大型語言模型在法規關聯性分析與發債用途一致性之應用

Analyzing Regulatory Compliance and Bond Purpose Consistency with Large Language Models

學生: 黃鈺婷
指導教授: 戴天時 教授
報告日期: 2025/09/03

法規條文關聯

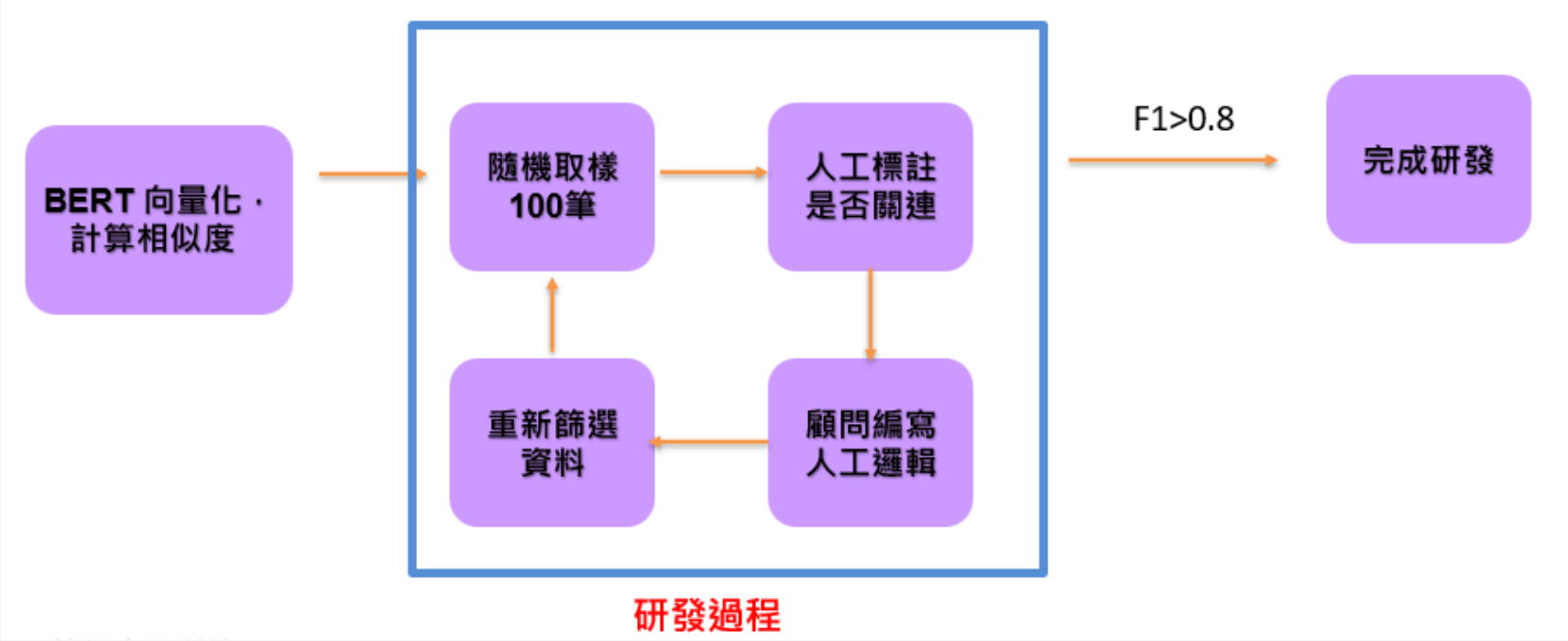
債券說明書和財報分析

研究背景與動機

- 研究背景：金融機構面臨日益複雜的法規遵循挑戰
  - 外規：金融業需遵守的外規超過5000部
  - 內規：超過800部，且需對應外規制定 ，以落實法遵精神
- 核心痛點：
  - 人工比對的困境：傳統上，法遵人員需手動比對數千條內外規，此過程耗時、效率低落且易因人為疏忽產生錯誤
  - 法遵風險：一旦外規變動，若未能即時更新對應內規，將產生嚴重的法遵風險
  - 資安限制：內規具機敏性，無法上雲處理，依賴雲端運算資源進行大型語言模型訓練和部署的方案面臨挑戰
- 研究目標：建構一個基於大型語言模型(LLMs)的AI關聯模型，自動化、高效且精準地建立內外規的關聯表

研究背景與動機

- 現有的AI關聯模型
  - 沒有辦法 retrain
  - BERT的結果有改善空間



文獻回顧 – 關鍵技術

- 本研究旨在建構一個能精準識別金融內外規關聯的AI模型，因此將回顧三項關鍵技術：
  1. 資訊檢索基礎: BM25 模型
  2. 深度學習語意相似度模型: Bi-encoder
  3. 資料標註策略: 主動學習 (Active Learning)

文獻回顧 – BM25模型

- 原理：基於詞頻的統計模型，計算速度快
- 缺點：
  - 無法捕捉深層語意關係，僅依賴關鍵字匹配
  - 在中文處理上，高度依賴分詞工具的準確性
    → 使用LLM(如GPT-4o)輔助分詞
    → 因外規可上雲，所以用LLM對外規進行分詞，並將分詞結果加入Jieba 的自定義字典中

文獻回顧 – Bi-encoder模型

- 原理：將內、外規文本**獨立編碼**為高維向量，再計算其語意相似度
- 為何選擇 Bi-encoder：相較於需要將文本對共同編碼的 Cross-encoder，Bi-encoder 在處理大規模資料(近200萬組文本對)時，推論效率具有壓倒性優勢
- 基礎模型：採用在多語言語意搜尋表現優異的 infloat/multilingual-e5-large

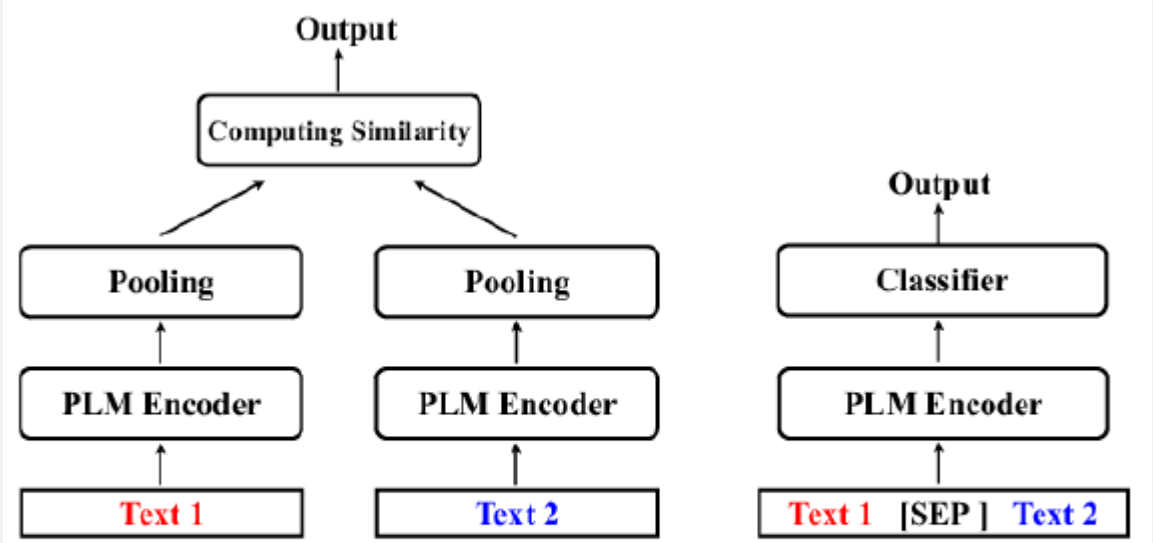


圖 1: Bi-encoder 和 Cross-emcoder 比較 [7]

文獻回顧 – 主動學習

- 面臨的挑戰：模型依賴大量高品質的標註資料，但法規文本的專業性使得大規模人工標註成本極高
- 解決方案：主動學習 (Active Learning)
- 概念：一種半監督學習方法，旨在用最少的標註成本達到最佳的模型性能
  - 策略:
    - 不確定性抽樣：挑選模型最「困惑」、最難以判斷的樣本
      →本研究使用取BM25分數和cosine similarity 兩者排名差異較大的樣本
    - 請求專家標註：將這些最有價值的樣本交由專家進行標註
      →本研究為了減少公司進行人工標註的人力，嘗試利用 LLMs(如Gemini)，來取代傳統的人工標註

