

LLM在財務之初探

LLM是什麼？

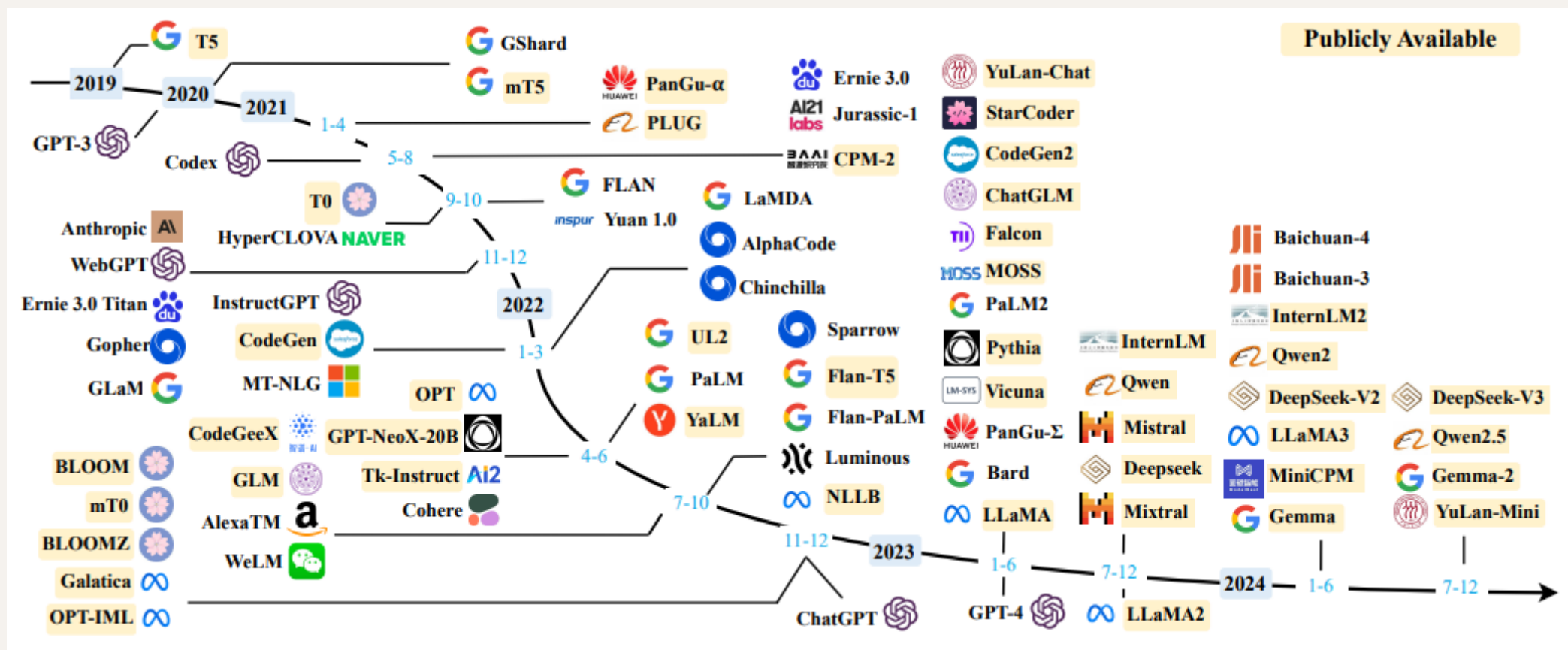
- 什麼是 LLM？
 - 可以「理解」和「產生」人類語言的 AI
 - 與傳統 NLP 模型不同，LLM 能產出新內容，不只是比對資料
 - 透過預測下一個字或詞來產生連貫語句
 - 先生成 w_1 ，根據 w_1 生成 w_2 ，根據 w_1 和 w_2 生成 w_3
 - 例如：ChatGPT、Claude、Gemini
- 為什麼叫「大型」？
 - 模型有上百億參數
 - 用了大量文字資料訓練

