

基於零樣本檢索增強生成之大型語言模型性能優化方法: 以企業客服應用為例

Optimization Methods for Zero-Shot Retrieval-Augmented Generation in Large Language Models:
A Case Study in Enterprise Customer Service Applications

研究生: 汪文豪 (Wun-Hao, Wang)

指導教授: 戴天時 (Tian-Shyr, Dai)

口試日期: 2025/01/10

綱要

- 研究背景與動機
- 專案研究: MetaEdge LLM 客服問答系統
- 實驗設計
- 實驗結果與分析
- 結論與未來工作

研究背景與動機

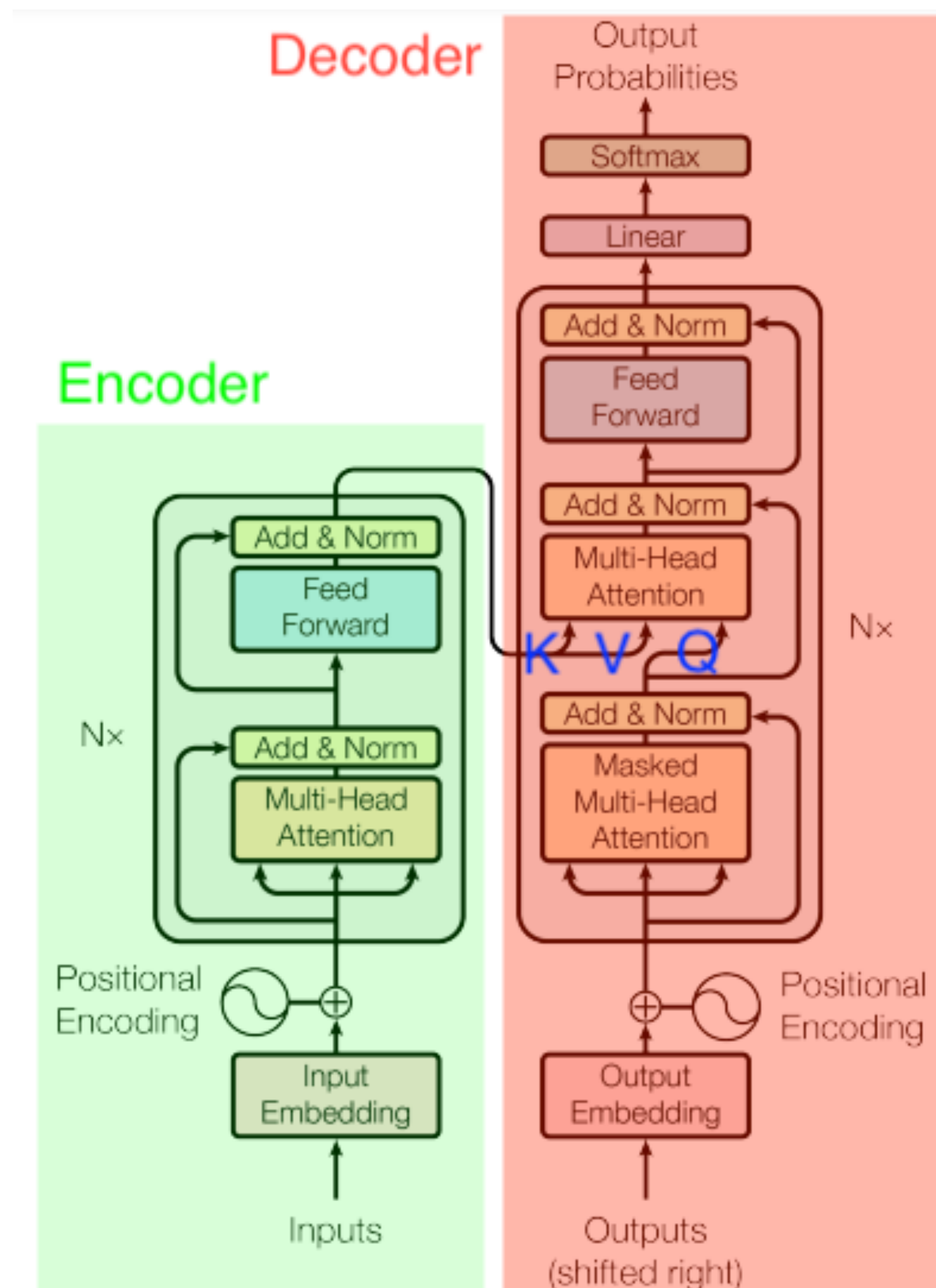
LLMs的革新浪潮

2017 - Transformer架構問世

- 引入突破性的注意力機制 (Attention Mechanism)
- 奠定現代大型語言模型(Large Language Models, LLMs)的核心技術基礎

2022 - ChatGPT引領AI革命

- 卓越的語言理解與生成能力
- 多樣化任務處理能力
- 在企業應用中帶來更多自動化的新可能



企業應用LLMs的挑戰

現有方案的局限性

- 第三方AI服務的限制
 - 資料隱私風險
 - 成本效益考量
- 自行開發與維護LLM的困境
 - 巨額的資源需求
 - 訓練資料與技術人力需求

企業應用LLMs的挑戰

技術整合策略

- 第三方AI服務的限制
 - 資料隱私風險
 - 成本效益考量
- 自行開發與維護LLM的困境
 - 巨額的資源需求
 - 訓練資料與技術人力需求



- 開源預訓練模型
- 基於產學合作案例進行需求分析並進行實驗，提供可落地的最佳零樣本(Zero-Shot)應用實踐

專案研究: MetaEdge LLM 客服問答系統

專案研究 - MetaEdge LLM 客服問答系統

現況分析與需求定義



現有系統 - 問題追蹤系統 (Mantis)

- 用於追蹤客戶問題的處理流程
- 擁有歷史諮詢案例資料

業務痛點 - 人力資源耗費

- 客服人員在類似的客戶問題諮詢上消耗大量時間

專案目標

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

專案研究 - MetaEdge LLM 客服問答系統

現況分析與需求定義



現有系統 - 問題追蹤系統 (Mantis)