研究方法 - 資料集建構

- 目前有的資料
  - 內規：6部共1623段
  - 外規：6部共1226段
  - 現有BERT產生的關聯表：4720組
    - 共有1,989,798組，其中1,985,078組不相關，4720組相關

Ex. 金融機構防制洗錢辦法

第 2 條

本辦法用詞定義如下：

一、金融機構：包括下列之銀行業、證券期貨業、保險業及其他經金融監督管理委員會（以下簡稱本會）指定之金融機構：

（一）銀行業：包括銀行、信用合作社、辦理儲金匯兌之郵政機構、票券金融公司、信用卡公司及信託業。

（二）證券期貨業：包括證券商、證券投資信託事業、證券金融事業、證券投資顧問事業、證券集中保管事業、期貨商。

（三）保險業：包括保險公司、專業再保險公司及辦理簡易人壽保險業務之郵政機構。

A	B	C	D	E	F	G
內規部	內規條	內規段	外規部	外規條	外規段	相似度
5	3	2	832	21	2	0.873646379
5	3	2	834	16	11	0.872581124
5	3	3	834	15	12	0.968946517
5	3	3	833	9	6	0.940811276
5	3	3	832	19	11	0.938995957
5	3	3	835	7	6	0.93470329
5	4	1	831	9	3	0.883011222
5	4	1	834	13	23	0.834484875
5	4	2	835	5	3	0.959981382
5	4	2	833	7	3	0.959981382
5	4	2	835	6	4	0.911621213
5	4	2	834	14	4	0.911621213
5	4	2	833	8	4	0.911621213
5	4	2	832	18	3	0.911621213

研究方法 - 資料集建構

- 測試資料集(Test Data)：
  - 目的：評估最終模型效能
  - 流程：從BERT關聯表中加權抽樣 →刪除一些重複的內外規組合後，留下136筆 → 使用LLM進行初步標註 → 由公司進行人工確認 → 91組相關＋45組不相關 → 從不在BERT關聯表的組合挑選64筆不相關

區間範圍	原始資料筆數	權重	權重比例	應抽取 250 筆中的數量
(0.809, 0.846]	518	5	0.333333	83
(0.846, 0.882]	570	4	0.266667	67
(0.882, 0.918]	671	3	0.200000	50
(0.918, 0.954]	718	2	0.133333	33
(0.954, 0.99]	817	1	0.066667	17

  - 最終組成：200筆資料 (91筆相關 vs. 109筆不相關)

研究方法 - 資料集建構

- 訓練與驗證資料集 (Train/Validation Data)：
  - 同樣經人工嚴謹篩選與確認，建構出訓練集 (660筆)與驗證集(66筆)，且維持正負樣本1:1的比例

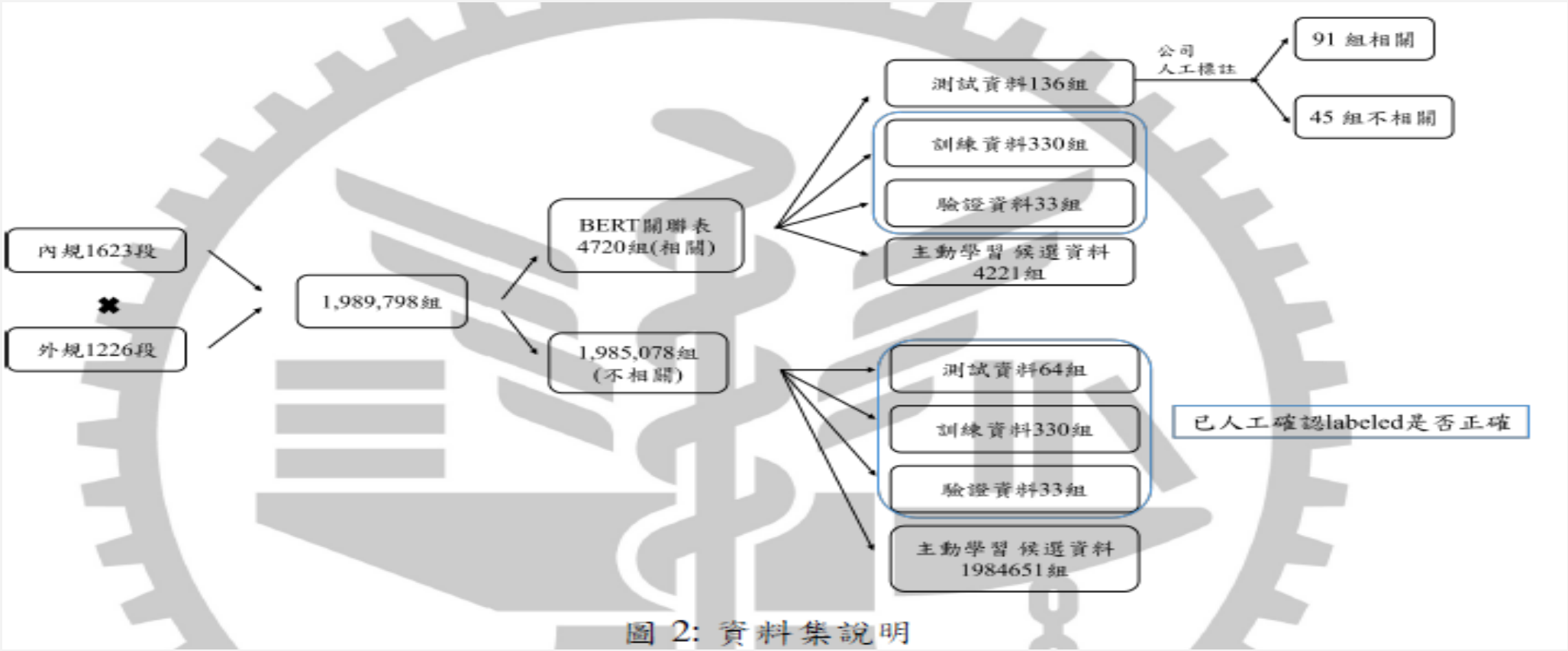


圖 2: 資料集說明

研究方法 – 實驗目標

- 輸入：一段外規
- 輸出：所有有相關的內規段落
  - 給定一段外規相關，可能輸出**多段**內規
  - 每一段外規輸出的內規段數並**不一致**

研究方法 – 實驗設計

- 目標：評估不同訓練策略與損失函數對Bi-encoder模型效能的影響
- 五組實驗設計：
  1. 實驗一：基準評估
     - 使用未經微調的 Bi-encoder 模型，評估其原始效能
  2. 實驗二： MNRL 微調
     - 使用 Multiple Negative Ranking Loss 進行微調
  3. 實驗三： Contrastive Loss 微調
     - 使用 Contrastive Loss 進行微調
  4. 實驗四：主動學習 + MNRL
     - 結合 BM25 檢索與主動學習策略，擴充訓練資料後，再以 MNRL 進行微調
  5. 實驗五：主動學習 + Contrastive Loss
     - 結合 BM25 檢索與主動學習策略，擴充訓練資料後，再以 Contrastive Loss 進行微調

實驗一：未經微調 Bi-encoder 的基準評估

- 目標：建立一個Baseline，評估預訓練模型在未經任何微調下的原始能力
- 方法：
  1. 直接載入預訓練的 Bi-encoder 模型
  2. 將測試資料集中的每一組內外規文本輸入模型，計算其餘弦相似度 (Cosine Similarity)
  3. 遍歷不同的相似度閾值 (0.00-1.00)，找出能讓 F1-Score 最高的最佳閾值
  4. 根據閾值將其餘弦相似度轉換為二元標籤
  5. 與ground truth比較，計算confusion matrix等指標
- 目的：
  - 作為後續所有微調實驗的對照組
  - 量化後續的微調步驟到底帶來了多少效能上的提升