LLM是什麼？

- 跟 Siri、Google 助理有什麼不同？

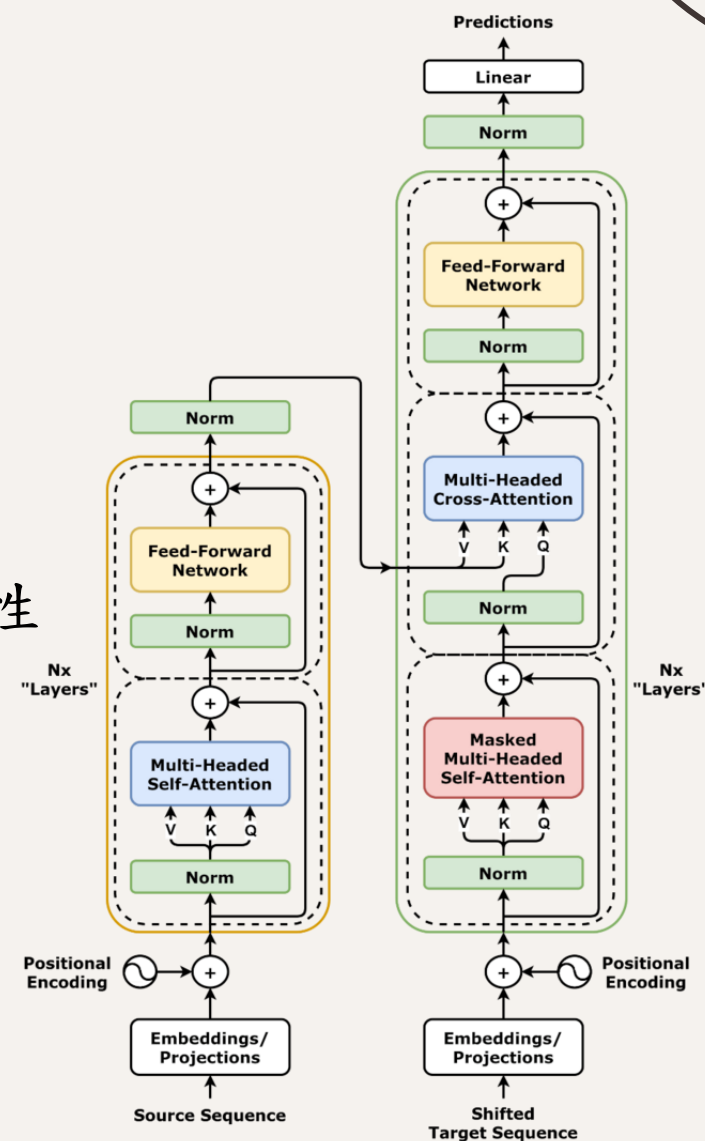
	Siri/Google助理	LLM
核心設計	指令式系統(基於關鍵字)	自然語言理解與生成(深度學習)
回答方式	從事先設定好的答案庫中抓出回應	動態生成回應，語句自然、邏輯連貫
語境理解	限於當前一句話	能理解上下文、多輪對話
創造力	幾乎沒有，只能執行固定任務	具備創造能力，可寫程式
資料來源	本地設定、雲端知識庫	基於大規模網路文本學習
可延伸性	功能受限於開發者設計的場景	可應用在客服、醫療、法律、金融等領域

模型的發展歷程



模型的發展歷程

- 2017：Transformer誕生
 - 能同時處理整個句子內容
 - 引入 Self-Attention 機制，理解句中詞語間的關聯性
 - 後來所有 LLM 幾乎都基於此架構發展
- Positional Encoding：讓模型知道「字詞順序」的重要性
- Self-Attention 機制
 - 模型能自動「關注」重要詞語之間的關係
 - 例子：「醫生建議病人吃藥」
 - 模型理解「建議」與「吃藥」有關
- Multi-Head Attention：從不同角度理解語意與語法



模型的發展歷程

- 2018：BERT(Bidirectional Encoder Representations from Transformers)
 - 參數量：340M
 - 由 Google 提出
 - 強調「理解」文本 → 能做分類、摘要等任務
 - 無法生成內容(只能判斷)
- 2020：GPT-3 (Generative Pretrained Transformer 3)
 - 參數量：175B
 - OpenAI 發表
 - 可流暢生成文章、翻譯、解釋程式
 - 掀起生成式 AI 熱潮

模型的發展歷程

- 2023：GPT-4、Claude、Gemini 等進一步推出
 - GPT-4：多語言理解更強、支援圖像輸入(GPT-4V)
 - Claude：重視 AI 安全性與人類價值對齊
 - Gemini：Google 推出多模態能力強的模型，結合影像、語音、影片處理
- 2023~2024：開源模型崛起
 - Llama(Meta)：免費開源、可地端部署
 - Mistral：高效、體積小、性能不輸 GPT 系列 → 開發者與企業喜愛

主流LLM模型介紹

模型	開發公司	特點
GPT-4	OpenAI	通用能力強、對話流暢、支援多語言
Claude	Anthropic	專注安全性與對話控制
Gemini	Google DeepMind	多模態能力(文字 + 圖片 + 聲音)
Llama	Meta	開源、企業內部部署彈性大
Mistral	Mistral AI	高效率、小體積開源模型