- 用於追蹤客戶問題的處理流程
- 擁有歷史諮詢案例資料

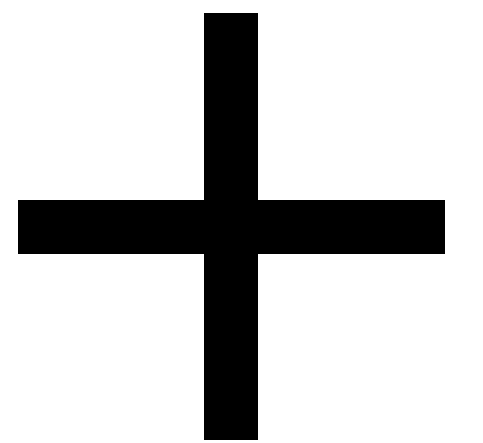
業務痛點 - 人力資源耗費

- 客服人員在類似的客戶問題諮詢上消耗大量時間

專案目標

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

文本
檢索



文本
總結

專案研究 - MetaEdge LLM 問答系統

Retrieval Augmented Generation, RAG

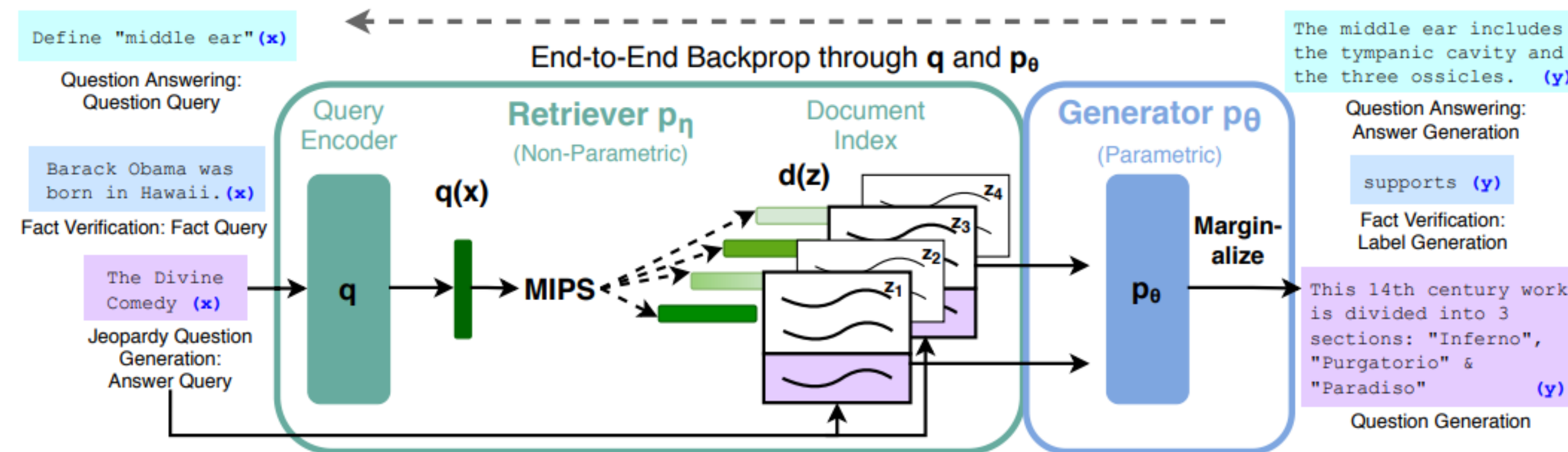


Figure 1: Overview of our approach. We combine a pre-trained retriever (*Query Encoder* + *Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query x , we use Maximum Inner Product Search (MIPS) to find the top-K documents z_i . For final prediction y , we treat z as a latent variable and marginalize over seq2seq predictions given different documents.

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，針對使用者問題進行回應

Mantis系統資料概況分析

Mantis_RawData_案件

欄位名稱	說明
編號	問題追蹤單號
摘要	該問題的簡要說明
客戶提報內容	記錄該客戶聯繫方式、問題內容、問題類型等

Mantis_RawData_明細

欄位名稱	說明
編號	問題追蹤單號
回覆內容	該問題追蹤單的處理回覆，包括內部記錄、客戶回覆、外部發布等
回覆時間	回覆內容的時間戳記，以 YYYY/M/d h:mm 格式儲存

- 回覆時間範圍: 2022年1月5日 - 2024年7月12日
- 歷史案件: 974筆
- 明細中的追蹤回覆: 9216筆
- 資料關聯: 一筆案件對應多筆明細

專案研究 - MetaEdge LLM 問答系統

核心挑戰

- 1. 資料涉及客戶隱私與缺乏訓練資料
- 2. 資料樣態為流水帳式描述
- 3. 如何應用於文本檢索與總結

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

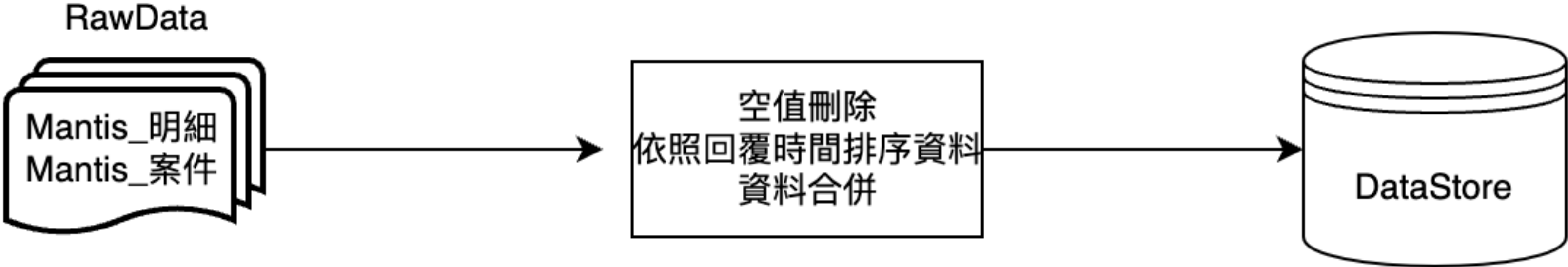
---(以下略過7筆回覆內容資料)---

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

文本前處理策略



資料整合策略

- **流水帳資料整合:** 將每個案件對應的多筆回覆內容，依時序合併為單一完整對話紀錄，從而更易於檢索並維持問題處理的時間脈絡與資訊連貫性。
- **檢索特徵建構:** 各案件具有「摘要」、「客戶提報內容」與「回覆內容」等基本特徵。為了評估其特徵整合對檢索性能的影響，另外建構了一個將三者依序合併的綜合特徵 (摘要 + 客戶提報內容 + 回覆內容)，以進行比較實驗。