實驗一：未經微調 Bi-encoder 的基準評估

- 流程圖：

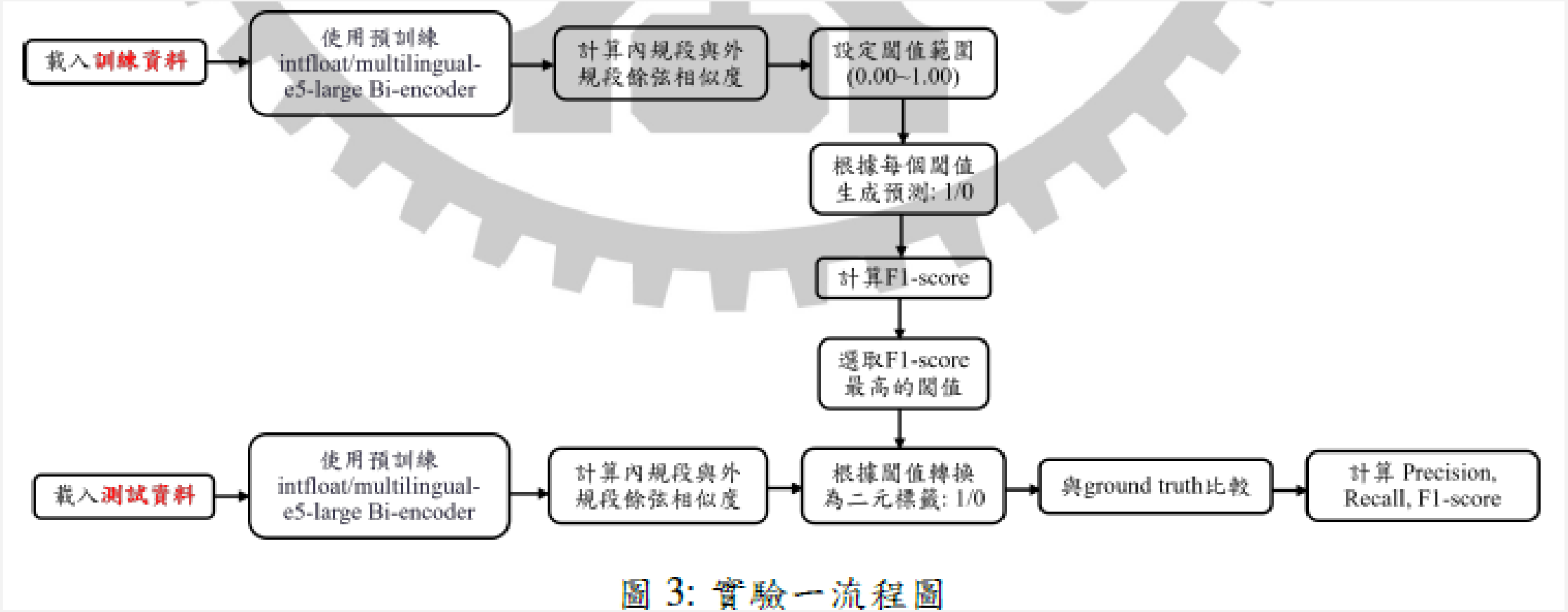


圖 3: 實驗一流程圖

實驗二：基於 MNRL 的 Bi-encoder 微調

- 目標：評估使用Multiple Negative Ranking Loss (MNRL) 損失函數進行微調後，模型效能的提升程度
- MNRL核心概念：
  - 運作方式：最大化「正樣本對」的相似度，同時最小化其與「**批次內其他樣本 (in-batch negatives)**」的相似度
  - 優點: 無需複雜的負樣本採樣策略，直接利用批次內的數據，效率高
  - 訓練資料：此方法僅需使用標註為相關 (labeled=1) 的正樣本對進行訓練

實驗二：基於 MNRL 的 Bi-encoder 微調

- 流程圖：

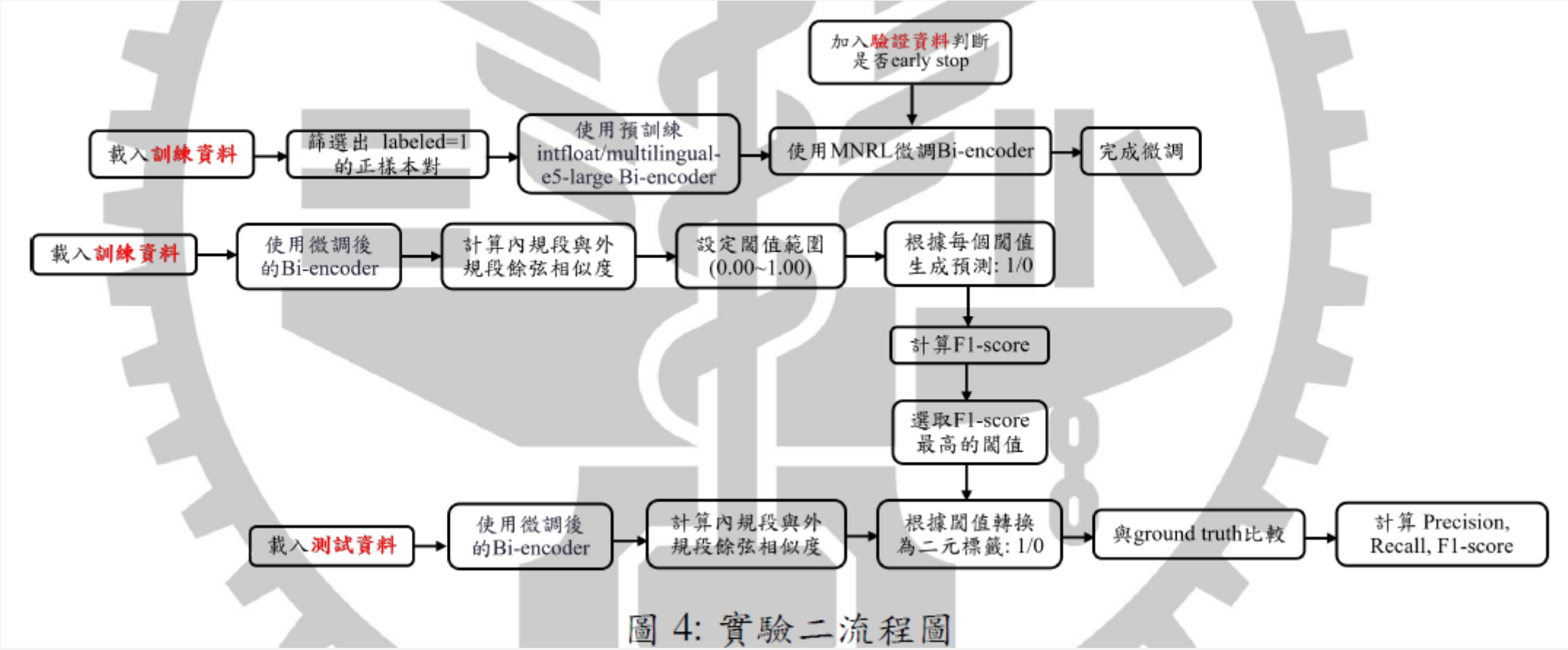


圖 4: 實驗二流程圖

實驗三：基於 Contrastive Loss 的 Bi-encoder 微調

- 目標：評估使用傳統且直觀的 Contrastive Loss 進行微調後的模型效能，並與實驗二的 MNRL 進行比較
- Contrastive Loss核心概念：
  - 運作方式：
    - 拉近「正樣本對」在向量空間中的距離
    - 推開「負樣本對」的距離，使其大於一個預設的邊界值 (margin)
  - 訓練資料：需要同時使用相關 (labeled=1) 與不相關 (labeled=0) 的樣本對進行訓練