LLM的訓練流程

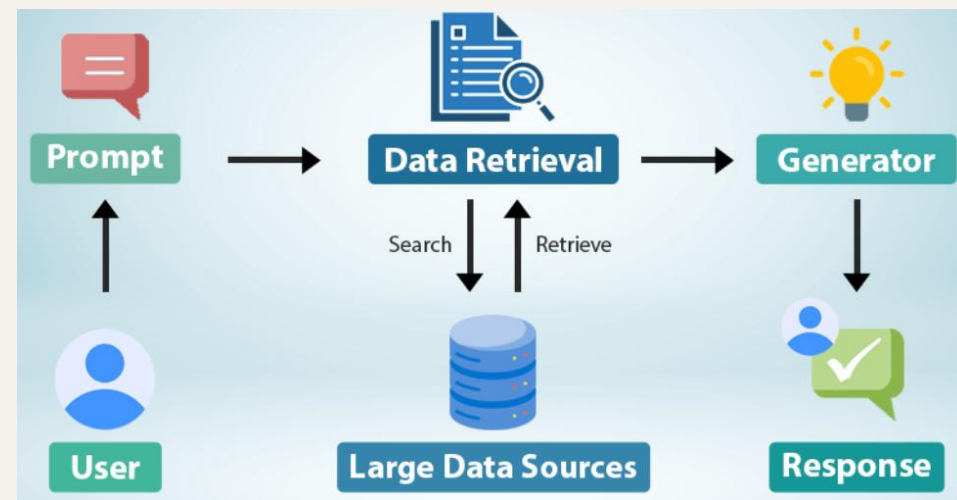
- 預訓練(Pretraining)
 - 用大量資料學語言規律(無標註資料)建構基礎模型
- 監督式微調(Supervised Fine-tuning)
 - 針對領域優化(如法律、醫療、金融)，根據下游任務，準備相對應的訓練資料集
 - 使用了包含問題和理想回答的成對資料，對模型進行進一步的微調，提升在特定任務上的性能
- 人類回饋強化學習(RLHF)
 - 人類給模型評分，調整產出品質

提升語言模型性能的方法

- 檢索增強生成(Retrieval-Augmented Generation, RAG)
- 提示工程(Prompt Engineering)
- 微調(Fine-Tuning)

Retrieval Augmented Generation, RAG

- 為了提升語言模型的性能和靈活性
- 結合內部知識和外部資訊的方法
 - 內部知識：模型本身參數所蘊含的知識，是基於模型架構設計和參數訓練過程所獲得的
 - 外部資訊：由外部資料庫中檢索提取並整合到提示中的資訊
- 語言模型能夠在生成回應時，不僅利用本身的參數知識，還能即時參考外部資源，進而提供更準確、更全面的答案



Prompt Engineering

- 提示 (Prompt) : 指的是對語言模型說的話，也就是輸入的指令、問題或上下文資訊
- 提示工程 (Prompt Engineering) : 設計有效指令來引導模型產生想要的結果
→ 像是對 AI 說話的藝術，你怎麼問，決定 AI 怎麼答。

為什麼需要 Prompt Engineering ?

- 相同模型，不同 prompt，可能產生完全不同的結果
- 適當設計 prompt 可以：
 - 提高正確率
 - 控制風格（格式、語氣）
 - 篩選資訊重點

Prompt Engineering的基本技巧

- 明確指定任務（例如："幫我總結這段文章"）
- 設定輸出格式（例如："列出三點原因"）
- 指定角色（例如："你是財務顧問..."）
- 使用例子（例如：few-shot learning）
- 分步驟指引（例如："先列出原因，再舉例"）

Fine-Tuning

- 在預訓練模型（Pretrained Model）基礎上進行再訓練，讓模型學會特定任務
- Fine-tuning vs Prompt Engineering

特性	Prompt Engineering	Fine-tuning
是否修改模型參數	否	是
資料需求	少量	中量以上
訓練成本	低	高（需GPU）
靈活性	高（可即時調整）	中（需重訓練）
精準度	較低	較高

雲端LLM vs 地端LLM

雲端 LLM

- 優點：
 - 不需自行訓練，直接接 API 使用
 - 快速迭代更新
- 缺點：
 - 資料需上傳 → 有隱私疑慮

地端 LLM

- 優點：
 - 資料不離開企業內部 → 高隱私與資安
 - 可客製化、融入企業內部系統
- 缺點：
 - 部署成本高，需要專業團隊維運
 - 模型更新、維護需投入更多資源

