免發生不必要的轉換錯誤。