實驗三：基於 Contrastive Loss 的 Bi-encoder 微調

- 流程圖：

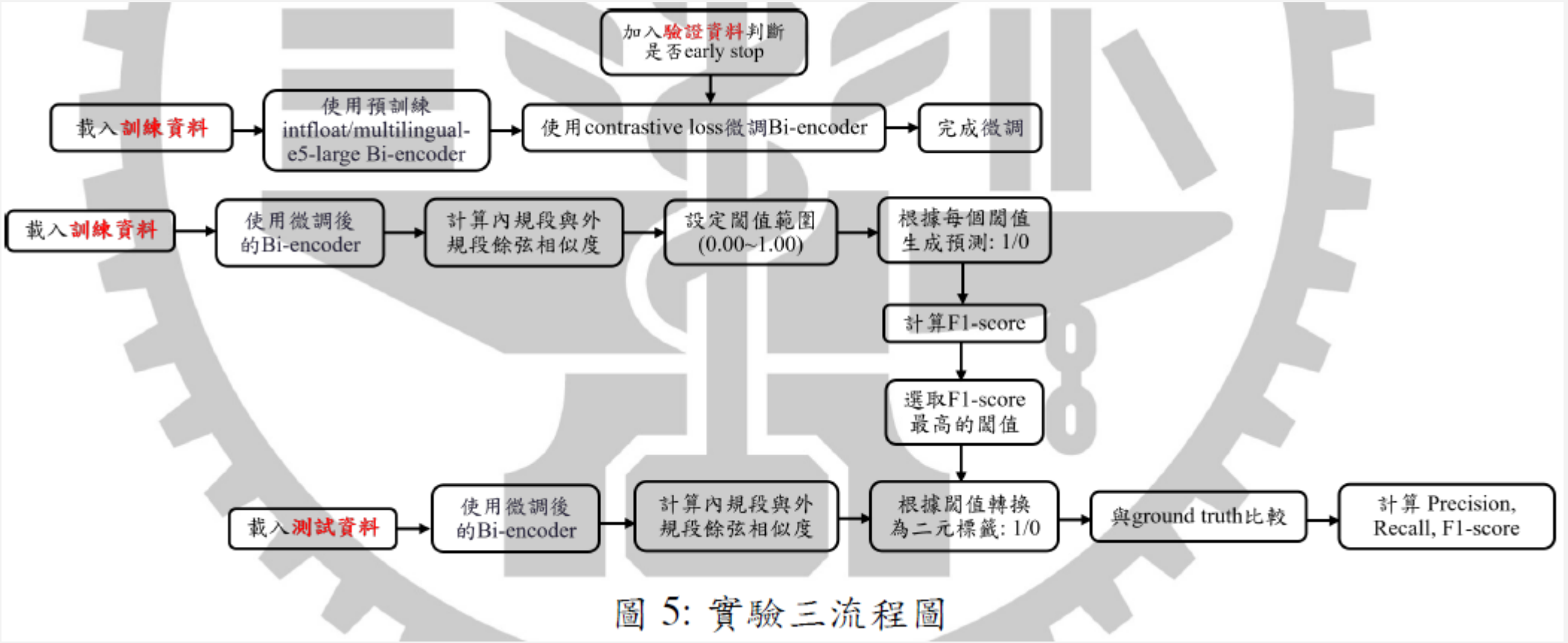


圖 5: 實驗三流程圖

實驗四與五：結合主動學習的微調

- 目標：探討在更大規模的未標註資料上，結合主動學習策略，是否能進一步提升模型效能
- 主動學習抽樣策略：
  1. 雙重評分：使用 BM25 (關鍵字匹配) 和已微調的 Bi-encoder (語意匹配) 分別對剩餘的大量未標註資料進行評分與排名
  2. 選取模糊樣本：找出兩種方法排名差異最大的前 250 筆資料，這些通常是「語意相關但關鍵字不重疊」或「關鍵字重疊但語意無關」的困難樣本，對模型學習最有價值
- 標註與迭代微調：
  - 使用 LLM (Gemini) 和Cross-encoder 模型輔助標註這 250 筆樣本
  - 將這些新標註的資料加入原始訓練集
  - 實驗四：使用擴充後的資料集，以 **MNRL** 繼續微調
  - 實驗五：使用擴充後的資料集，以 **Contrastive Loss** 繼續微調

