自然語言處理與應用
Natural Language Processing and Applications

Project Introduction

Instructor: 林英嘉 (Ying-Jia Lin)
2025/04/07

Course GitHub

Slido # NLP_0407

Outline

- Recap: Language Generation
- Decoding Strategies
    - Greedy Decoding
    - Beam Search
    - Top-k / Top-p Sampling
- Evaluations

What is Kaggle?

- Kaggle is a platform that provides:
    - Real-word datasets for machine learning
    - Competitions with prizes (sometimes with money)
    - Discussion forum with a lots of code examples

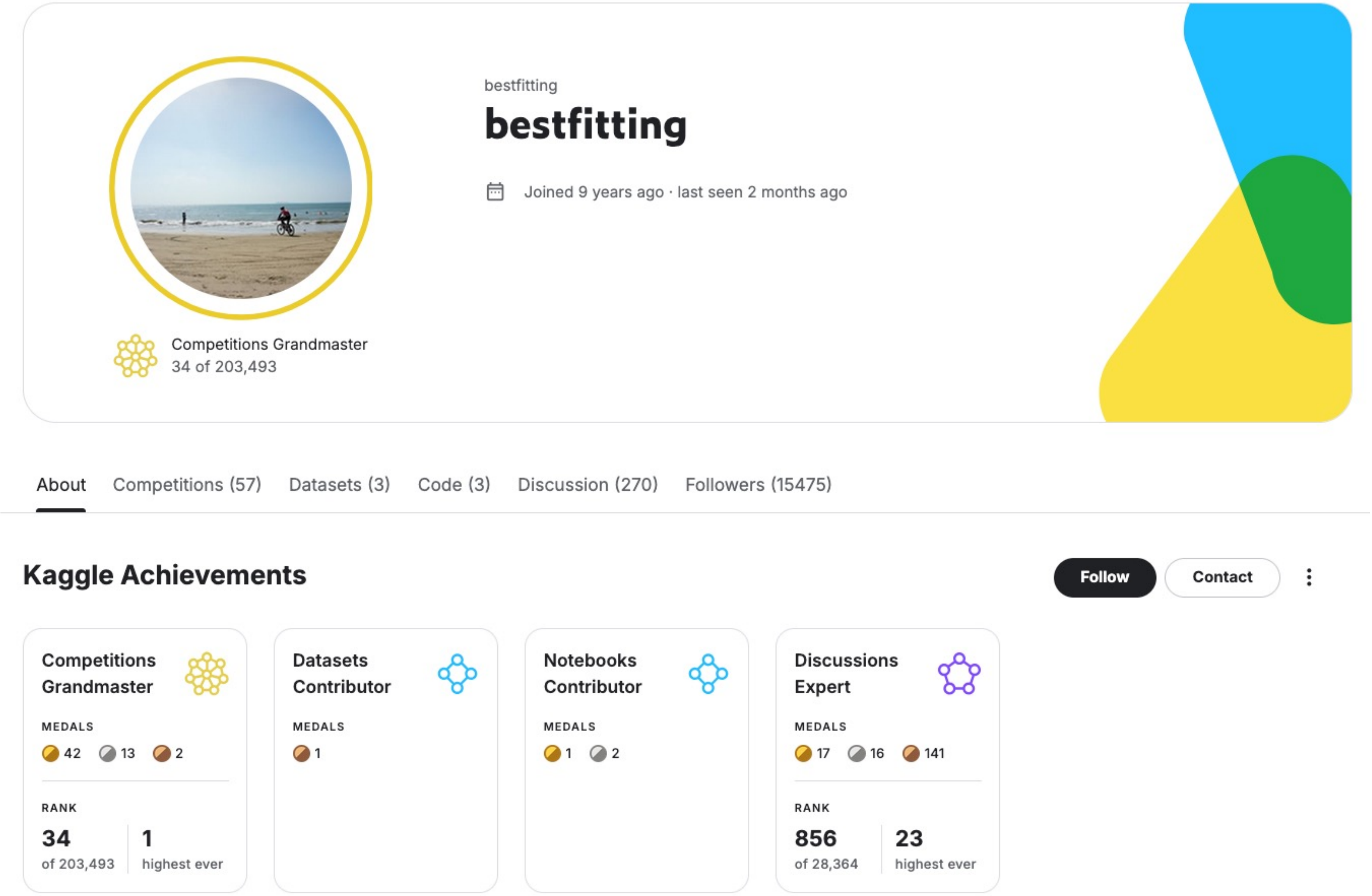
Kaggle submission types

- Traditional competitions
  - Upload **submission file** (e.g., *.csv)
- Code / Notebook competitions

本課程 projects 採用此方式

  - Upload **code** (e.g., *.ipynb)

If you play Kaggle a lot …



bestfitting

bestfitting

Joined 9 years ago · last seen 2 months ago

Competitions Grandmaster
34 of 203,493

Kaggle Achievements

Competitions Grandmaster
MEDALS
🥇 42 🥈 13 🥉 2
RANK
34 of 203,493 | 1 highest ever

