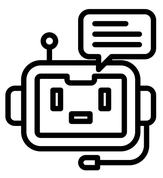
自然語言處理 NLP 2024 Term Project

Task: LLM Classification Finetuning **Team ID: 87** **Team Member: 李宥萱 吳定霖 吳宗樺**

任務介紹



- 此任務目的是預測使用者更偏好哪一個大型語言模型所回答的答案，幫助縮短 LLM 能力與人類偏好之間的差距。

- 與Reinforcement Learning from Human Feedback (RLHF)的概念密切相關，將有助於改善聊天機器人與人類的互動。

資料集介紹

訓練資料包含55,000筆資料，而測試集約25,000筆

欄位	描述
id	A unique identifier for the row.
model_[a/b]	The identity of model_[a/b]. Included in train.csv but not test.csv.
prompt	The prompt that was given as an input (to both models).
response_[a/b]	The response from model_[a/b] to the given prompt.
winner_model_[a/b/tie]	Binary columns marking the judge's selection. The ground truth target column.

輸入/輸出/評估方式

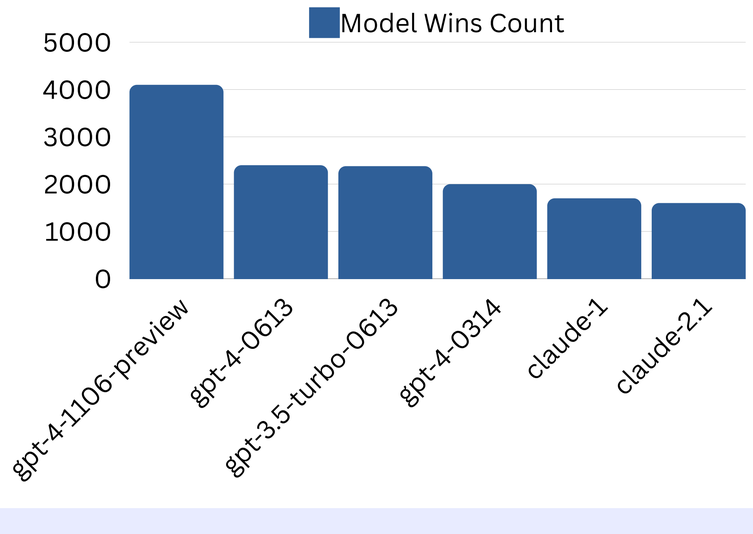
- 輸入: prompt, response_[a/b]
- 輸出: a 模型回應較好的機率、b 模型回應較好的機率、兩模型回應一樣好的機率 (三個機率值總和為 1)
- 評估方式: 三個預測機率值與正確答案的 Log Loss

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log(p_i) + (1-y_i)\log(1-p_i)\right)$$

  - N : 樣本的數量
  - Y_i : 是真實標籤
  - P_i : 是該樣本屬於正類的預測機率

EDA

- 共64種模型
- 平手的情形共出現17761次 (31%)
- 最常被兩兩比較的模型
  - gpt-4-1106-preview V.S. claude-2.1
  - gpt-4-1106-preview V.S. gpt-4-0613
  - claude-1 V.S. claude-2.1
- 模型贏的次數最多的前五名



研究方法

資料前處理

- 文本預處理
  - 使用 NLTK 將文本分詞
  - 移除停用詞與特殊字符
  - 詞形還原
- 資料格式化
  - 建立Prompt-Response配對
  - 將winner_model_[a/b/tie] 轉換為數字標籤

模型設計

DeBERTa-v3 small

- max_length = 512, epochs = 1, batch_size = 16, lr = 2e-5
- 各自的隱層輸出經過平均池化（mean-pooling)。
- 將兩者拼接後，通過一層全連接層進行分類。
- 最終使用 softmax 將 logits 轉化為概率分布。

訓練與驗證

- 損失函數與優化器
  - CrossEntropyLoss 作為分類任務的損失函數。
  - 使用 AdamW 並透過線性學習率調度器控制學習率。
- 訓練流程
  - 使用 torch.cuda.amp 提高混合精度訓練效率。
  - 前向傳播：計算 logits 和損失。
  - 反向傳播：通過梯度縮放 (amp.GradScaler) 避免數值不穩定。
  - 優化器更新權重。
- 推理過程
  - 應用 softmax，將 logits 轉換為 [0, 1] 範圍的概率。

研究結果

模型輸出loss後直接應用 softmax	inference 的時候再應用 softmax	epoch	score
X	O	1	1.04219
X	O	3	1.07467
O	O	1	1.04398
O (log_softmax)	O	1	1.07435
O	X	1	1.35201
O	X	3	1.58984

如果在計算損失前應用了 softmax，會導致兩個問題：

- 數值不穩定：當 logits 的值非常大或非常小時，指數函數容易導致溢出或下溢。
- 冗餘計算：CrossEntropyLoss 內部會再一次計算 log(softmax)，重複操作可能會降低效率。

後續研究方向

- 透過增強數據多樣性、分割特徵差異解決overfitting的問題
- 結合其他LLM，預期能有效綜合兩個模型的優勢，減少單一模型的偏誤。