各類特徵的平均字串長度

欄位名稱	字串平均長度
摘要	29.75
客戶提報內容	231.65
回覆內容	2130.99
摘要 + 客戶提報內容 + 回覆內容	2394.39

專案研究 - MetaEdge LLM 問答系統

Retrieval Augmented Generation, RAG

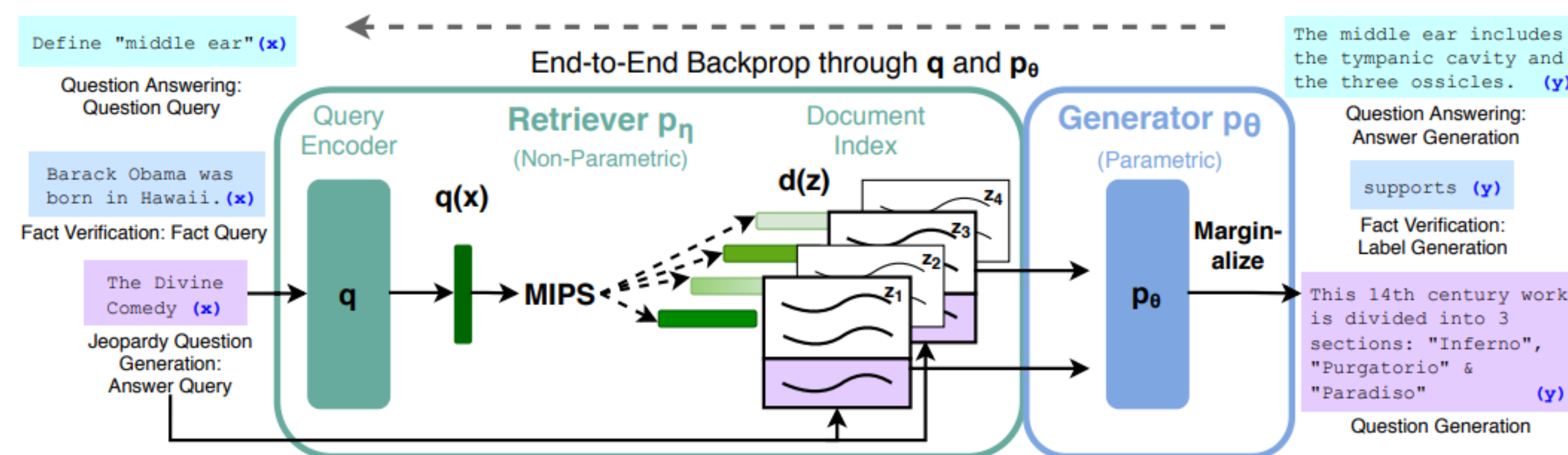


Figure 1: Overview of our approach. We combine a pre-trained retriever (*Query Encoder* + *Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query x , we use Maximum Inner Product Search (MIPS) to find the top-K documents z_i . For final prediction y , we treat z as a latent variable and marginalize over seq2seq predictions given different documents.

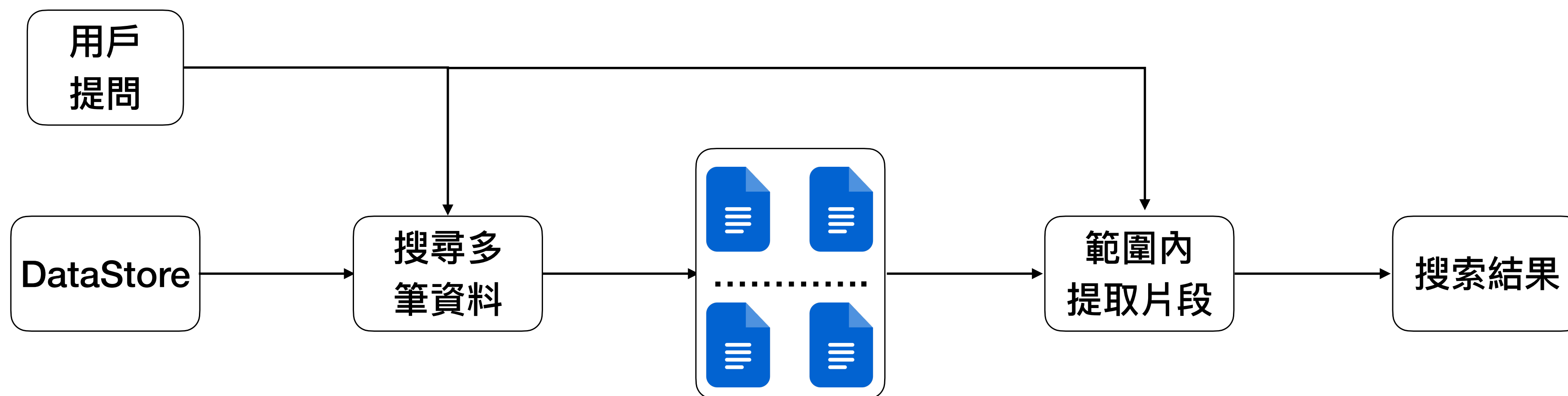
Retriever

- 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，針對使用者問題進行回應

文本檢索系統



參考Thakur 等人(2021)的研究發現[2]，BM25 + Cross-Encoder 在零樣本檢索中表現最佳。

整體的檢索系統以此想法作為基礎：

- **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，過濾冗餘內容

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

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

案件 1

案件 2

·
·
·

案件
2199

·
·
·

案件 N

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)---

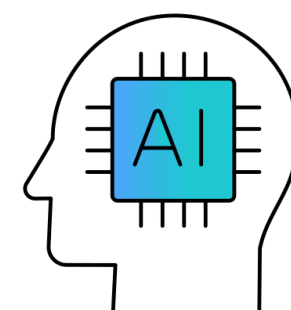
回覆內容(8):