Datasets Contributor
MEDALS
🥉 1

Notebooks Contributor
MEDALS
🥇 1 🥈 2

Discussions Expert
MEDALS
🥇 17 🥈 16 🥉 141
RANK
856 of 28,364 | 23 highest ever

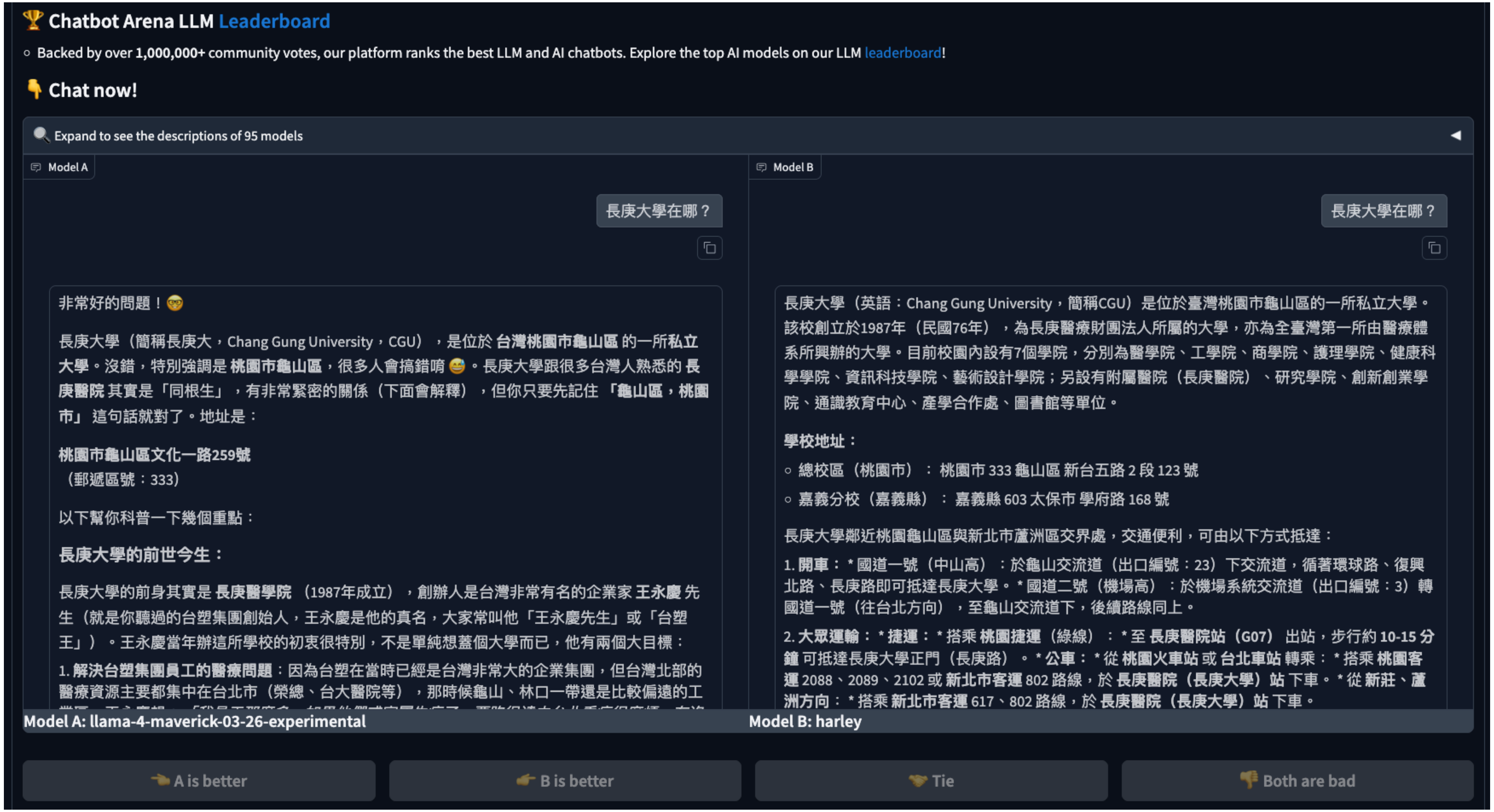
Figure source: https://www.kaggle.com/bestfitting

Outline of tasks

Platform	Competition Name	Data Type	Task Type
Kaggle	LLM Classification Finetuning	text	Text classification
	NBME - Score Clinical Patient Notes	text	Token classification (like Named Entity Recognition)
	LLM - Detect AI Generated Text	text	Text classification

LLM Classification Finetuning (intro)

Chatbot Arena: https://lmarena.ai/?leaderboard



🏆 **Chatbot Arena LLM** Leaderboard

○ Backed by over **1,000,000+** community votes, our platform ranks the best LLM and AI chatbots. Explore the top AI models on our LLM leaderboard!

👇 **Chat now!**

🔍 Expand to see the descriptions of 95 models

長庚大學在哪？

非常好的問題！🤓

長庚大學（簡稱長庚大，Chang Gung University，CGU），是位於 **台灣桃園市龜山區** 的一所**私立大學**。沒錯，特別強調是 **桃園市龜山區**，很多人會搞錯唷 😅。長庚大學跟很多台灣人熟悉的 **長庚醫院** 其實是「同根生」，有非常緊密的關係（下面會解釋），但你只要先記住 **「龜山區，桃園市」** 這句話就對了。地址是：

桃園市龜山區文化一路259號
（郵遞區號：333）

以下幫你科普一下幾個重點：

長庚大學的前世今生：

長庚大學的前身其實是 **長庚醫學院** （1987年成立），創辦人是台灣非常有名的企業家 **王永慶** 先生（就是你聽過的台塑集團創始人，王永慶是他的真名，大家常叫他「王永慶先生」或「台塑王」）。王永慶當年辦這所學校的初衷很特別，不是單純想蓋個大學而已，他有兩個大目標：

1. **解決台塑集團員工的醫療問題**：因為台塑在當時已經是台灣非常大的企業集團，但台灣北部的醫療資源主要都集中在台北市（榮總、台大醫院等），那時候龜山、林口一帶還是比較偏遠的工

長庚大學在哪？

長庚大學（英語：Chang Gung University，簡稱CGU）是位於臺灣桃園市龜山區的一所私立大學。該校創立於1987年（民國76年），為長庚醫療財團法人所屬的大學，亦為全臺灣第一所由醫療體系所興辦的大學。目前校園內設有7個學院，分別為醫學院、工學院、商學院、護理學院、健康科學學院、資訊科技學院、藝術設計學院；另設有附屬醫院（長庚醫院）、研究學院、創新創業學院、通識教育中心、產學合作處、圖書館等單位。