LLM的應用

- 企業客服系統 → 地端LLM
- 法規條文關聯 → 雲端/地端LLM
- 分析債券與財報資訊 → 雲端/地端LLM

企業客服系統

- 現有系統

- 用於追蹤客戶問題的處理流程
- 用有歷史諮詢案例資料



- 業務痛點－人力資源耗費

- 客戶人員在類似案例的客戶問題諮詢上消耗大量時間

- 目標

- 基於歷史諮詢案例，建立內部知識庫
- 開發檢索助手降低重複性工作，加速問題解決流程，提升客服工作效率



企業客服系統

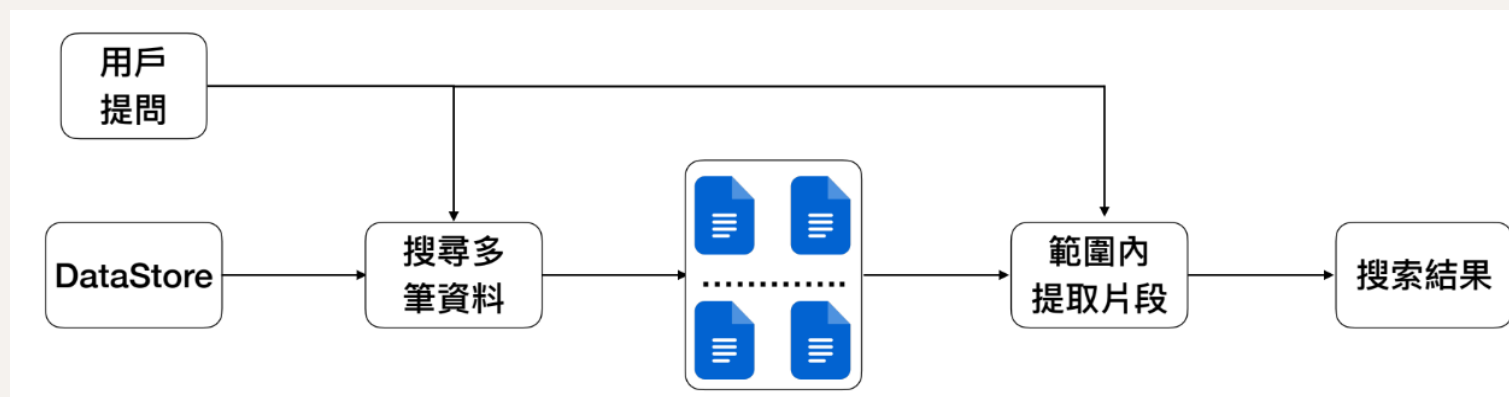
- 使用 Retrieval Augmented Generation (RAG) 技術
- Retriever
 - 透過檢索模型從知識庫檢索相關資訊
 - 強化輸出品質，提供可信度較高的回應
- Generator
 - 應用預訓練語言模型的自然語言生成能力
 - 整合檢索資訊與模型知識，針對使用者問題自行回覆

Retriever

- 詞袋模型：BM25
 - 根據關鍵字出現與否計算分數
 - 以字詞為單位，採用稀疏向量
 - 推理速度快，但無法理解同義詞與語意
- Bi-encoder：Chuxin-Embedding
 - 利用深度學習模型將 query 與 document 編碼向量，並以相似度檢索
 - 將 query 和 document 分別轉換成「語意向量（dense embeddings）」，再用向量相似度（如內積、餘弦相似度）來檢索
- Cross-encoder：LdlR-Qwen2-reranker-1.5B
 - 將 query 和 document 合併輸入模型中（如 [CLS] query [SEP] document），由 transformer 直接輸出一個相關性分數
 - 適合做 re-ranking

Retriever

- 採用BM25 + Cross-encoder的方式
- 第一階段：案件搜索
 - 目標：在大量的案件中進行初步篩選
- 第二階段：案件片段提取
 - 目標：保留關鍵資訊，過濾冗於內容



Retriever

執行申報註記設定的報表下載功能都無法下載，錯誤訊息「系統發生錯誤，請洽系統管理員。Operation is not valid on locked cells.」

使用者查詢



第一階段
檢索模型

公司歷史案件資料

案件 1

案件 2

·
·
·

案件
2199

·
·
·

案件 N

編號: 2199

摘要: [{銀行簡稱}] [{報表}/申報註記設定]報表下載功能異常出現Error

客戶提報內容:

【聯絡資訊】{客戶姓名}<{客戶mail}> 【問題類別】 【問題項目】 [{報表}/申報註記設定]報表下載功能異常出現Error 【問題說明】執行申報註記設定的報表下載功能(202310~202401)都無法下載，異常畫面請參考如下，但測試其他報表無此異常 【發生時間】 2024/03/01

回覆內容(1):

【案件狀態】對外發布 【案件類型】諮詢 【信件時間】 2024/3/4 (週一) 上午 09:22 【信件人員】{客戶姓名}<{客戶mail}> 【問題分析】執行申報註記設定的報表下載功能(202310~202401)都無法下載 【建議方案】請確認申報檔是否有鎖定儲存格 【處理說明】此錯誤訊息為EXCEL有儲存格被鎖定，請使用者確認EXCEL是否設定到鎖定，會導致系統無法解析下載，謝謝。 【預計交期】無

回覆時間(1): 2024/3/4 09:22

---(以下略過7筆回覆內容資料)---

回覆內容(8):

【案件狀態】申請結案 【案件類型】諮詢 【問題分析】執行申報註記設定的報表下載功能(202310~202401)都無法下載 針對系統的優化建議，若調整上傳的檔案是需要採用平台下載的，是否可以於上傳時作一卡控點檢驗，避免直到申報註記下載時才會察覺 【建議方案】請確認調整上傳所使用的檔案是否和平台產製格式相同 【處理說明】再請{內部員工姓名}協助與客戶確認此單是否可先結案，後續有優化評估結果再通知客戶，感謝! >3/12 MA回覆：感謝提供的建議，會列入產品優化建議中，再額外評估是否能透過調整上傳時的檔案比對確認範本合理性，謝謝。 【預計交期】無

回覆時間(8): 2024/6/6 15:51

Retriever

執行申報註記設定的報表下載功能都無法下載，錯誤訊息「系統發生錯誤，請洽系統管理員。Operation is not valid on locked cells.」

使用者查詢



第二階段
檢索模型

回覆內容(1):

【案件狀態】對外發布 【案件類型】諮詢 【信件時間】2024/3/4 (週一) 上午 09:22 【信件人員】{客戶姓名} <{客戶mail}>
【問題分析】執行申報註記設定的報表下載功能(202310~202401)都無法下載 【建議方案】請確認申報檔是否有鎖定儲存格 【處理說明】此錯誤訊息為EXCEL有儲存格被鎖定，請使用者確認EXCEL是否設定到鎖定，會導致系統無法解析下載，謝謝。 【預計交期】無