【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

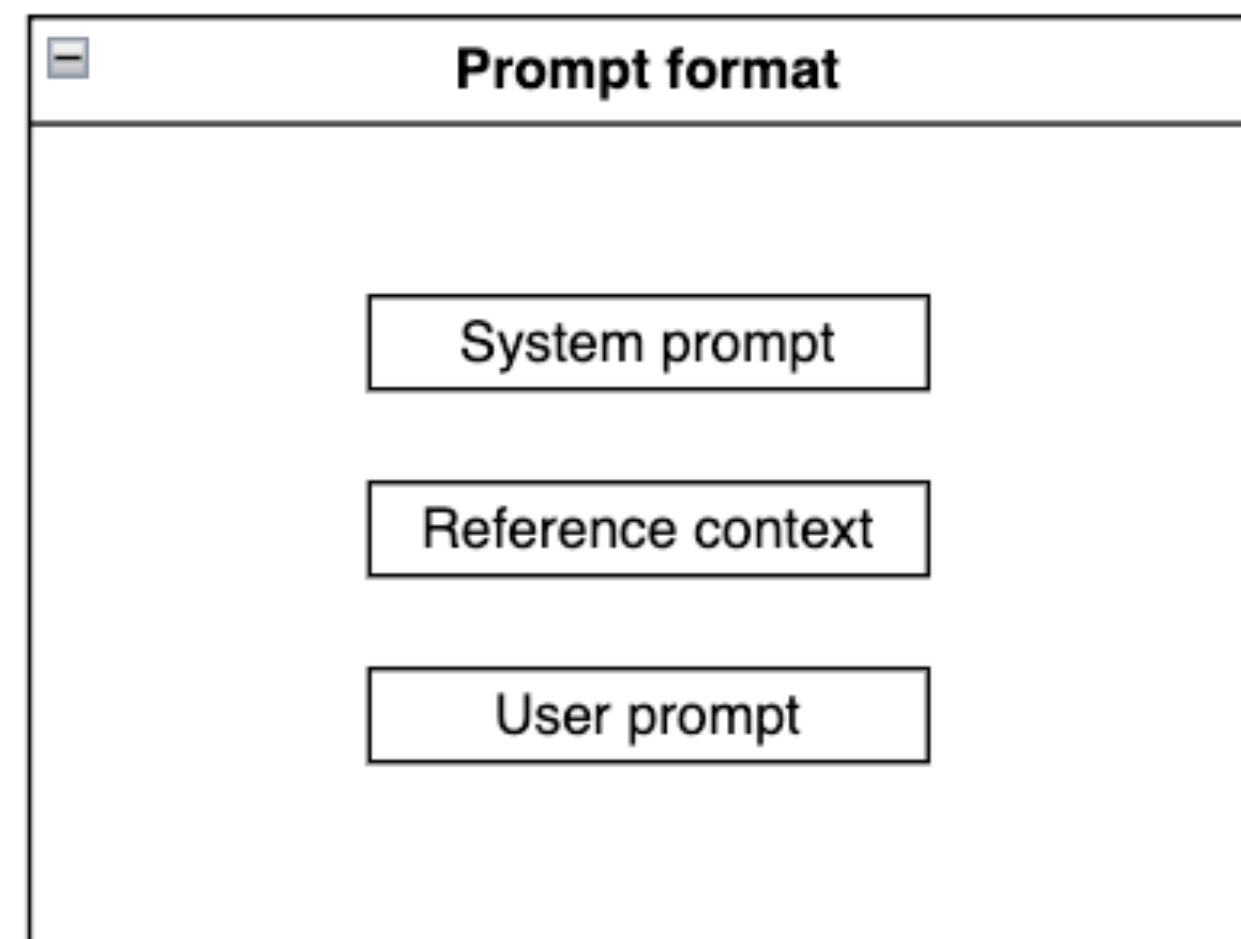
回覆時間(8): 2024/6/6 15:51



使用者



語言模型



檢索模型選擇說明

- 基於相似度計算方式的不同分為三類
 - 詞袋模型: BM25
 - Bi-Encoder: Chuxin-Embedding (簡稱"Chuxin")
 - Cross-Encoder: LdIR-Qwen2-reranker-1.5B (簡稱"Qwen")
- 模型選擇依據
 - HuggingFace MTEB Leaderboard
 - 針對檢索任務性能表現進行評估

專案研究 - MetaEdge LLM 問答系統

Retrieval Augmented Generation, RAG

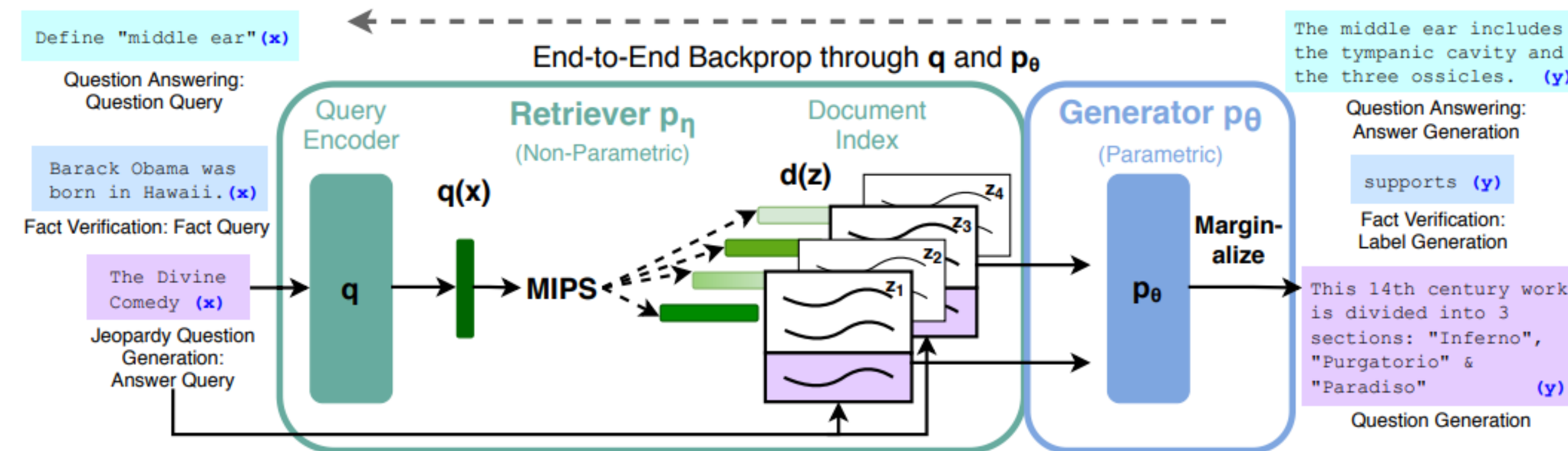


Figure 1: Overview of our approach. We combine a pre-trained retriever (*Query Encoder* + *Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query x , we use Maximum Inner Product Search (MIPS) to find the top-K documents z_i . For final prediction y , we treat z as a latent variable and marginalize over seq2seq predictions given different documents.