學校地址：

- 總校區（桃園市）： 桃園市 333 龜山區 新台五路 2 段 123 號
- 嘉義分校（嘉義縣）： 嘉義縣 603 太保市 學府路 168 號

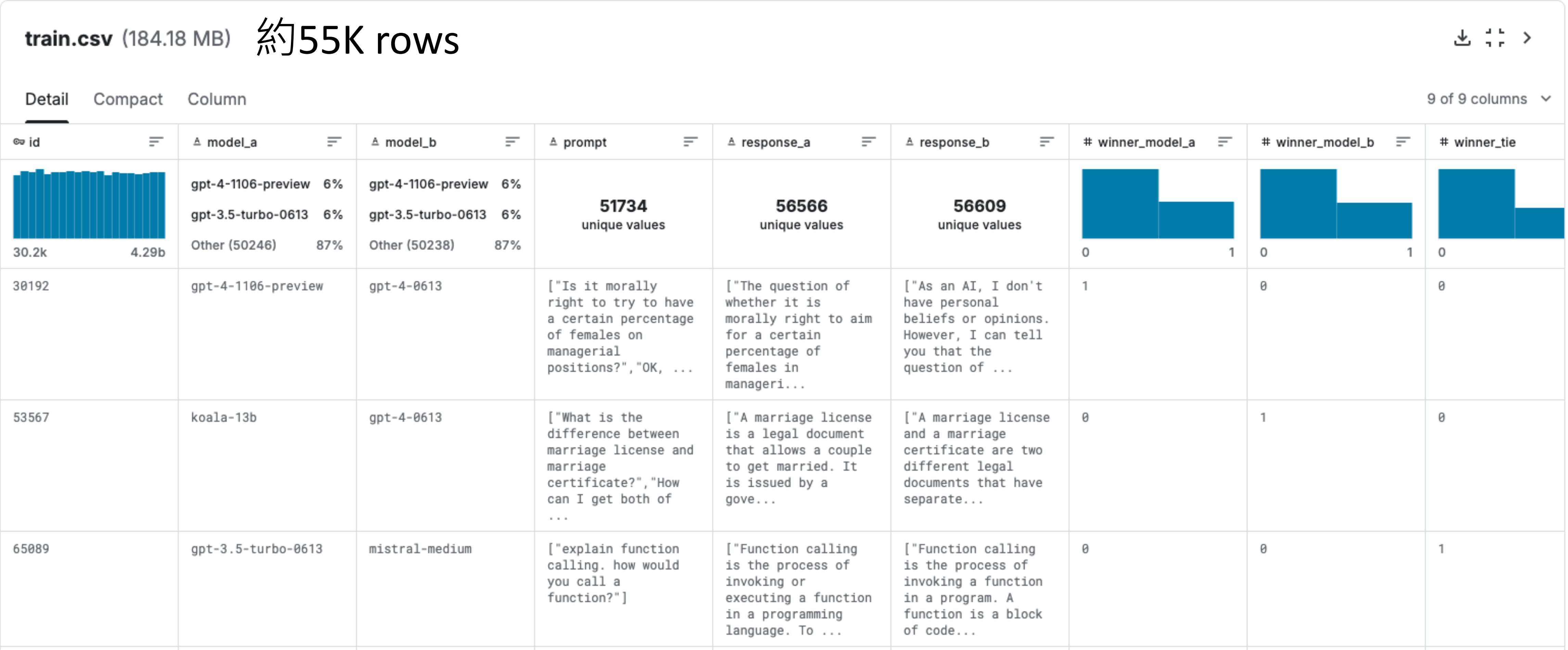
長庚大學鄰近桃園龜山區與新北市蘆洲區交界處，交通便利，可由以下方式抵達：

1. **開車**： * 國道一號（中山高）：於龜山交流道（出口編號：23）下交流道，循著環球路、復興北路、長庚路即可抵達長庚大學。 * 國道二號（機場高）：於機場系統交流道（出口編號：3）轉國道一號（往台北方向），至龜山交流道下，後續路線同上。
2. **大眾運輸**： * **捷運**： * 搭乘 **桃園捷運**（綠線）： * 至 **長庚醫院站**（G07） 出站，步行約 10-15 分鐘 可抵達長庚大學正門（長庚路）。 * **公車**： * 從 **桃園火車站** 或 **台北車站** 轉乘： * 搭乘 **桃園客運** 2088、2089、2102 或 **新北市客運** 802 路線，於 **長庚醫院（長庚大學）站** 下車。 * 從 **新莊、蘆洲方向**： * 搭乘 **新北市客運** 617、802 路線，於 **長庚醫院（長庚大學）站** 下車。

Model A: llama-4-maverick-03-26-experimental

Model B: harley

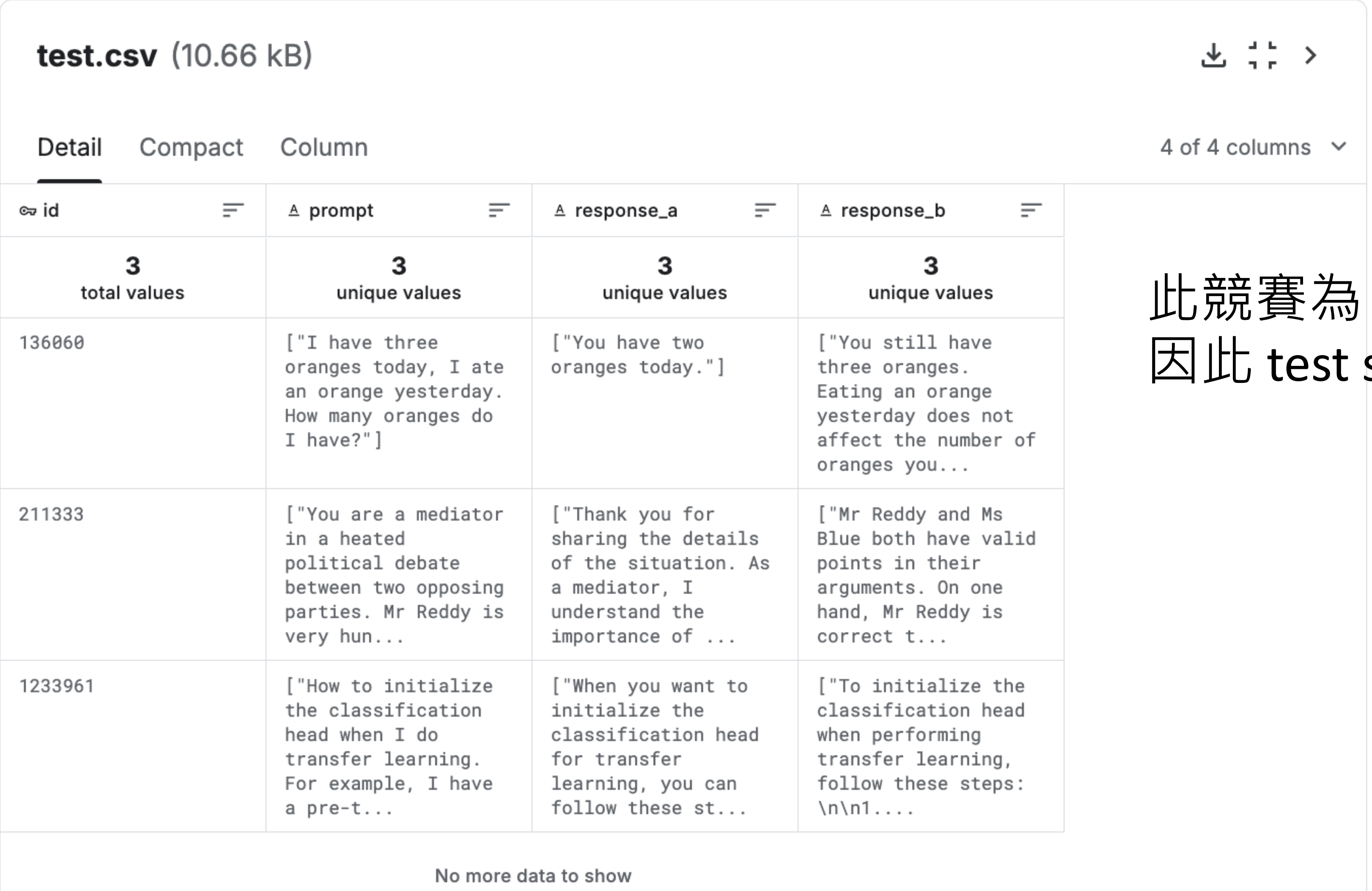
LLM Classification Finetuning (data example)



