人物關係擷取模型開發之研究 會議記錄

會議	專案啟動會議
日期時間	2024/03/08 14:00~15:30
地點	中央大學工程館 B323
出席人員	國安局: 董先生、王先生 WIDM 實驗室: 張嘉惠老師、洪閔昭、葉季儒
議程	二月份進度討論-親屬關係生成實測結果討論
結論事項	1. 再從 Common Crawl 數據中清洗出足夠的 positive data 訓練模型，若需要 10,000 筆 positive 資料，約需要原本資料量的 10 倍，也就是約 common crawl 2023-50 的 1%。 2. 並將原本模型標記 union 的 1098 筆作為 test data 的 positive 資料，原本的 26,293 筆則為 test data 總數。 3. 將新的清洗出來的數據，交給 Gemini 標記，標記完成後，切割出和 test data 相同數量作為 valid data。 4. 將 LLM 生成的各種關係，用於微調本地端的關係生成模型(嘗試 mt0 和 gemma)，實做測試將所有關係不過濾直接訓練。 5. 做一個關鍵字篩選器，拉除只有稱謂的情形，比較在拉除稱謂前的效能以及拉除稱謂後的效能。 6. 對於輸出格式不可控的部分，可以參考使用 guidance 來控制模型的自定義格式。 7. 如果 GPU 設備資源不足，可以參考唐鳳 github 上的 GGUF 方式。
備註	聯絡方式： 0228891088 董先生 #55334 陳小姐 #55335 孫小姐 #55675 (上班時間無法使用手機，比較急時打電話聯絡)