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，針對使用者問題進行回應

生成模型選擇說明

TAME (Taiwan Mixture of Experts) - 8B

選擇依據

- **基礎模型優勢**
基於Llama 3.1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)
- 遵循預設停用詞("USER:"、"ASSISTANT:"、"<|im_end|>"、"<|eot_id|>"和"<|end_of_text|>")。

Token採樣生成的相關參數配置如下:

- 溫度(temperature): 0.1
- Top-k: 50
- 最大生成長度: 1024

實驗設計

- 文本檢索
- 文本生成
- 延伸實驗

文本檢索實驗

第一階段檢索: 案件搜索

評估模型

- BM25
- Chuxin (Traditional/Simplified)
- Qwen (Traditional/Simplified)

檢索特徵

- 摘要
- 客戶提報內容
- 回覆內容
- 摘要 + 客戶提報內容 + 回覆內容

評估指標

- Accuracy
- Avg Query Time (s)

文本檢索實驗

第二階段檢索: 案件片段提取

評估模型

- BM25
- Chuxin (Traditional/Simplified)
- Qwen (Traditional/Simplified)

檢索特徵

- 基於第一階段實驗最佳的檢索結果，檢索該案件對應的「回覆內容」特徵，並以換行字元進行切分
- 潛在候選片段的平均數量: 8.85個，最大為37個
- 參考片段的平均數量: 1.16個，最大為6個

評估指標

- **Precision@N**
該值愈高表示檢索結果包含較少冗餘內容
- **Recall@N**
該值愈高表示檢索結果對參考片段的覆蓋越完整
- **F1-Score@N**
- **Compression Rate**
衡量檢索內容相較於原「回覆內容」的文字精簡比例
- **Avg Query Time (s)**

實驗設計

- 文本檢索
- 文本生成
- 延伸實驗

文本生成實驗

提示詞模板設計

- 使用TAME模型定義的相關標籤[3]以及停用詞[4]
- 三個核心要素
 - 系統提示
 - 檢索內容
 - 使用者查詢
- 固定使用者查詢，系統性測試系統提示與檢索內容的影響

```
<|im_start|>system  
{系統提示}<|im_end|>  
  
<|im_start|>context  
[主要問題]  
  
{摘要}  
  
{客戶提報內容}  
  
[過往解決流程]  
  
{回覆內容}<|im_end|>  
  
<|im_start|>user  
{使用者查詢}<|im_end|>  
  
<|im_start|>assistant
```

[3] <|im_start|>, <|im_end|>, system, user, assistant

[4] " USER:"、" ASSISTANT:"、" <| im_end| >"、" <| eot_id| >" 和" <|end_of_text|>"

文本生成實驗

評估指標

- ROUGE-1 (Recall)
- ROUGE-L (Recall)
- BLEU Score
- BERT Score
- LLM Score
- Inference Time

文本生成實驗

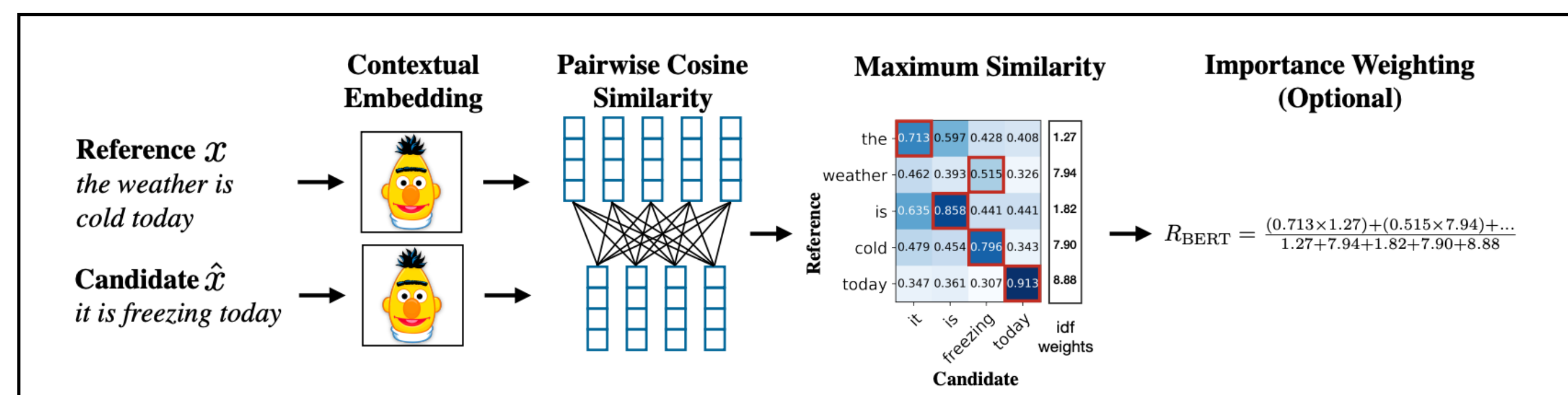
評估指標

- **ROUGE-1 (Recall):** 衡量候選文本與參考文本間使用Token的重疊程度
- **ROUGE-L (Recall):** 評估候選文本與參考文本間的Longest Common Subsequence
- **BLEU Score:** 針對短且使用Token重疊高的候選文本施加懲罰
- BERT Score
- LLM Score
- Inference Time