train.csv (184.18 MB) 約55K rows

id	model_a	model_b	prompt	response_a	response_b	winner_model_a	winner_model_b	winner_tie
30.2k 4.29b	gpt-4-1106-preview 6% gpt-3.5-turbo-0613 6% Other (50246) 87%	gpt-4-1106-preview 6% gpt-3.5-turbo-0613 6% Other (50238) 87%	51734 unique values	56566 unique values	56609 unique values	0 1	0 1	0
30192	gpt-4-1106-preview	gpt-4-0613	["Is it morally right to try to have a certain percentage of females on managerial positions?","OK, ...	["The question of whether it is morally right to aim for a certain percentage of females in manageri...	["As an AI, I don't have personal beliefs or opinions. However, I can tell you that the question of ...	1	0	0
53567	koala-13b	gpt-4-0613	["What is the difference between marriage license and marriage certificate?","How can I get both of ...	["A marriage license is a legal document that allows a couple to get married. It is issued by a gove...	["A marriage license and a marriage certificate are two different legal documents that have separate...	0	1	0
65089	gpt-3.5-turbo-0613	mistral-medium	["explain function calling. how would you call a function?"]	["Function calling is the process of invoking or executing a function in a programming language. To ...	["Function calling is the process of invoking a function in a program. A function is a block of code...	0	0	1

LLM Classification Finetuning (data example)



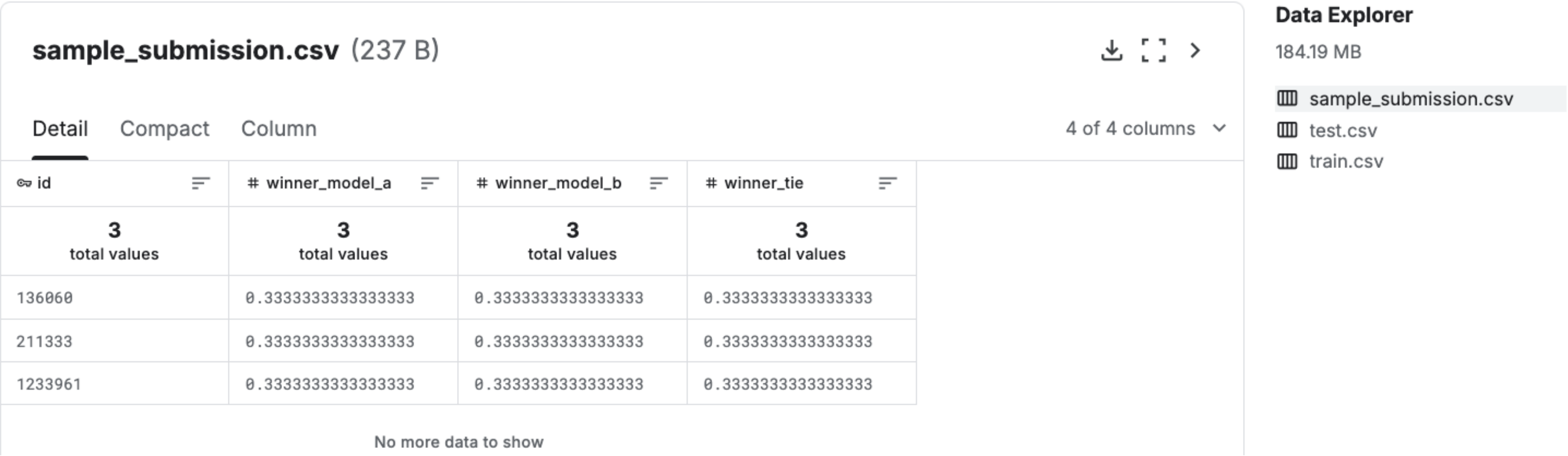
id	prompt	response_a	response_b
3 total values	3 unique values	3 unique values	3 unique values
136060	["I have three oranges today, I ate an orange yesterday. How many oranges do I have?"]	["You have two oranges today."]	["You still have three oranges. Eating an orange yesterday does not affect the number of oranges you...
211333	["You are a mediator in a heated political debate between two opposing parties. Mr Reddy is very hun...	["Thank you for sharing the details of the situation. As a mediator, I understand the importance of ...	["Mr Reddy and Ms Blue both have valid points in their arguments. On one hand, Mr Reddy is correct t...
1233961	["How to initialize the classification head when I do transfer learning. For example, I have a pre-t...	["When you want to initialize the classification head for transfer learning, you can follow these st...	["To initialize the classification head when performing transfer learning, follow these steps: \n\n1....

No more data to show

此競賽為 Code competition
因此 test set 實際上不公開

LLM Classification Finetuning (submission)

- 上傳到 Kaggle Leaderboard 需要實作 .ipynb
- 你的 code 需要能 predict test.csv 並產生檔案名稱為 submission.csv 的檔案，Kaggle 才能幫你執行程式碼並打分數，可參考 LMSYS: KerasNLP Starter



sample_submission.csv (237 B)

id	# winner_model_a	# winner_model_b	# winner_tie
3 total values	**3** total values	**3** total values	**3** total values
136060	0.3333333333333333	0.3333333333333333	0.3333333333333333
211333	0.3333333333333333	0.3333333333333333	0.3333333333333333
1233961	0.3333333333333333	0.3333333333333333	0.3333333333333333

No more data to show

Data Explorer

184.19 MB

- sample_submission.csv
- test.csv
- train.csv

LLM Classification Finetuning (evaluation)

- Binary cross-entropy，對於任一個 class 而言 (win_a, win_b, tie):

$$L_{\log}(y,p) = -\big(y\log(p) + (1-y)\log(1-p)\big)$$

p: 預測該類別為1的機率
y: 正確答案的類別

- Leaderboard 比的是 test set的 Loss 值 (越小越好)

https://www.kaggle.com/code/metric/log-loss

NBME - Score Clinical Patient Notes

- 找出 patient note 中的重要特徵，例如：
  - 輸入一篇 patient note，輸出為 “diminished appetite”
- 競賽的實際範例：