實驗四與五：結合主動學習的微調

- 實驗四流程圖：

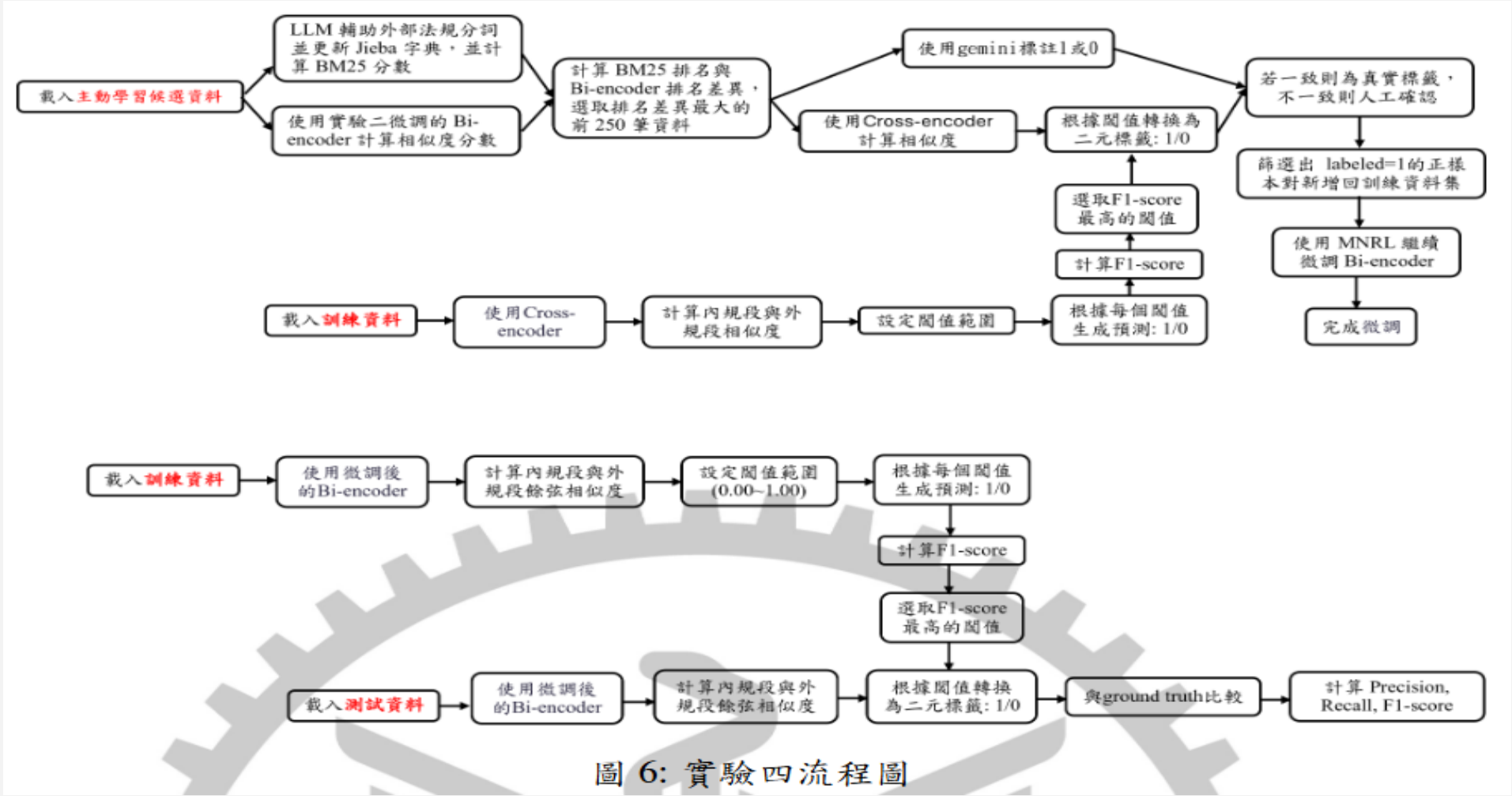


圖 6: 實驗四流程圖

實驗四與五：結合主動學習的微調

- 實驗五流程圖：

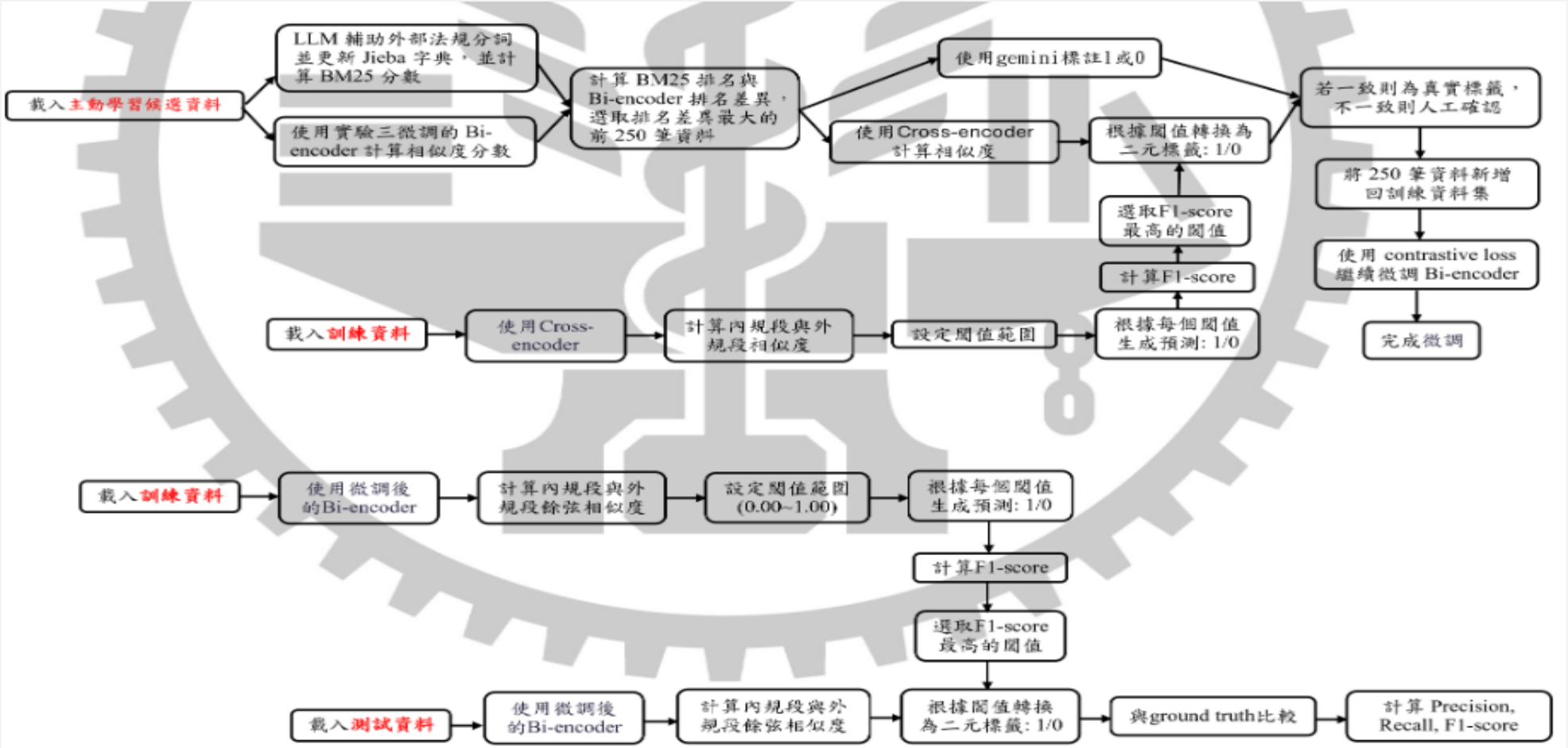


圖 7: 實驗五流程圖

實驗結果 - 評估指標

- 混淆矩陣 (Confusion Matrix)：視覺化呈現模型預測結果與實際標籤的關係
  - 真陽性 (TP)：模型預測相關，實際也相關
  - 偽陽性 (FP)：模型預測相關，實際不相關(增加人工審核成本)
  - 真陰性 (TN)：模型預測不相關，實際也不相關
  - 偽陰性 (FN)：模型預測不相關，實際卻相關 (最嚴重的錯誤，會導致法遵風險)

實驗結果 - 評估指標

- 精確率 (Precision)：模型預測為「相關」的結果中，有多少是真正相關的，重要性在於控制「誤報」的成本
  - $Precision = \frac{TP}{TP+FP}$
- 召回率 (Recall)：所有真正「相關」的樣本中，有多少被模型成功辨識，重要性在於控制「漏報」的成本
  - $Recall = \frac{TP}{TP+FN}$
- F1-Score：精確率與召回率的調和平均數，提供一個能夠平衡兩者表現的綜合性指標
  - $F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} = \frac{2TP}{2TP+FP+FN}$

實驗結果分析 -評估指標的選擇考量

- 為何Recall不能作為BERT的Benchmark？
- Recall 的重要性：
  - 在法規關聯判斷的應用情境中，偽陰性(FN)代表模型遺漏了應找出的真實關聯，可能導致嚴重的法遵風險，因此召回率(Recall)應該要是特別關注的指標
- BERT 模型 Recall 指標的限制：
  - 根據表2、4、6、8、10的數據，作為比較基準的BERT模型，其召回率在所有測試情境下皆為 1
  - 原因：本研究的資料集建構方式 -- 所有被標記為「相關」的樣本均來自於既有的BERT關聯表，不會產生任何偽陰性(FN=0)，導致召回率指標虛高
- BERT 的召回率無法作為一個有意義的benchmark，後續的比較將更側重於精確率與綜合性的 F1-Score，這兩項指標能更好地反映模型在抑制偽陽性(FP)、提升預測結果可信度方面的實際能力

改善BERT的Recall無法當作Benchmark的問題

- 加入FN的樣本(不在BERT關聯表但實際相關)
- 因為不相關的樣本數太多，如果用隨機抽樣，可能抽不到真正相關的樣本
- Bi encoder標不在BERT關聯表裡的樣本
- 用哪個bi-encoder?
  - 是否可以用與後面實驗相同的或是要用不同的bi-encoder
- 挑出n筆bi-encoder分數最高的資料
- 將n筆資料用LLM和cross-encoder標記相關或不相關
  - 兩種結果都相關就直接視為FN
  - 不一致的交給公司進行人工標註

接續改善BERT的Recall無法當作Benchmark的問題__未來研究方向

- 目前已從所有資料扣除驗證、訓練、測試的子資料集中找到不在BERT關聯表的不相關資料
- 因為目標是尋找 FN，所以希望是能找到BERT覺得不相關但實際是相關。
  - 目前是先利用實驗一未經微調的 bi-encoder 以及其對應的 threshold (0.92)找出 12965 筆應被標為相關的樣本
- 如何將 n 筆資料(= 12965) 進行標注
  - 是否需將這 n 筆先進一步縮減成更小的範圍
  - 初步想法是透過不同的模型找共同標註為「相關」的資料（由前面實驗所用模型或其他LLM）。
- 接續可能的問題
  - 模型標註的基礎邏輯(單一模型的表現(是否微調過)、多個模型的交集比較)

實驗結果分析 - BERT vs. 未微調 Bi-encoder

		預測標籤	
		相關 (Positive)	不相關 (Negative)
實驗一			
標籤資料	相關 (Positive)	88 (96.7%)	3 (3.3%)
	不相關 (Negative)	38 (34.9%)	71 (65.1%)
BERT[4]			
標籤資料	相關 (Positive)	91 (100%)	0 (0%)
	不相關 (Negative)	45 (41.3%)	64 (58.7%)

表 1: 各實驗在測試集上的混淆矩陣比較

	Accuracy	Precision	Recall	F1 score
實驗一	0.7950	0.6984	0.9670	0.8111
實驗二	0.8050	0.7031	0.9890	0.8219
實驗三	0.7850	0.7000	0.9231	0.7962
實驗四	0.7950	0.6984	0.9670	0.8111
實驗五	0.7900	0.6992	0.9451	0.8037
BERT	0.7750	0.6691	~~1.0000~~	0.8018

表 2: 各實驗在測試集上的效能指標

- 即使未經微調，先進的 Bi-encoder 模型在抑制「誤報」(FP) 方面的基礎能力，已優於作為基準的 BERT 模型

實驗結果分析 -不同損失函數的微調效果

		預測標籤	
		相關 (Positive)	不相關 (Negative)
實驗二			
標籤資料	相關 (Positive)	90 (98.9%)	1 (1.1%)
	不相關 (Negative)	38 (34.9%)	71 (65.1%)
實驗三			
標籤資料	相關 (Positive)	84 (92.3%)	7 (7.7%)
	不相關 (Negative)	36 (33.0%)	73 (67.0%)
BERT			
標籤資料	相關 (Positive)	91 (100%)	0 (0%)
	不相關 (Negative)	45 (41.3%)	64 (58.7%)

表 1: 各實驗在測試集上的混淆矩陣比較

	Accuracy	Precision	Recall	F1 score
實驗一	0.7950	0.6984	0.9670	0.8111
實驗二	0.8050	0.7031	0.9890	0.8219
實驗三	0.7850	0.7000	0.9231	0.7962
實驗四	0.7950	0.6984	0.9670	0.8111
實驗五	0.7900	0.6992	0.9451	0.8037
BERT	0.7750	0.6691	~~1.0000~~	0.8018

