**LLM Pairwise/Tie Classification – 最終整合報告**（基於 2025-08-07 的實驗與討論）

**資料來源補充**

本研究數據來自 2025 年 Kaggle 競賽 **「LLM Chat Response Pairwise Preference」**。官方資料集包含 ~71k 條對話記錄，每條樣本提供

### **1. 數據概況**

| **類別** | **訓練集樣本數** | **驗證集樣本數** | **佔比** |
| --- | --- | --- | --- |
| non-tie (0) | 4969 | 4 969×0.1 ≈ 497 | ≈ 70 % |
| tie (1) | 2165 | 2 165×0.1 ≈ 217 | ≈ 30 % |

**關鍵問題**：資料嚴重不平衡（≈ 7 : 3），且 *tie* 樣本本身較少，導致模型在該類別上難以學習到穩定判斷邊界。

### **2. 實驗設置（關鍵超參數）**

* **Backbone**：roberta-base, deberta-v3-large
* num\_train\_epochs = 2 - 5、per\_device\_train\_batch\_size = 8
* learning\_rate = 3 e-6（tie 分支）／2 e-5（pairwise 分支）
* fp16 + gradient\_accumulation\_steps = 2
* 嘗試 **class weight** 與 **增加 dropout (0.4)** — 效果有限
* 進一步的 **data cleaning / 擴充大型模型** → 仍無顯著改善，且可能降低整體準確率、耗時大
* Epochs = 2；Batch = 8 -16

### 

### 

### 

### 

### **3. 驗證集最終結果**

| **指標** | **總體** | **non-tie (0)** | **tie (1)** |
| --- | --- | --- | --- |
| **Accurac y** | **0.6995** |  |  |
| Precision |  | 0.7109 | 0.5240 |
| Recall |  | 0.9581 | **0.1058** |
| F1-score |  | 0.8162 | 0.1760 |
| Support | 7 134 | 4 969 | 2 165 |

**Confusion Matrix**

Pred 0 Pred 1

True 0 (non-tie) 4 761 208

True 1 (tie) 1 936 229

* **優勢**：對 non-tie 召回率高 (≈ 96 %)，整體 loss 維持 0.60 左右。
* **瓶頸**：tie 類別召回率僅 ≈ 11 %，F1 ≈ 0.18——主因仍是資料偏斜 + tie 樣本不足。

### **4. 已嘗試但收益有限的方向**

| **方法** | **結果 / 評論** |
| --- | --- |
| **Class weight** (balanced) | tie 召回率提升 < 1 %，整體波動不大 |
| **更大模型** | 記憶體緊張，acc 無明顯提升，耗時翻倍 |
| **強化 data cleaning** | 刪除長度極端、含噪文本 → 整體樣本減少，acc & F1 反而小幅下降 |
| **增大 dropout / 降低 LR** | overfitting 略緩，但 tie 指標仍 < 0.2 |
| **延長訓練 epoch** | > 4 epoch 後開始明顯 overfit，valid loss 上升 |

### 

### **5. 結論與建議**

1. **在目前資料規模與分佈下，70 % accuracy 已接近可達上限**。
2. 若業務對 *tie* 誤檢容忍度高（只需保證非 tie 準確判斷），當前模型可先行部署。
3. 若需要顯著提高 *tie* 召回 / F1，**必須從數據層面切入**：
   * 收集或標註更多高質量 tie 樣本（推薦 ≥ 倍增），或採半監督/弱監督擴充。
   * 考慮文本增廣（back-translation、EDA 等）專門針對 tie 類別。
4. 其他可選技巧（收益預估有限）：Focal Loss、閾值調優、過／欠採樣。

**備註**：完整訓練-評估（含 confusion matrix）雖耗時，但對定位不平衡問題至關重要；未來若再次縮時訓練，建議至少保留小批次 eval + confusion matrix 以迅速觀察類別召回情形。

**LLM Pairwise / Tie Classification – Final Consolidated Report (English)***(Experiments run up to 2025-08-07)*

### **Additional information about the data source**

### **The data for this study comes from the 2025 Kaggle competition “LLM Chat Response Pairwise Preference.” The official dataset contains approximately 71,000 conversation records, with each sample providing**

### 

### **1 · Data Overview**

| **Class** | **Train samples** | **Valid samples** | **Share** |
| --- | --- | --- | --- |
| **non-tie (0)** | 4 969 | ≈ 497 | ≈ 70 % |
| **tie (1)** | 2 165 | ≈ 217 | ≈ 30 % |

**Key issue:** The dataset is strongly imbalanced (~ 7 : 3). The minority *tie* class is under-represented, making it hard for the model to learn good decision boundaries.

### **2 · Experimental Setup (main hyper-parameters)**

* **Backbone**：roberta-base, deberta-v3-large
* num\_train\_epochs = 2 - 5、per\_device\_train\_batch\_size = 8
* learning\_rate = 3 e-6（tie 分支）／2 e-5（pairwise 分支）
* fp16 + gradient\_accumulation\_steps = 2
* 嘗試 **class weight** 與 **增加 dropout (0.4)** — 效果有限
* 進一步的 **data cleaning / 擴充大型模型** → 仍無顯著改善，且可能降低整體準確率、耗時大
* Epochs = 2；Batch = 8 -16

### 

### 

### **3 · Validation Results**

| **Metric** | **Overall** | **non-tie (0)** | **tie (1)** |
| --- | --- | --- | --- |
| **Accuracy** | **0.6995** |  |  |
| Precision |  | 0.7109 | 0.5240 |
| Recall |  | 0.9581 | **0.1058** |
| F1-score |  | 0.8162 | 0.1760 |
| Support | 7 134 | 4 969 | 2 165 |

**Confusion Matrix**

Pred 0 Pred 1

True 0 4 761 208

True 1 1 936 229

* Strength: very high recall on **non-tie** (≈ 96 %); validation loss ≈ 0.60.
* Weakness: only ~10 % recall on **tie**; F1 ≈ 0.18 – a direct consequence of data skew and limited tie samples.

### **4 · Approaches Already Tested (little or no gain)**

| **Method** | **Outcome / Notes** |
| --- | --- |
| **Class weights** (balanced) | tie recall ↑ < 1 %; negligible impact |
| **Larger backbone** | VRAM pressure, no clear accuracy gain, training time ×2 |
| **Aggressive data cleaning** | Fewer samples, slight drop in accuracy & F1 |
| **Higher dropout / lower LR** | Slightly less overfitting, tie metrics still < 0.20 |
| **More epochs** (> 4) | Clear overfitting, validation loss rises |

### 

### 

### 

### **5 · Conclusions & Recommendations**

1. **~70 % overall accuracy is close to the attainable ceiling on the current data.**
2. If downstream use-cases mainly need reliable *non-tie* detection, the model is deployable as-is.
3. To substantially boost *tie* recall/F1, **data work is essential**:
   * Collect / label many more high-quality *tie* examples (ideally doubling them) or apply semi-/weak-supervised augmentation.
   * Use targeted text augmentation (back-translation, EDA, etc.) for *tie* only.
4. Other techniques (focal loss, threshold tuning, over/under-sampling) may yield incremental gains but cannot solve the root imbalance alone.

**Note:** Even quick pilot runs should always produce a (mini) confusion matrix; it is the fastest way to reveal class-imbalance issues without waiting for full training.