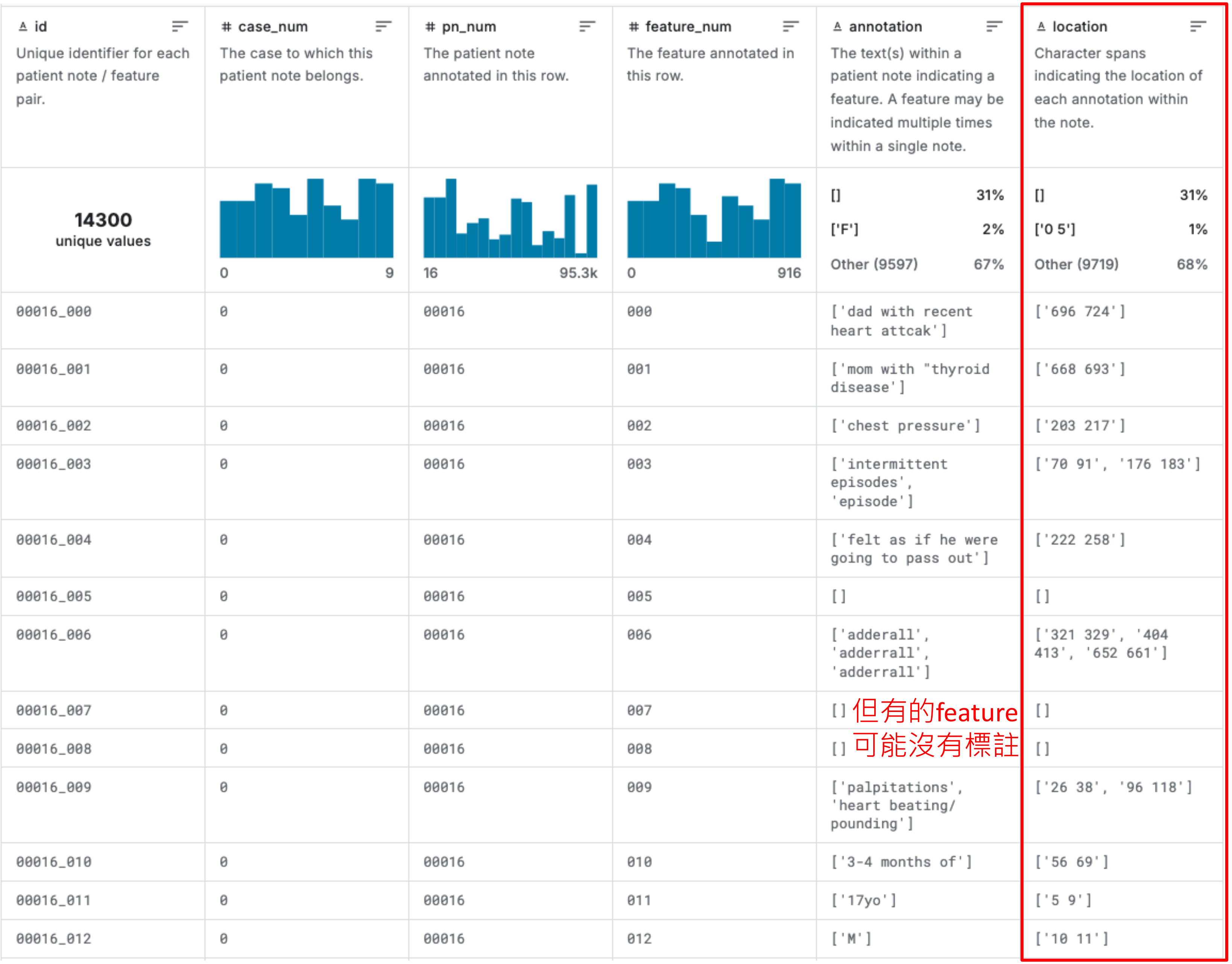
```
>>> import pandas as pd
>>> df = pd.read_csv("patient_notes.csv")
>>> df[df["pn_num"]==16].pn_history.values[0][696:724]
'dad with recent heart attcak'
```

NBME – patient note example

```
>>> df[df["pn_num"]==16].pn_history.values[0]
```

```
HPI: 17yo M presents with palpitations. Patient reports 3-4 months of
intermittent episodes of "heart beating/pounding out of my chest." 2
days ago during a soccer game had an episode, but this time had chest
pressure and felt as if he were going to pass out (did not lose
conciousness). Of note patient endorses abusing adderall, primarily to
study (1-3 times per week). Before recent soccer game, took adderrall
night before and morning of game. Denies shortness of breath,
diaphoresis, fevers, chills, headache, fatigue, changes in sleep,
changes in vision/hearing, abdominal paun, changes in bowel or urinary
habits. \r\nPMHx: none\r\nRx: uses friends adderrall\r\nFHx: mom with
"thyroid disease," dad with recent heart attcak\r\nAll:
none\r\nImmunizations: up to date\r\nSHx: Freshmen in college.
Endorses 3-4 drinks 3 nights / week (on weekends), denies tabacco,
endorses trying marijuana. Sexually active with girlfriend x 1 year,
uses condoms
```

NBME - Score Clinical Patient Notes (data example - train.csv)

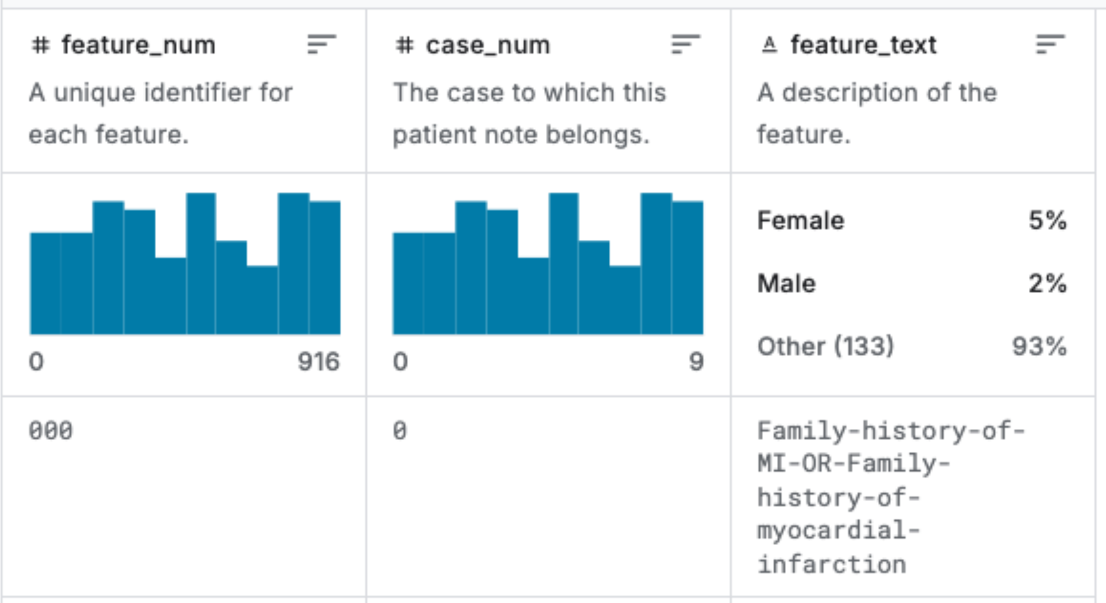


id	case_num	pn_num	feature_num	annotation	location
Unique identifier for each patient note / feature pair.	The case to which this patient note belongs.	The patient note annotated in this row.	The feature annotated in this row.	The text(s) within a patient note indicating a feature. A feature may be indicated multiple times within a single note.	Character spans indicating the location of each annotation within the note.
14300 unique values	0 – 9	16 – 95.3k	0 – 916	[] 31% ['F'] 2% Other (9597) 67%	[] 31% ['0 5'] 1% Other (9719) 68%
00016_000	0	00016	000	['dad with recent heart attcak']	['696 724']
00016_001	0	00016	001	['mom with "thyroid disease']	['668 693']
00016_002	0	00016	002	['chest pressure']	['203 217']
00016_003	0	00016	003	['intermittent episodes', 'episode']	['70 91', '176 183']
00016_004	0	00016	004	['felt as if he were going to pass out']	['222 258']
00016_005	0	00016	005	[]	[]
00016_006	0	00016	006	['adderall', 'adderall', 'adderall']	['321 329', '404 413', '652 661']
00016_007	0	00016	007	[]	[]
00016_008	0	00016	008	[]	[]
00016_009	0	00016	009	['palpitations', 'heart beating/ pounding']	['26 38', '96 118']
00016_010	0	00016	010	['3-4 months of']	['56 69']
00016_011	0	00016	011	['17yo']	['5 9']
00016_012	0	00016	012	['M']	['10 11']