表 2: 各實驗在測試集上的效能指標

- MNRL 在綜合效能與風險控制上，均展現了比 Contrastive Loss 更適合本任務的特性

實驗結果分析 – 主動學習

	Accuracy	Precision	Recall	F1 score
實驗一	0.7950	0.6984	0.9670	0.8111
實驗二	0.8050	0.7031	0.9890	0.8219
實驗三	0.7850	0.7000	0.9231	0.7962
實驗四	0.7950	0.6984	0.9670	0.8111
實驗五	0.7900	0.6992	0.9451	0.8037
BERT	0.7750	0.6691	~~1.0000~~	0.8018

表 2: 各實驗在測試集上的效能指標

- 實驗四 (MNRL + Active Learning)：
  - 結果： 效能與實驗一相同，沒有變化
  - 原因： MNRL 的訓練機制僅利用「相關」樣本 ，若主動學習挑選的新資料中，「相關」樣本比例過低，則對模型訓練的貢獻將十分有限

		預測標籤 相關 (Positive)	預測標籤 不相關 (Negative)
實驗一			
真實標籤	相關 (Positive)	88 (96.7%)	3 (3.3%)
真實標籤	不相關 (Negative)	38 (34.9%)	71 (65.1%)
實驗二			
真實標籤	相關 (Positive)	90 (98.9%)	1 (1.1%)
真實標籤	不相關 (Negative)	38 (34.9%)	71 (65.1%)
實驗三			
真實標籤	相關 (Positive)	84 (92.3%)	7 (7.7%)
真實標籤	不相關 (Negative)	36 (33.0%)	73 (67.0%)
實驗四			
真實標籤	相關 (Positive)	88 (96.7%)	3 (3.3%)
真實標籤	不相關 (Negative)	38 (34.9%)	71 (65.1%)
實驗五			
真實標籤	相關 (Positive)	86 (94.5%)	5 (5.5%)
真實標籤	不相關 (Negative)	37 (33.9%)	72 (66.1%)
BERT[4]			
真實標籤	相關 (Positive)	91 (100%)	0 (0%)
真實標籤	不相關 (Negative)	45 (41.3%)	64 (58.7%)

表 1: 各實驗在測試集上的混淆矩陣比較

實驗結果分析 – 主動學習

	Accuracy	Precision	Recall	F1 score
實驗一	0.7950	0.6984	0.9670	0.8111
實驗二	0.8050	0.7031	0.9890	0.8219
實驗三	0.7850	0.7000	0.9231	0.7962
實驗四	0.7950	0.6984	0.9670	0.8111
實驗五	0.7900	0.6992	0.9451	0.8037
BERT	0.7750	0.6691	~~1.0000~~	0.8018

表 2: 各實驗在測試集上的效能指標

- 實驗五 (Contrastive Loss + Active Learning)：
  - 結果： 效能提升 (F1-Score 從 0.7962 → 0.8037，FN 從 7 → 5)
  - 原因： Contrastive Loss 同時利用「相關」與「不相關」樣本。這表明主動學習引入的新樣本，確實為模型提供了有價值的資訊，幫助其學習區分正負樣本的邊界

		預測標籤	
		相關 (Positive)	不相關 (Negative)
實驗一			
真實標籤	相關 (Positive)	88 (96.7%)	3 (3.3%)
	不相關 (Negative)	38 (34.9%)	71 (65.1%)
實驗二			
真實標籤	相關 (Positive)	90 (98.9%)	1 (1.1%)
	不相關 (Negative)	38 (34.9%)	71 (65.1%)
實驗三			
真實標籤	相關 (Positive)	84 (92.3%)	7 (7.7%)
	不相關 (Negative)	36 (33.0%)	73 (67.0%)
實驗四			
真實標籤	相關 (Positive)	88 (96.7%)	3 (3.3%)
	不相關 (Negative)	38 (34.9%)	71 (65.1%)
實驗五			
真實標籤	相關 (Positive)	86 (94.5%)	5 (5.5%)
	不相關 (Negative)	37 (33.9%)	72 (66.1%)
BERT[4]			
真實標籤	相關 (Positive)	91 (100%)	0 (0%)
	不相關 (Negative)	45 (41.3%)	64 (58.7%)

表 1: 各實驗在測試集上的混淆矩陣比較

實驗結果分析

	Accuracy	Precision	Recall	F1 score
實驗一	0.7950	0.6984	0.9670	0.8111
實驗二	0.8050	0.7031	0.9890	0.8219
實驗三	0.7850	0.7000	0.9231	0.7962
實驗四	0.7950	0.6984	0.9670	0.8111
實驗五	0.7900	0.6992	0.9451	0.8037
BERT	0.7750	0.6691	~~1.0000~~	0.8018

表 2: 各實驗在測試集上的效能指標

- 實驗一到五比BERT好，但微調的結果沒很穩定
  - 可能是資料集的問題
  - 將資料集打散重做一次

		預測標籤	
		相關 (Positive)	不相關 (Negative)
實驗一			
標準答案	相關 (Positive)	88 (96.7%)	3 (3.3%)
	不相關 (Negative)	38 (34.9%)	71 (65.1%)
實驗二			
標準答案	相關 (Positive)	90 (98.9%)	1 (1.1%)
	不相關 (Negative)	38 (34.9%)	71 (65.1%)
實驗三			
標準答案	相關 (Positive)	84 (92.3%)	7 (7.7%)
	不相關 (Negative)	36 (33.0%)	73 (67.0%)
實驗四			
標準答案	相關 (Positive)	88 (96.7%)	3 (3.3%)
	不相關 (Negative)	38 (34.9%)	71 (65.1%)
實驗五			
標準答案	相關 (Positive)	86 (94.5%)	5 (5.5%)
	不相關 (Negative)	37 (33.9%)	72 (66.1%)
BERT[4]			
標準答案	相關 (Positive)	91 (100%)	0 (0%)
	不相關 (Negative)	45 (41.3%)	64 (58.7%)

表 1: 各實驗在測試集上的混淆矩陣比較

實驗結果分析－交叉驗證

- 訓練資料的組成是優化模型的關鍵
- 目標：為了進一步驗證實驗結果的穩健性，並確認模型的效能是否受到特定資料集劃分的影響
- 核心原則：在保持與原始資料集(訓練、驗證、測試集) 相同總筆數及正負樣本比例的前提下，對資料進行重新劃分與抽樣
- 四種資料集重抽樣策略
  1. 保留80% 測試集
  2. 保留60% 測試集
  3. 保留40% 測試集
  4. 保留20% 測試集