回覆時間(1): 2024/3/4 09:22

---(以下略過7筆回覆內容資料)---

回覆內容(8):

【案件狀態】申請結案 【案件類型】諮詢 【問題分析】執行申報註記設定的報表下載功能(202310~202401)都無法下載 針對系統的優化建議，若調整上傳的檔案是需要採用平台下載的，是否可以於上傳時作一卡控點檢驗，避免直到申報註記下載時才會察覺 【建議方案】請確認調整上傳所使用的檔案是否和平台產製格式相同 【處理說明】再請{內部員工姓名}協助與客戶確認此單是否可先結案，後續有優化評估結果再通知客戶，感謝! >3/12 MA回覆：感謝提供的建議，會列入產品優化建議中，再額外評估是否能透過調整上傳時的檔案比對確認範本合理性，謝謝。 【預計交期】無

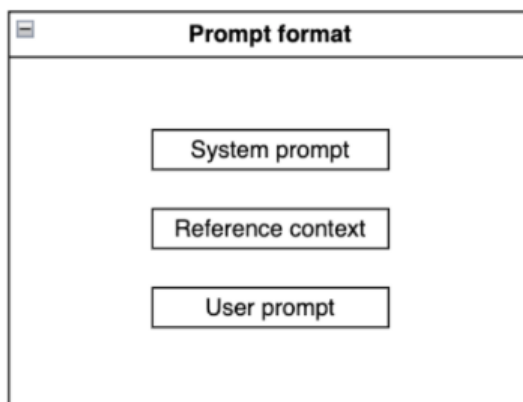
回覆時間(8): 2024/6/6 15:51



使用者



語言模型



Generator : TAME(Taiwan Mixture of Experts) – 8B

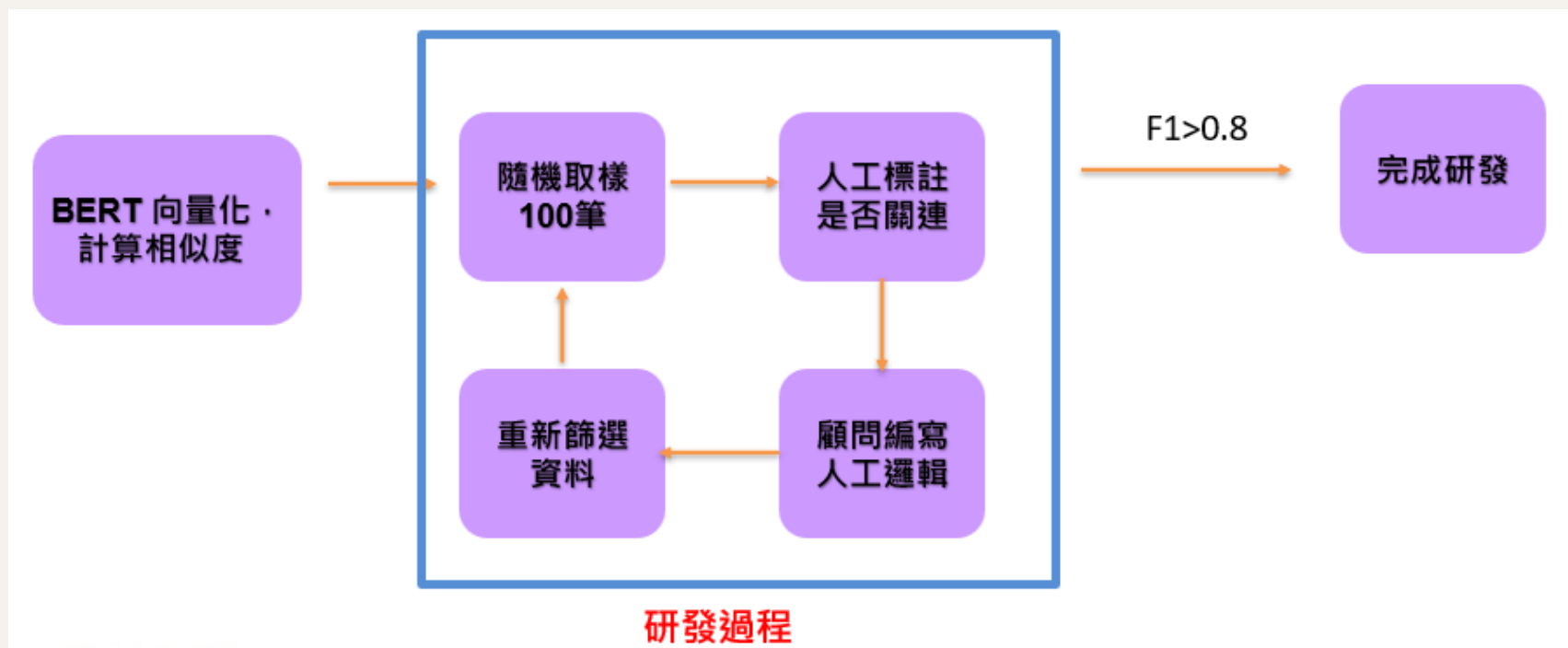
- 模型優勢
 - 基於Llama3.1進行微調
- 微調資料優勢，包含多家台灣企業資料
 - 具備繁體中文處理能力，同時適合處理台灣本土文化語境
- 符合需求的參數量
- 模型設定
 - 提示結構將遵循相關的標籤定義(<im_start>, <im_end>, system, user, assistant等)