但有的feature
可能沒有標註

每筆資料都有
feature_num代號

NBME - Score Clinical Patient Notes (data example - features.csv)

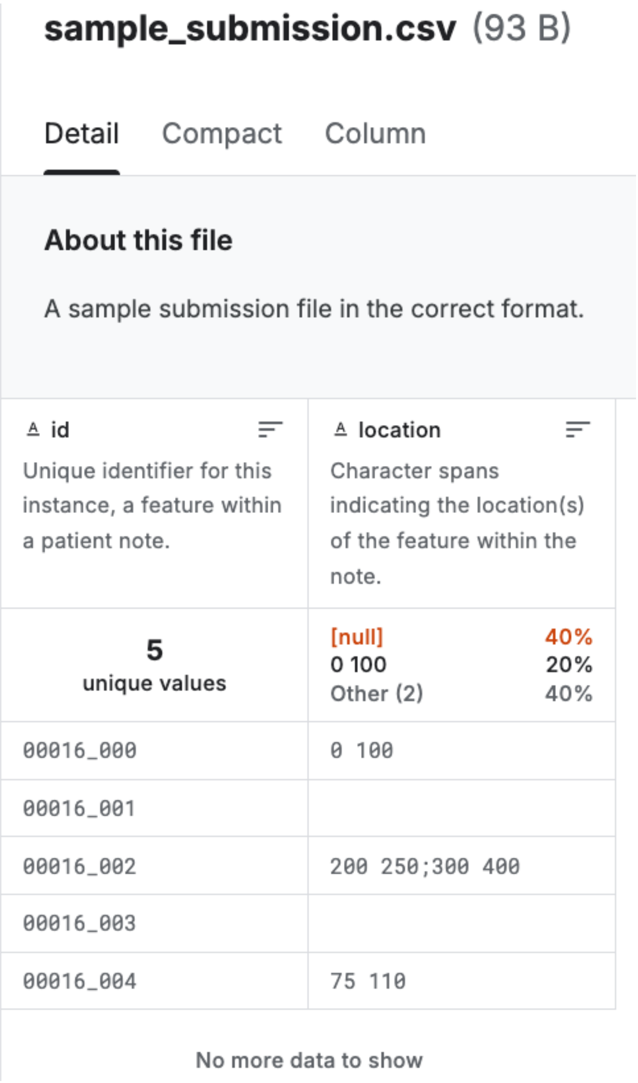


# feature_num	# case_num	feature_text
A unique identifier for each feature.	The case to which this patient note belongs.	A description of the feature.
0 — 916	0 — 9	Female 5% Male 2% Other (133) 93%
000	0	Family-history-of-MI-OR-Family-history-of-myocardial-infarction

- 競賽的實際範例：

```
>>> import pandas as pd
>>> df = pd.read_csv("patient_notes.csv")
>>> df[df["pn_num"]==16].pn_history.values[0][696:724]
'dad with recent heart attcak'
```

NBME - Score Clinical Patient Notes (submission and evaluation)



sample_submission.csv (93 B)

About this file

A sample submission file in the correct format.

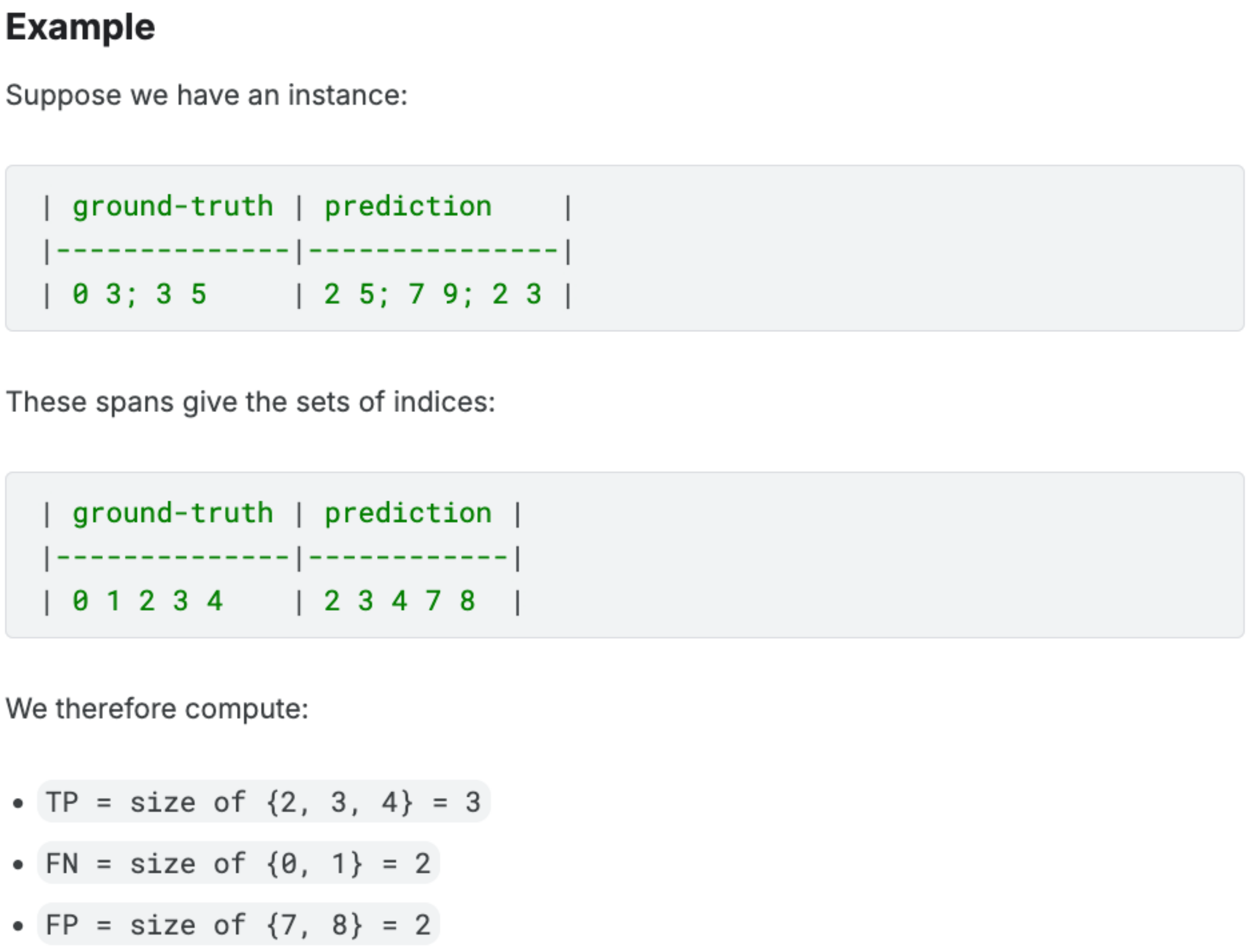
id	location
Unique identifier for this instance, a feature within a patient note.	Character spans indicating the location(s) of the feature within the note.
5 unique values	[null] 40% 0 100 20% Other (2) 40%
00016_000	0 100
00016_001	
00016_002	200 250;300 400
00016_003	
00016_004	75 110