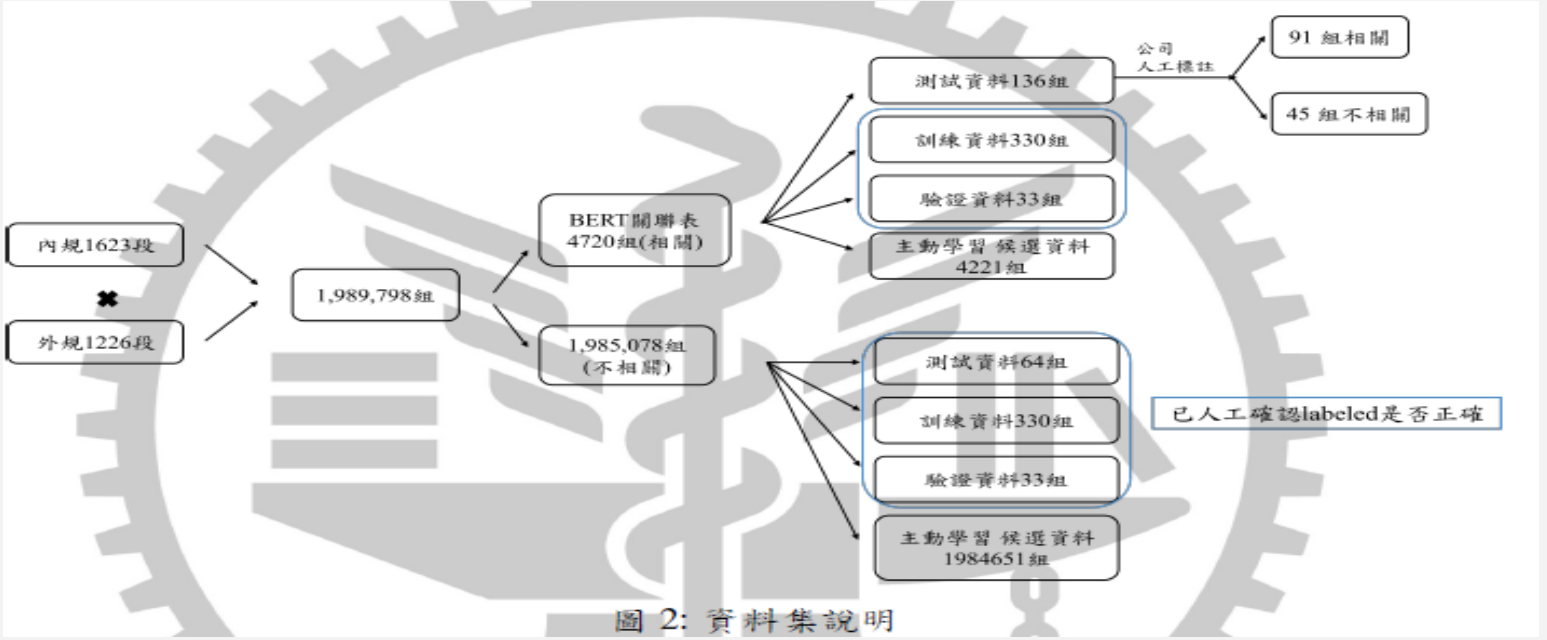


圖 2: 資料集說明

實驗結果分析 – 交叉驗證

- 情境一：保留 80% 測試集
  - 現象： 此情境下，新的訓練集僅混入約 20% 的不確定性資料。實驗二(MNRL)的表現最佳，F1-Score與精確率均顯著優於其他實驗
  - 在此設定下，微調能有效將 FN 數量降至 0。然而，實驗三(Contrastive Loss)的 FP 不減反增，導致整體表現不佳
  - 初步結論： 少量的『不確定性資料』可能對 MNRL 的微調有正面效益，但對 Contrastive Loss 則可能產生干擾

	Accuracy	Precision	Recall	F1 score
實驗一	0.8300	0.7395	0.9670	0.8381
實驗二	0.8700	0.7778	1.0000	0.8750
實驗三	0.7900	0.6842	1.0000	0.8125
BERT	0.8150	0.7109	~~1.0000~~	0.8311

表 4: 各實驗在「保留 80% 測試集」情境下，於新測試資料集上的效能指標

		預測標籤	
		相關 (Positive)	不相關 (Negative)
實驗一			
真實標籤	相關 (Positive)	88 (96.7%)	3 (3.3%)
	不相關 (Negative)	31 (28.4%)	78 (71.6%)
實驗二			
真實標籤	相關 (Positive)	91 (100%)	0 (0.0%)
	不相關 (Negative)	26 (23.9%)	83 (76.1%)
實驗三			
真實標籤	相關 (Positive)	91 (100%)	0 (0.0%)
	不相關 (Negative)	42 (38.5%)	67 (61.5%)
BERT			
真實標籤	相關 (Positive)	91 (100%)	0 (0.0%)
	不相關 (Negative)	37 (33.9%)	72 (66.1%)

表 3: 各實驗在「保留 80% 測試集」情境下，於新測試資料集上的混淆矩陣比較

實驗結果分析－交叉驗證

- 情境二：保留 60% 測試集 → 最佳平衡點
  - 在此情境下，實驗一、二、三的 F1-Score 與精確率全面優於 BERT
  - MNRL 優勢凸顯： 實驗二(MNRL)的表現尤其突出，不僅 FP 數量為所有實驗中最低，FN 數量也與未微調的實驗一持平，取得了最佳的綜合表現
  - 與情境一比較： 若將本情境的精確率(表6)與情境一的精確率(表4)進行比較，可以發現本情境下各實驗的精確率普遍更高

→在訓練集中策略性地引入約 40% 的不確定性資料，有助於模型更有效地學習，達到一個抑制 FP 與控制 FN 的最佳平衡點

	Accuracy	Precision	Recall	F1 score
實驗一	0.8450	0.7586	0.9670	0.8502
實驗二	0.8650	0.7857	0.9670	0.8670
實驗三	0.8450	0.7542	0.9780	0.8517
BERT	0.8300	0.7280	~~1.0000~~	0.8426

表 6: 各實驗在「保留 60% 測試集」情境下，於新測試資料集上的效能指標

		預測標籤	
		相關 (Positive)	不相關 (Negative)
實驗一			
標籤	相關 (Positive)	88 (96.7%)	3 (3.3%)
標籤	不相關 (Negative)	28 (25.7%)	81 (74.3%)
實驗二			
標籤	相關 (Positive)	88 (96.7%)	3 (3.3%)
標籤	不相關 (Negative)	24 (22.0%)	85 (78.0%)
實驗三			
標籤	相關 (Positive)	89 (97.8%)	2 (2.2%)
標籤	不相關 (Negative)	29 (26.6%)	80 (73.4%)
BERT			
標籤	相關 (Positive)	91 (100%)	0 (0.0%)
標籤	不相關 (Negative)	34 (31.2%)	75 (68.8%)

表 5: 各實驗在「保留 60% 測試集」情境下，於新測試資料集上的混淆矩陣比較

實驗結果分析 – 交叉驗證

- 引入過多不確定性資料的影響
- 情境三：保留 40% 測試集
  - 此情境讓訓練集混入了約 60% 的不確定性資料。雖然各實驗的精確率仍優於 BERT，但微調的效果已不顯著
  - 問題： 實驗二的 FP 甚至比實驗一還多，而實驗三的 FN 也增加
  - 初步結論： 當引入過多不確定性資料時，在總資料筆數有限的情況下，可能反而會干擾模型的穩定收斂

	Accuracy	Precision	Recall	F1 score
實驗一	0.8950	0.8241	0.9780	0.8945
實驗二	0.8900	0.8165	0.9780	0.8900
實驗三	0.8900	0.8224	0.9670	0.8889
BERT	0.8900	0.8053	~~1.0000~~	0.8922

表 8: 各實驗在「保留 40% 測試集」情境下，於新測試資料集上的效能指標

		預測標籤	
		相關 (Positive)	不相關 (Negative)
實驗一			
真實標籤	相關 (Positive)	89 (97.8%)	2 (2.2%)
	不相關 (Negative)	19 (17.4%)	90 (82.6%)
實驗二			
真實標籤	相關 (Positive)	89 (97.8%)	2 (2.2%)
	不相關 (Negative)	20 (18.3%)	89 (81.7%)
實驗三			
真實標籤	相關 (Positive)	88 (96.7%)	3 (3.3%)
	不相關 (Negative)	19 (17.4%)	90 (82.6%)
BERT			
真實標籤	相關 (Positive)	91 (100%)	0 (0.0%)
	不相關 (Negative)	22 (20.2%)	87 (79.8%)

表 7: 各實驗在「保留 40% 測試集」情境下，於新測試資料集上的混淆矩陣比較

實驗結果分析 – 交叉驗證

- 引入過多不確定性資料的影響
- 情境四：保留 20% 測試集
  - 此情境讓訓練集混入了約 80% 的不確定性資料經微調的實驗二與實驗三，其 F1-Score 與精確率皆優於 BERT 及未微調的實驗一
  - 微調在此情境下，同時降低了 FN 與 FP 的數量，展現了正面效益，但減少的幅度並不明顯
  - 初步結論：雖然仍有正面效果，但效益已開始遞減，不如「保留60%測試集」情境顯著

	Accuracy	Precision	Recall	F1 score
實驗一	0.9450	0.9082	0.9780	0.9418
實驗二	0.9550	0.9184	0.9890	0.9524
實驗三	0.9500	0.9175	0.9780	0.9468
BERT	0.9550	0.9100	~~1.0000~~	0.9529

表 10: 各實驗在「保留 20% 測試集」情境下，於新測試資料集上的效能指標

		預測標籤	
		相關 (Positive)	不相關 (Negative)
實驗一			
標籤資料	相關 (Positive)	89 (97.8%)	2 (2.2%)
	不相關 (Negative)	9 (8.3%)	100 (91.7%)
實驗二			
標籤資料	相關 (Positive)	90 (98.9%)	1 (1.1%)
	不相關 (Negative)	8 (7.3%)	101 (92.7%)
實驗三			
標籤資料	相關 (Positive)	89 (97.8%)	2 (2.2%)
	不相關 (Negative)	8 (7.3%)	101 (92.7%)
BERT			
標籤資料	相關 (Positive)	91 (100%)	0 (0.0%)
	不相關 (Negative)	9 (8.3%)	100 (91.7%)

表 9: 各實驗在「保留 20% 測試集」情境下，於新測試資料集上的混淆矩陣比較

實驗結果分析 – 交叉驗證總結

1. 微調策略普遍有效：
   - 綜合所有情境，微調策略(特別是MNRL)普遍能提升模型效能，尤其在抑制偽陽性(FP)方面
2. 效能高度依賴資料組成：
   - 模型的最終表現，極度敏感於訓練資料的組成結構
3. 在訓練集中策略性地引入一定比例(約40%)的「不確定性資料」，似乎能最有效地提升微調成效
   - 太少，模型學習有限
   - 太多，則可能因總資料量不足而干擾收斂

   →實驗四五的active learning應該要挑多少樣本?