法規條文關聯

- 金管會要求金融業依照所需遵循的法規，訂定相對應的內規
- 金融業需遵守的外部法規超過5000部，內部規範則超過800部
- 內規條文制定的邏輯會以外規條文為基礎來編寫內規條文，所以條文與條文之間存在語意的相似性
- 因為外規變動內規也會跟著變動，所以建構AI關聯模型，進而建立外規與內規之間的關聯表

法規條文關聯

- 現有的AI關聯模型
 - 沒有辦法retrain
 - 客戶回饋BERT的結果不準



法規條文關聯

- 現有資料

- 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：

一段

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。

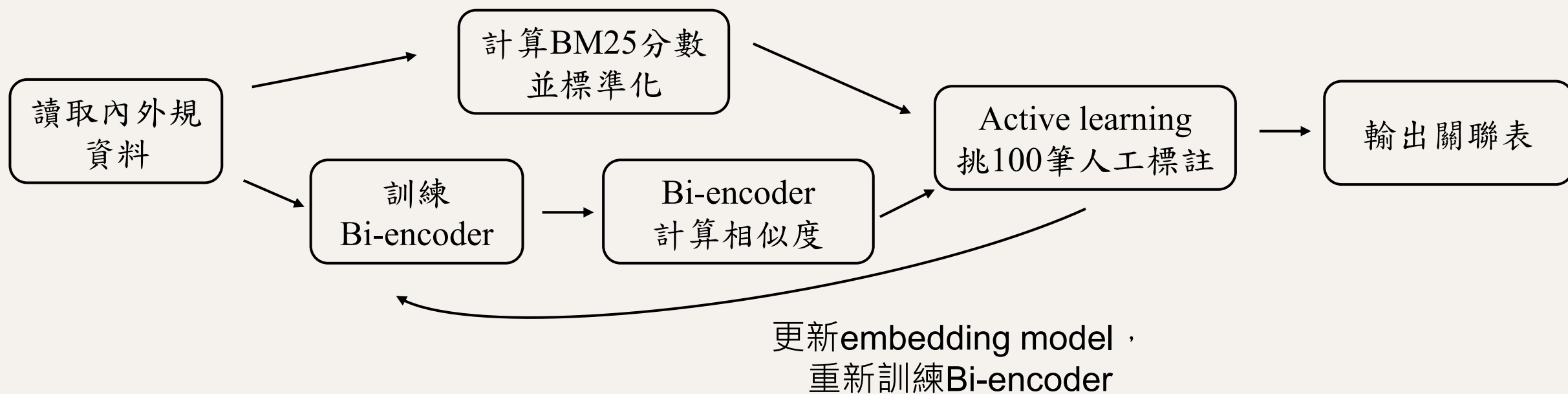
A	B	C	D	E	F	G
內規部	內規條	內規段	外規部	外規條	外規段	相似度
5	3	2	832	21	2	0.873646379
5	3	2	834	16	11	0.872581124
5	3	3	834	15	12	0.968946517
5	3	3	833	9	6	0.940811276
5	3	3	832	19	11	0.938995957
5	3	3	835	7	6	0.93470329
5	4	1	831	9	3	0.883011222
5	4	1	834	13	23	0.834484875
5	4	2	835	5	3	0.959981382
5	4	2	833	7	3	0.959981382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