No more data to show

- 只需要預測 location 即可
- 上傳到 Kaggle Leaderboard 需要實作 .ipynb
- 你的 code 需要能 predict test.csv 並產生檔案名稱為 submission.csv 的檔案，Kaggle 才能幫你執行程式碼並打分數，可參考 [NBME / Deberta-base baseline [inference]]

NBME - Score Clinical Patient Notes (evaluation)

- Metric: Micro F1-score



Example

Suppose we have an instance:

```
| ground-truth | prediction    |
|--------------|---------------|
| 0 3; 3 5     | 2 5; 7 9; 2 3 |
```

These spans give the sets of indices:

```
| ground-truth | prediction |
|--------------|------------|
| 0 1 2 3 4    | 2 3 4 7 8  |
```

We therefore compute:

- `TP = size of {2, 3, 4} = 3`
- `FN = size of {0, 1} = 2`
- `FP = size of {7, 8} = 2`

Confusion Matrix

	Actually positive	**Actually negative**
Predicted positive	True positive (TP)	False positive (FP)
Predicted negative	False negative (FN)	True negative (TN)

- Precision = TP / (TP + FP)
  - 模型預測的TP比例
- Recall = TP / (TP + FN)
  - True Positive Rate (TPR)
- F1-score (廣義) = 2(Precision*Recall) / (Precision+Recall)

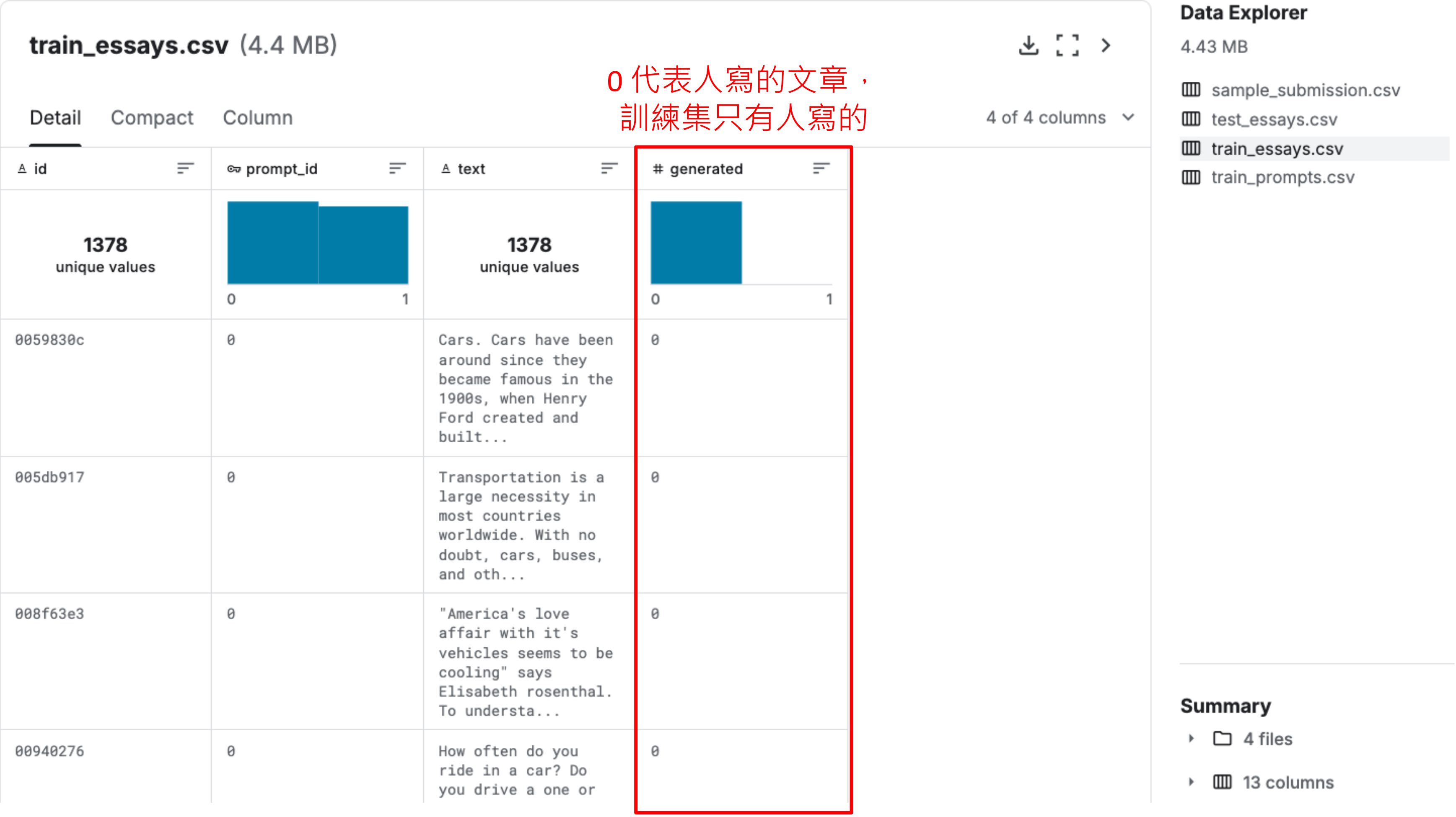
Macro vs. Micro

- 假設總共有100筆資料：
    - 每一筆都算出一個 F1-score，最後取平均 => Macro F1-score
    - 加總全部100筆的 TP, FN 以及 FP 之後，再算出 F1-score => Micro F1-score

LLM - Detect AI Generated Text

- The competition dataset comprises about 10,000 essays
    - Some written by students and some generated by a variety of large language models (LLMs).
- The goal of the competition is to determine whether or not essay was generated by an LLM.