結論

1. Bi-encoder 優於基準
   - 經微調的 Bi-encoder 模型，透過顯著減少偽陽性(FP)，在精確率與 F1-Score 上超越了既有的 BERT 模型，能有效降低法遵人員的人工審核負擔
2. MNRL 為首選策略
   - MNRL 損失函數在降低偽陰性(FN)風險上表現最佳，最符合法遵應用中，極力避免「錯放」任何潛在關聯的嚴苛需求
3. 資料組成的重要性
   - 適度在訓練集中引入「不確定性資料」，是有效提升模型泛化能力的關鍵，此發現也為導入主動學習策略提供了實證基礎
4. 研究限制
   - 由於資料集建構方式，本研究無法客觀衡量模型在改善偽陰性(FN)方面的真實能力，這是未來研究需要克服的挑戰

未來研究

1. 改善測試集
   - 目標： 克服當前研究限制，實現對召回率的公允評估
2. 導入主動學習
   - 目標： 以最少標註成本，最大化模型學習效率
   - 本研究已初步證實**適量**的「不確定性資料」可以幫助微調
     - 須衡量要給多少不確定性資料，可能需要加入一些非常確定的樣本
   - 目前實驗：
     - 挑選BM25和Bi-encoder分數排名差異大的前250筆
     - MNRL：250筆進行標註，只取標註為相關的樣本新增回訓練集
     - Contrastive：250筆標註後皆會新增回訓練集
   - 未來研究：
     - 可能也要加入確定的樣本，不能只用不確定性樣本
     - 較確定的樣本可以挑選BM25和Bi-encoder分數排名差異小的

法規條文關聯

債券說明書和財報分析

緒論

- 研究背景：企業透過舉債籌集的資金，其**運用方向**對自身信用風險有顯著影響
- 研究缺口：過往研究多從財務槓桿角度探討信用風險，較少深入追蹤舉債資金是否**實際流向**如併購 (M&A) 等高風險投資活動

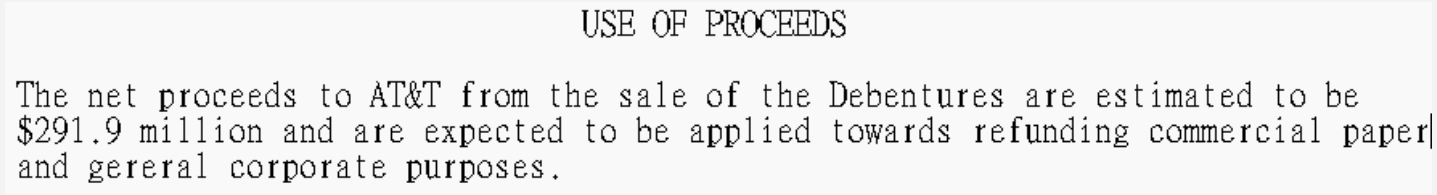
  →本研究將嚴格地鎖定為併購其他公司實體的行為，不包含 Property, Plant & Equipment(PP&E)等資產型投資，因為企業實體併購通常被視為涉及更高不確定性及潛在風險的投資行為
- 研究目標：
  1. 透過分析公開文件，識別出將發債資金明確用於併購的企業
  2. 驗證這些企業是否實際執行了併購行為

研究方法 – 資料來源

- 美國證券交易委員會 (SEC) 的公開資料
  - 債券說明書(424B2)：1,954 間公司，共 11,003 份
  - 公司年報(10-K)：對應公司的所有年報

研究方法－實驗設計

- 第一階段：確認債券發行目的
  - 目標：從債券說明書中，篩選出發債目的是「併購」的樣本
  - 方法：使用 LLM 分析說明書中的「Use of Proceeds (資金用途)」段落，自動化標記



USE OF PROCEEDS

The net proceeds to AT&T from the sale of the Debentures are estimated to be $291.9 million and are expected to be applied towards refunding commercial paper and gereral corporate purposes.

- 第二階段：確認債券實際用途
  - 目標：確認第一階段篩選出的公司，是否真的執行了併購
  - 方法：使用 LLM 分析對應公司年報中的「 Item 7：Management's Discussion and Analysis of Financial Condition and Results of Operations」(MD&A) 段落，核實併購活動的揭露情況

實驗結果與分析

- 第一階段成果：
  - 在 11,003 份債券說明書中，LLM 成功識別出 3,869 份明確提及將資金用於「併購」相關活動

```
prompt = ("The following are ten potential uses for a company's bond issuance. Based on the provided reference material, please determine if the company's plan includes these uses.\n"
        "Please answer in the following format:\n"
        "[1, 0, 1, 0, 0, 1, 0, 0, 1, 0] (1 indicates inclusion, 0 indicates exclusion. The order corresponds to the uses listed below.)\n"
        "Please provide only the list as your answer; do not offer any additional explanations or descriptions.\n"
        "If a determination cannot be made from the provided reference material, please answer [0, 0, 0, 0, 0, 0, 0, 0, 0, 0].\n"
        "Uses:\n"
        "1. Repayment of existing debt\n"
        "2. Capital expenditures\n"
        "3. Acquisitions\n"
        "4. Share buybacks\n"
        "5. Working capital\n"
        "6. Repayment of commercial paper\n"
        "7. General corporate purposes\n"
        "8. Short-term investments\n"
        "9. Increasing financial leverage\n"
        "10. Exchange offer\n"
        "The reference material is as follows:\n"
)
```

H	I	J	K	L	M	N	O	P	Q	R	S
Filename	acted Co	repay indebtedness	capital expenditure	acquisition	repurchase common stock	working capital	repay commercial paper	general corporate purpose	short-term investment	increase financial leverage	exchange offer
95_1581_5907_424B2_17.txt	The net pr	1	1	1	1	0	1	1	0	0	0

實驗結果與分析

- 第二階段關鍵發現：顯著的「資訊落差」
  - 現象：債券說明書中揭示的併購**意圖**，與公司年報中**實際揭露**的併購事件細節，存在巨大差距
  - 實例：微軟 (Microsoft)
    - 大型併購會揭露：如收購動視暴雪 (Activision Blizzard)和Nuance Communications
    - 中小型策略併購則未揭露：如收購網路安全公司 RiskIQ、物聯網安全公司 ReFirm Labs 等，在年報中均未找到詳細說明
  - 原因分析：
    1. 重大性原則 ：年報僅需揭露對財報有「重大」影響的事件
    2. 資訊彙總與省略：管理層在年報中可能選擇性地揭露資訊

結論

- 本研究揭示了不同監管文件間的資訊不對稱性
- 若僅依賴年報 (10-K) 來研究舉債併購，會因「資訊落差」而系統性地低估此類活動的真實規模，導致研究樣本偏差與結論失真
- 債券公開說明書 (Prospectus) 是捕捉企業真實財務意圖、構建更完整研究樣本的更可靠、更有效的途徑

未來研究

- 建立更完整的併購資料庫：整合重大事項報告 (8-K)、季度報告 (10-Q)、新聞稿或其他資料庫，以追蹤資金從意圖到實際運用的完整軌跡
- 拓展研究範疇：將分析對象從「企業實體併購」擴展至其他高風險資本支出 (如興建新廠房、技術升級)
- 探討資訊不對稱的市場意涵：研究市場對於在不同管道（說明書 vs. 年報）首次揭露的併購意圖，其反應與定價效率是否存在差異

Q&A