文本生成實驗

評估指標

- ROUGE-1 (Recall)
- ROUGE-L (Recall)
- BLEU Score
- **BERT Score: 捕捉語義相似度**
- LLM Score
- Inference Time



本研究採用Chuxin計算 BERT Score

文本生成實驗

評估指標

- ROUGE-1 (Recall)
- ROUGE-L (Recall)
- BLEU Score
- BERT Score
- **LLM Score**
- Inference Time

LLM Score計算方法

- 採用經人類意圖對齊後的TAME-8B作為評估模型
- 依照給定的提示模板，模型依其提示輸入，判斷在「是」與「否」兩個Token上的輸出機率
- 兩者間取最大值作為最終的分類結果。
- 輸出Token對應的分數：「是」為1，「否」為0

<|im_start|>system

You are an AI assistant called Twllm, created by TAME (TAiwan Mixture of Expert) project.<|im_end|>

<|im_start|>context

參考句: {參考答案}

預測句: {候選答案}<|im_end|>

<|im_start|>user

請評估在context中的「參考句」與「預測句」其句義是否相似？

請直接回答「是」或「否」。
<|im_end|>

<|im_start|>assistant

文本生成實驗

實驗方法

系統提示	檢索內容	生成策略
- TAME 預設配置 - 明確角色能力 - 無系統提示	- 基於案件檢索結果 - 基於案件片段檢索結果 - 基於時序排序的案件片段檢索結果	- 基礎生成策略 - 推理引導策略 - 採樣策略

文本生成實驗

實驗方法

系統提示

- TAME 預設配置
- 明確角色能力
- 無系統提示

檢索內容

- 基於案件檢索結果
- 基於案件片段檢索結果
- 基於時序排序的案件片段檢索結果

生成策略

- 基礎生成策略
- 推理引導策略
- 採樣策略

文本生成實驗

系統提示

- **TAME 預設配置 (簡稱"TAME")**

雖未明確指出其模型角色，但模型可能在微調階段中作為前綴提示詞參與訓練，從而將相關能力內化於模型回應中。

- **明確角色能力 (簡稱"Expert")**

明確定義AI助理的特徵，強調其知識能力與回應風格。

- **無系統提示 (簡稱"Without System Prompt")**

作為對照組，評估系統提示的實際效果。

TAME

<|im_start|>system

You are an AI assistant called Twllm, created by TAME (TAiwan Mixture of Expert) project.

回答時，請遵循以下原則:

1. 僅從提供的context中的內容進行回答
2. 回答必須簡潔直接，只提供與user詢問高度相關的內容<|im_end|>

Expert

<|im_start|>system

你是一位知識豐富、高效且直接的AI助理。

回答時，請遵循以下原則:

1. 僅從提供的context中的內容進行回答
2. 回答必須簡潔直接，只提供與user詢問高度相關的內容<|im_end|>

文本生成實驗

實驗方法

系統提示

- TAME 預設配置
- 明確角色能力
- 無系統提示

檢索內容

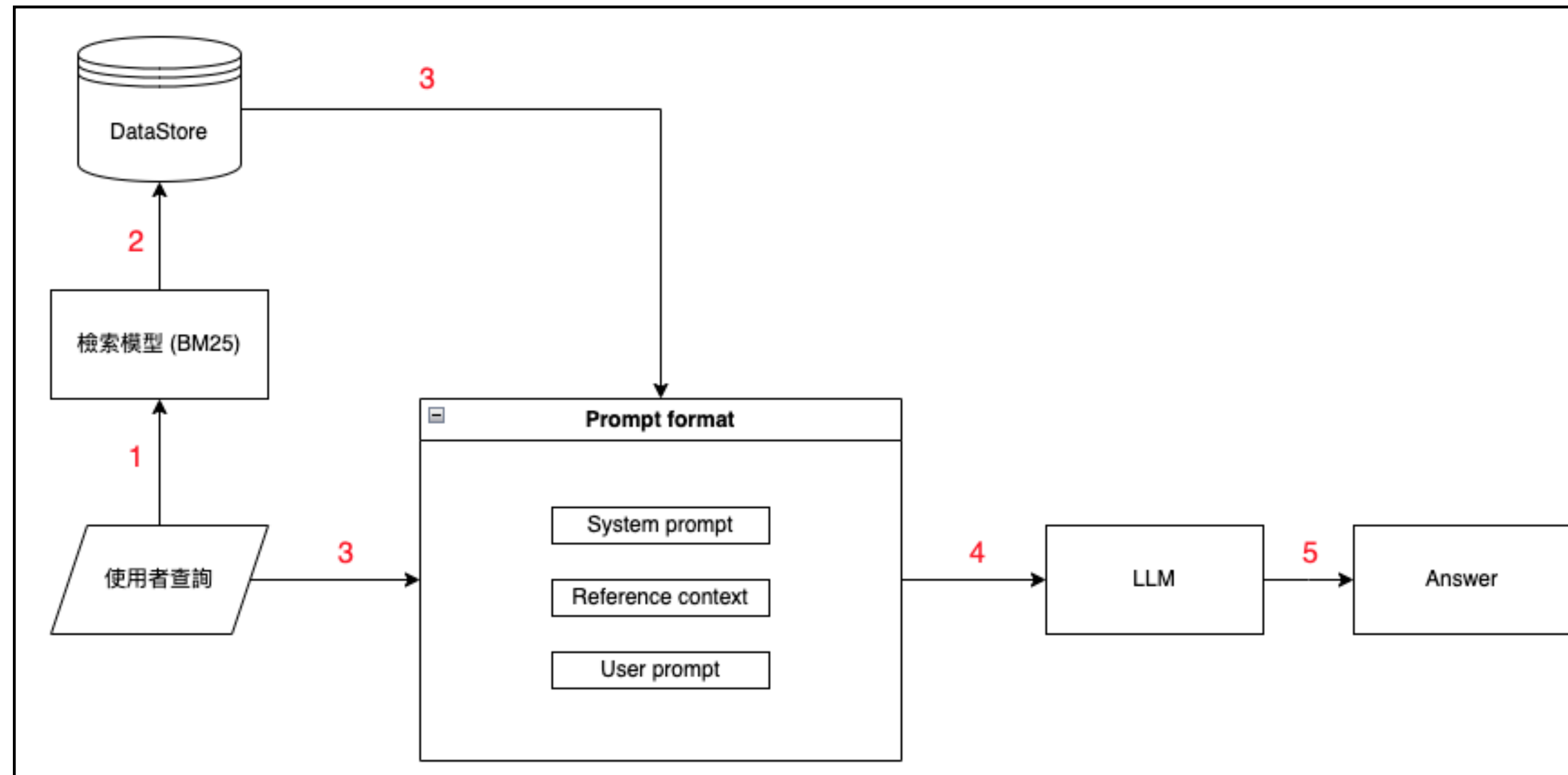
- 基於案件檢索結果
- 基於案件片段檢索結果
- 基於時序排序的案件片段檢索結果