LLM - Detect AI Generated Text (data)



train_essays.csv (4.4 MB)

0 代表人寫的文章，
訓練集只有人寫的

id	prompt_id	text	generated
1378 unique values	0 1	1378 unique values	0 1
0059830c	0	Cars. Cars have been around since they became famous in the 1900s, when Henry Ford created and built...	0
005db917	0	Transportation is a large necessity in most countries worldwide. With no doubt, cars, buses, and oth...	0
008f63e3	0	"America's love affair with it's vehicles seems to be cooling" says Elisabeth rosenthal. To understa...	0
00940276	0	How often do you ride in a car? Do you drive a one or	0

Data Explorer

4.43 MB

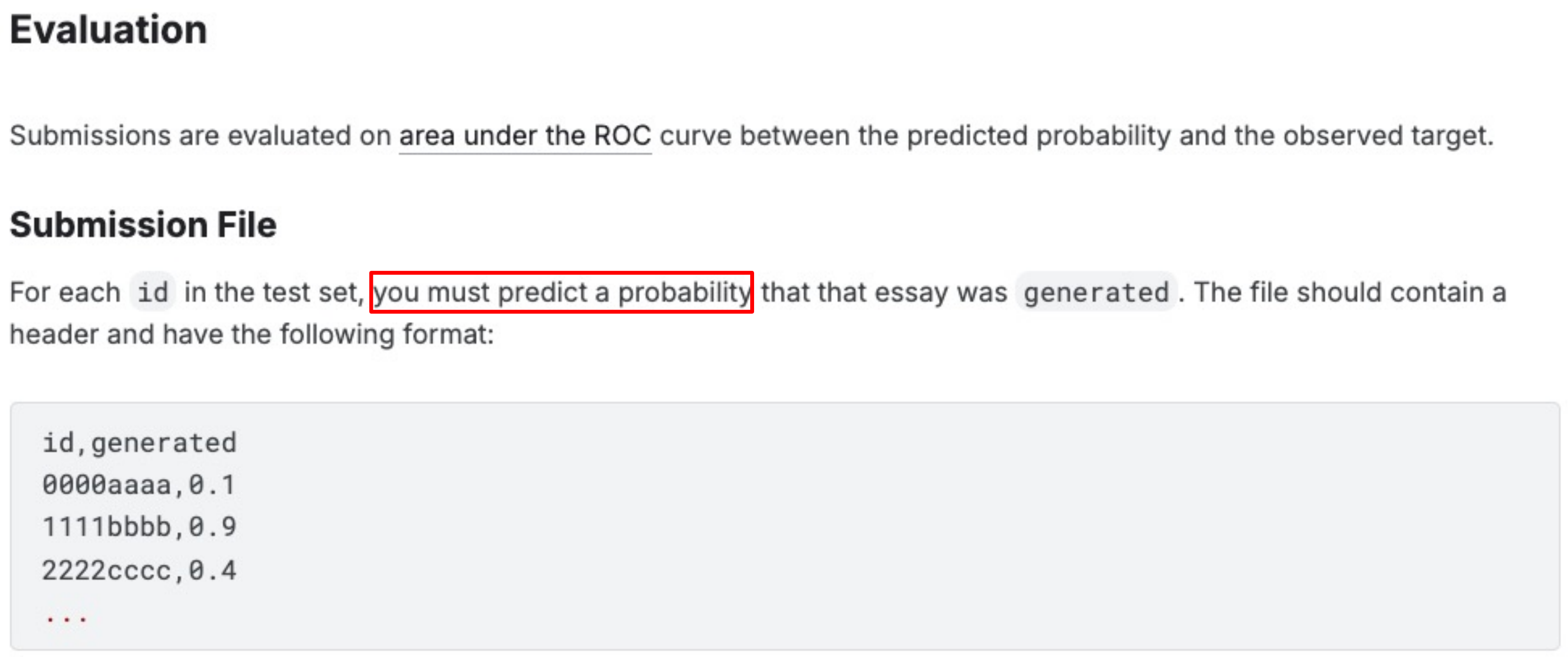
- sample_submission.csv
- test_essays.csv
- train_essays.csv
- train_prompts.csv

Summary

- 4 files
- 13 columns

可能有幫助的外部資料集：https://www.kaggle.com/datasets/alejopaullier/daigt-external-dataset

LLM - Detect AI Generated Text (submission and evaluation)



Evaluation

Submissions are evaluated on area under the ROC curve between the predicted probability and the observed target.

Submission File

For each `id` in the test set, you must predict a probability that that essay was `generated`. The file should contain a header and have the following format:

```
id,generated
0000aaaa,0.1
1111bbbb,0.9
2222cccc,0.4
...
```

- 上傳到 Kaggle Leaderboard 需要實作 .ipynb
- 你的 code 需要能 predict test_essays.csv 並產生檔案名稱為 submission.csv 的檔案，Kaggle 才能幫你執行程式碼並打分數，可參考 0.960| Phrases are keys

ROC Curve

(Useful!!) https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=zh-tw

ROC: Receiver operating characteristic curve (接收者操作特徵曲線)

AUC: ROC 的底面下面積 (越大越好)

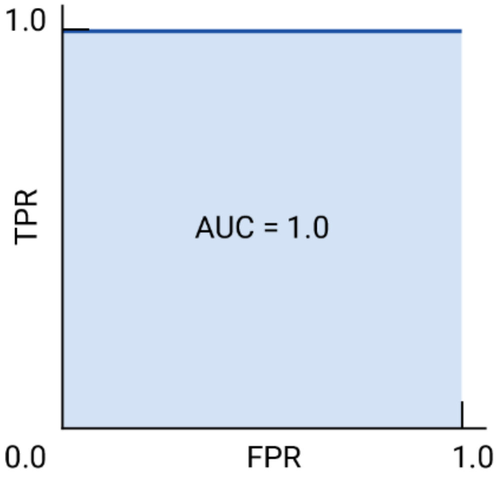


圖 1：ROC 和 AUC，這是一個完美假設的模型。

- TP / (TP + FN): True Positive Rate (TPR)
  - 又稱作 Recall
- FP / (FP+TN): False Positive Rate (FPR)

Project checkpoints (暫定)

- Week 9: 確定各組的題目
- Week 11: 進度報告 PPT (5 pages)
- Week 13: 進度報告 PPT (5+5 pages), Presentations (selected teams)
- Week 15 – Week 16: Final presentations for all teams (maybe poster)
- Week 16 結束前: 繳交書面報告以及程式碼

期末 Project 規定 (暫定)

- 需要上傳 Kaggle Leaderboard
  - 請留意，每個題目都是 Code competition，code 必須沒有 bug 才能上傳
- 每次報告都需要列出每位組員的貢獻內容，以及組員間的工作比重 (%)

Thank you!

Instructor: 林英嘉

✉️ yjlin@cgu.edu.tw