法規條文關聯

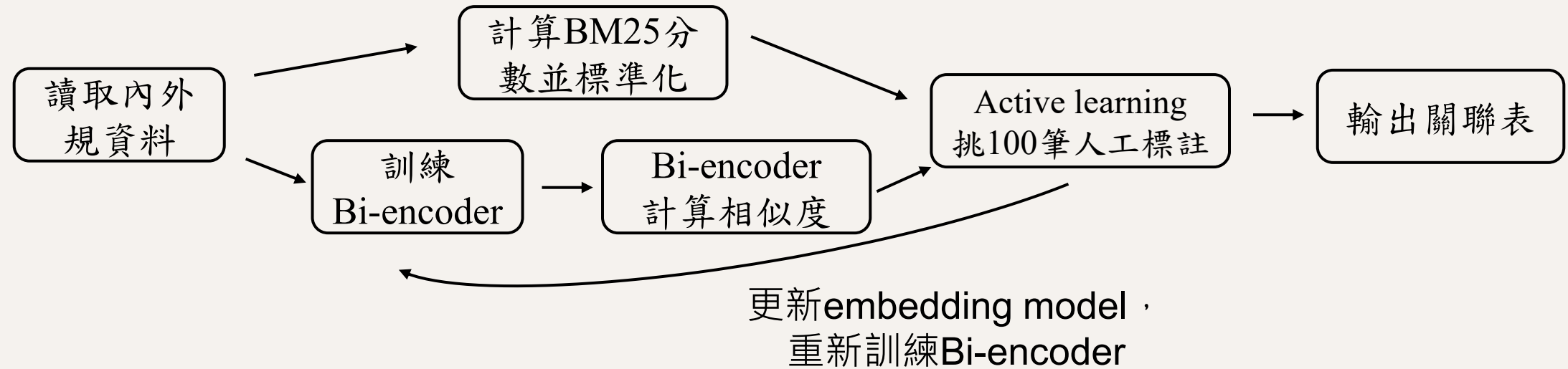
- 實驗目標
- 輸入:一段外規
- 輸出:所有有相關的內規段落

法規條文關聯 - 實驗流程

- 採用BM25 + Bi-encoder的方式並結合active learning，逐步提升法規段落配對的語意準確率

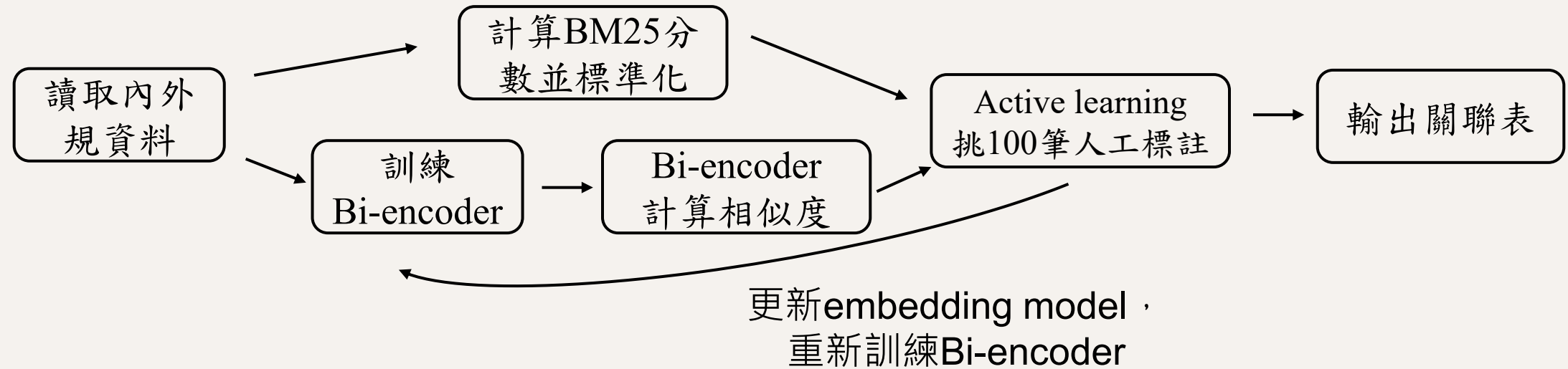


法規條文關聯 - 實驗流程



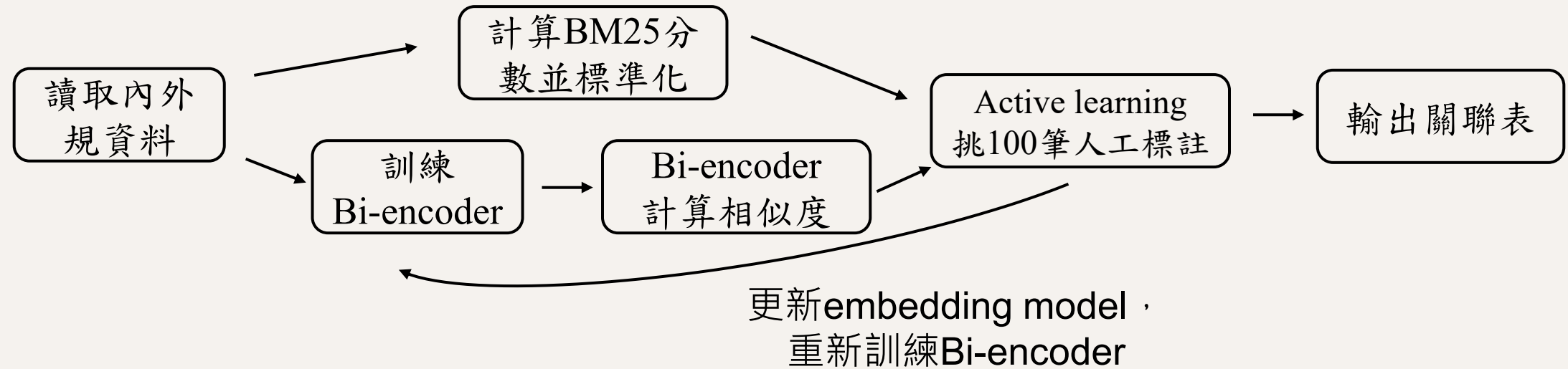
- 輸入資料：內規段落 × 外規段落：共 9440 組（正負樣本各半）
- 前處理：加入 query: 前綴
- Bi-encoder訓練
 - 模型：multilingual-e5-large
 - Loss function：BCE Loss，並將 cosine similarity 映射至 [0, 1]