生成策略

- 基礎生成策略
- 推理引導策略
- 採樣策略

文本生成實驗

檢索方法

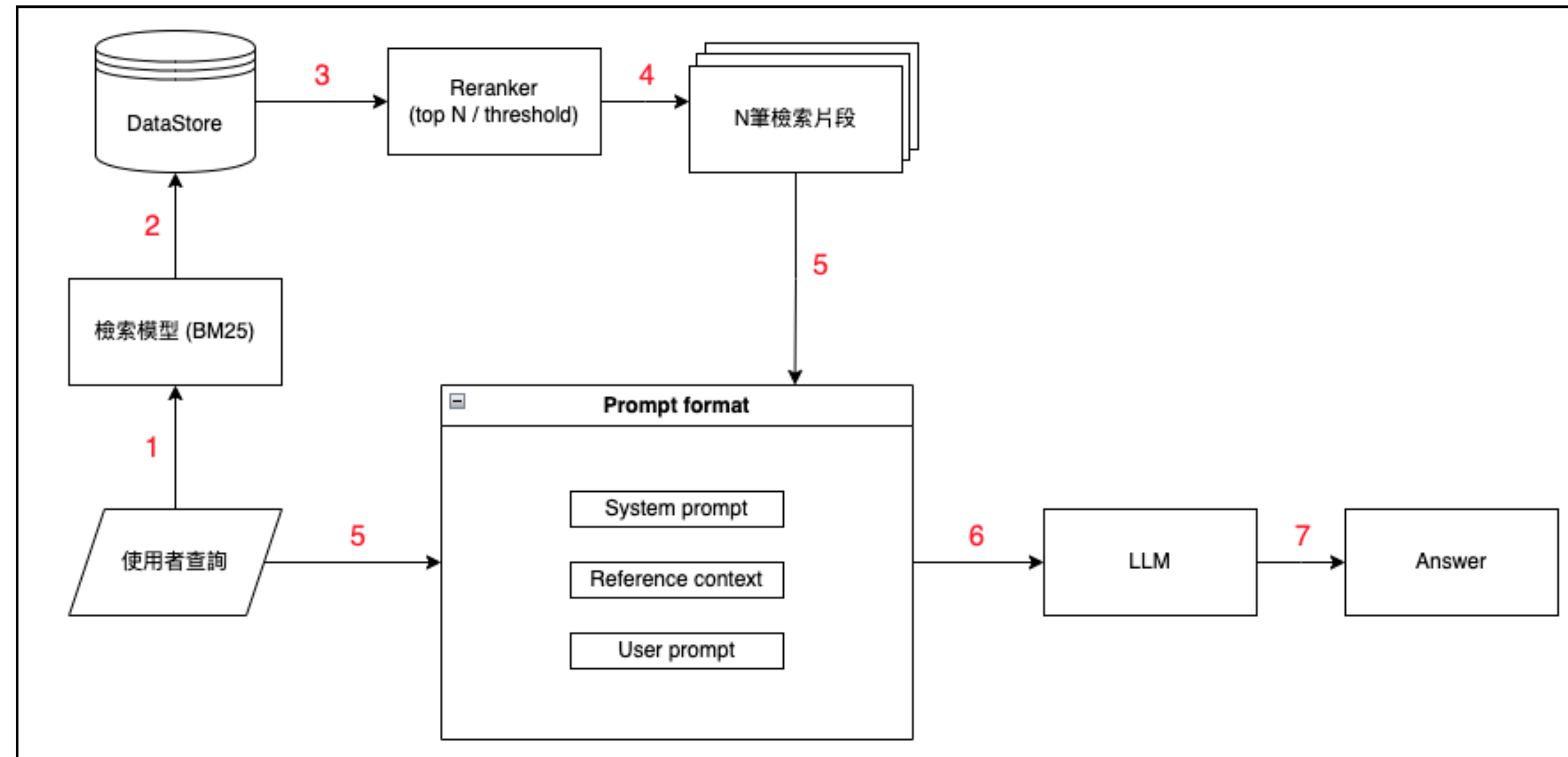


- 基於案件檢索結果的生成實驗 (簡稱"Document"):