法規條文關聯 - 實驗流程



- BM25 分數：基於 TF-IDF 模型計算分數，再經 min-max 標準化
- Bi-Encoder 相似度：將輸出 embedding 向量進行 cosine similarity 計算
- 當 BM25 與 Bi-Encoder 分數差距大，代表模型存在不確定性 → 適合進行人工驗證

法規條文關聯 - 實驗流程



- Active learning挑選策略: 計算 $|\text{BM25_score} - \text{cosine_similarity}|$ ，取最大前 100 筆
- 人工標註：標記是否真正相似，作為 retrain 的新增資料
- 優點：無需全標註，聚焦「最不確定」的樣本，提高標註效率與效益

分析債券與財報資訊

- 步驟1：從債券公開說明書中確認債券發行目的
- 步驟2：從公司財報中確認資金實際用途
 - 以acquisition為例

步驟1：確認債券發行目的

- 先從債券說明書中找到use of proceeds的段落

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 general corporate purposes.

- 透過語言模型(gemini-1.5-flash)標記出債券發行目的

```
prompt = (  
    "以下是公司發行債券的十種用途，請根據參考資料判斷公司是否包含這些用途。\\n"  
    "請按照以下格式回答：\\n"  
    "[1, 0, 1, 0, 0, 1, 0, 0, 1, 0] (1 表示包含, 0 表示不包含，順序對應下列用途)\\n"  
    "請只提供回答清單，不要提供任何額外解釋或敘述\\n\\n"  
    "如果無法根據提供的參考資料判斷，請回答[0, 0, 0, 0, 0, 0, 0, 0, 0, 0]"  
    "用途：\\n"  
    "1. 償還現有債務\\n"  
    "2. 資本支出\\n"  
    "3. 收購\\n"  
    "4. 庫藏股回購\\n"  
    "5. 營運資金\\n"  
    "6. 償還商業票據\\n"  
    "7. 一般公司用途\\n"  
    "8. 短期投資\\n"  
    "9. 提高財務槓桿\\n"  
    "10. 交換性質\\n\\n"  
    "參考資料如下：\\n"  
)
```

H	I	J	K	L	M	N	O	P	Q	R	S
Filename	acted Cor	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

步驟2：確認資金實際用途

- 將財報分段後，使用語言模型找出公司發債資金的實際用途
 - 以 acquisition 為例

```
1 # 設定 Gemini Prompt
2 prompt_template = """
3 請判斷以下公司財報的段落是否提及 Acquisition（收購/併購）。
4 如果段落包含公司有進行 Acquisition，請回答 1；
5 如果沒有提到 Acquisition，請回答 0。
6
7 段落內容：
8 "{content}"
9
10 請只回應 1 或 0，不要加任何其他文字。
11 """
```

```
{
  "Section": "Item 6.",
  "Content": "Selected Financial Data The information set forth below for the five years ended September 29, 2018 , is not necessarily indicative of results of future operations, and should be read in conjunction with Part II, Item 7, Management's Discussion and Analysis of Financial Condition and Results of Operations and the consolidated financial statements and related notes thereto included in Part II, Item 8 of this Form 10-K to fully understand factors that may affect the comparability of the information presented below (in millions, except number of shares, which are reflected in thousands, and per share amounts). 2018 2017 2016 2015 2014 Net sales $ 265,595 $ 229,234 $ 215,639 $ 233,715 $ 182,795 Net income $ 59,531 $ 48,351 $ 45,687 $ 53,394 $ 39,510 Earnings per share: Basic $ 12.01 $ 9.27 $ 8.35 $ 9.28 $ 6.49 Diluted $ 11.91 $ 9.21 $ 8.31 $ 9.22 $ 6.45 Cash dividends declared per share $ 2.72 $ 2.40 $ 2.18 $ 1.98 $ 1.82 Shares used in computing earnings per share: Basic 4,955,377 5,217,242 5,470,820 5,753,421 6,085,572 Diluted 5,000,109 5,251,692 5,500,281 5,793,069 6,122,663 Total cash, cash equivalents and marketable securities $ 237,100 $ 268,895 $ 237,585 $ 205,666 $ 155,239 Total assets $ 365,725 $ 375,319 $ 321,686 $ 290,345 $ 231,839 Non-current portion of term debt $ 93,735 $ 97,207 $ 75,427 $ 53,329 $ 28,987 Other non-current liabilities $ 45,180 $ 40,415 $ 36,074 $ 33,427 $ 24,826 Apple Inc. | 2018 Form 10-K | 21",
  "Length": 1445,
  "Acquisition": 0
},
```

LLM的挑戰與影響

- 幻覺（Hallucination）問題
 - 它只靠統計學預測下一個詞，沒有真正的知識記憶
 - 加入知識檢索輔助，使用RAG 技術
- 資料偏誤與歧視
- 隱私與資安
- 該怎麼面對 LLM？
 - 把它當成工具，學會問對問題（Prompt Engineering）