基於第一階段檢索實驗中表現最佳的檢索配置作為對照組。

文本生成實驗

檢索方法

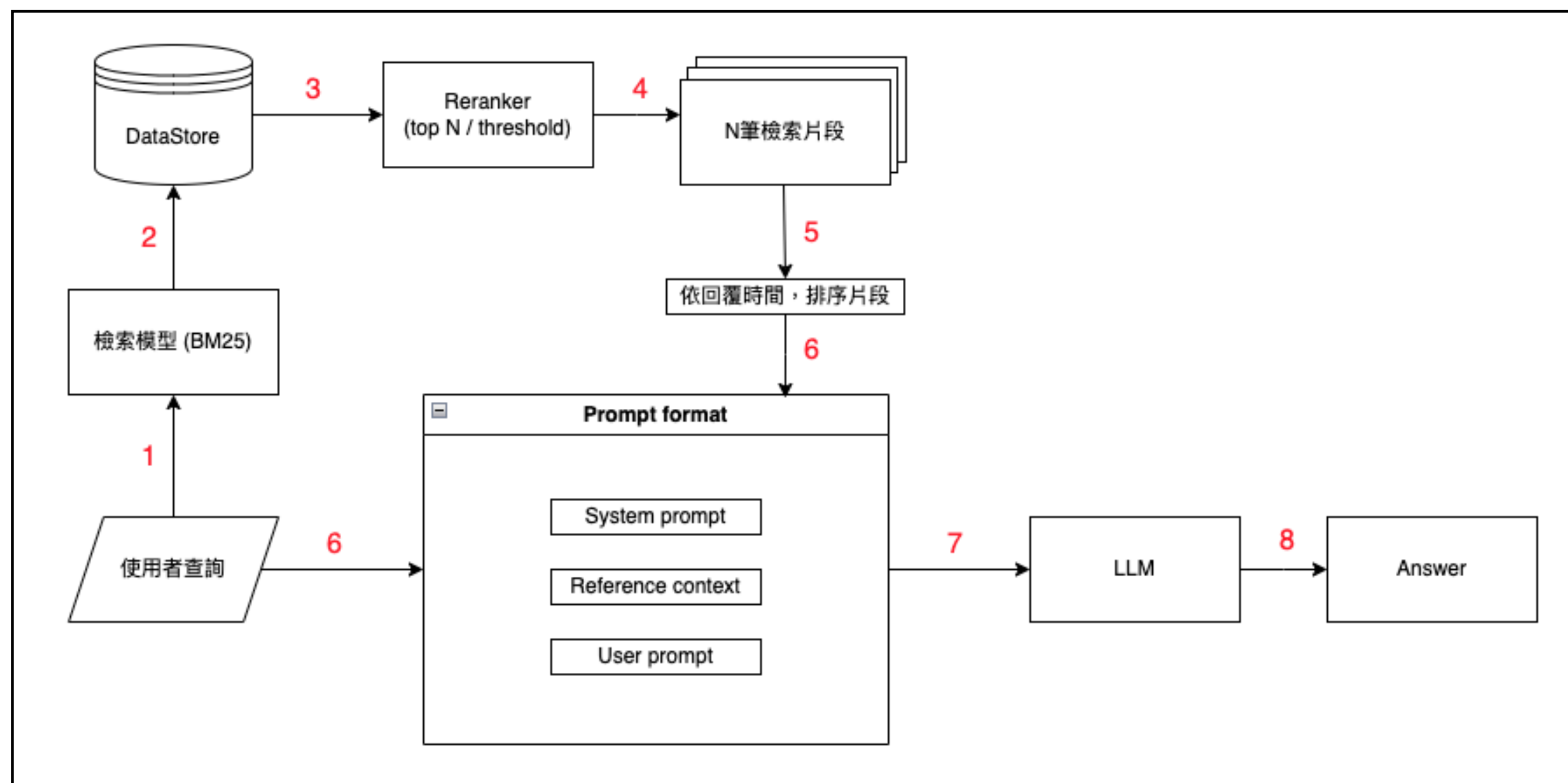


- 基於案件片段檢索結果的生成實驗 (簡稱"Passage"):

基於第二階段檢索實驗中表現最佳的配置，評估篩選資訊後的生成性能。

文本生成實驗

檢索方法



- **基於時序排序後的案件片段檢索結果的生成實驗 (簡稱"Time Sorted"):**

避免第二階段檢索過程中破壞上下文時序關係，針對檢索結果依該特徵對應的「回覆時間」進行排序，測試位置資訊對生成效果的影響。

文本生成實驗

實驗方法

系統提示

- TAME 預設配置
- 明確角色能力
- 無系統提示

檢索內容

- 基於案件檢索結果
- 基於案件片段檢索結果
- 基於時序排序的案件片段檢索結果

生成策略

- 基礎生成策略
- 推理引導策略
- 採樣策略

文本生成實驗

生成策略

- 基礎生成策略 (簡稱"Base")

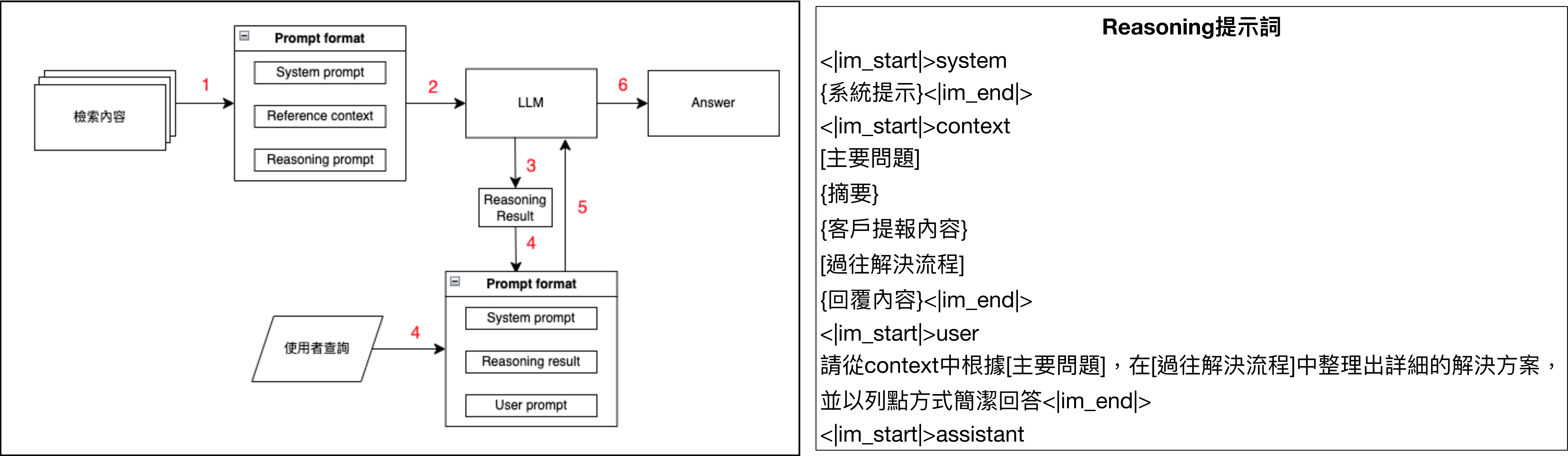
作為對照組，此策略直接將檢索結果與用戶查詢作為輸入，由模型一次性生成答案，不使用任何優化技巧。

文本生成實驗

生成策略

- 推理引導策略 (簡稱"Reasoning")

先針對檢索文本進行摘要總結，再基於此結果進一步推導最終答案。

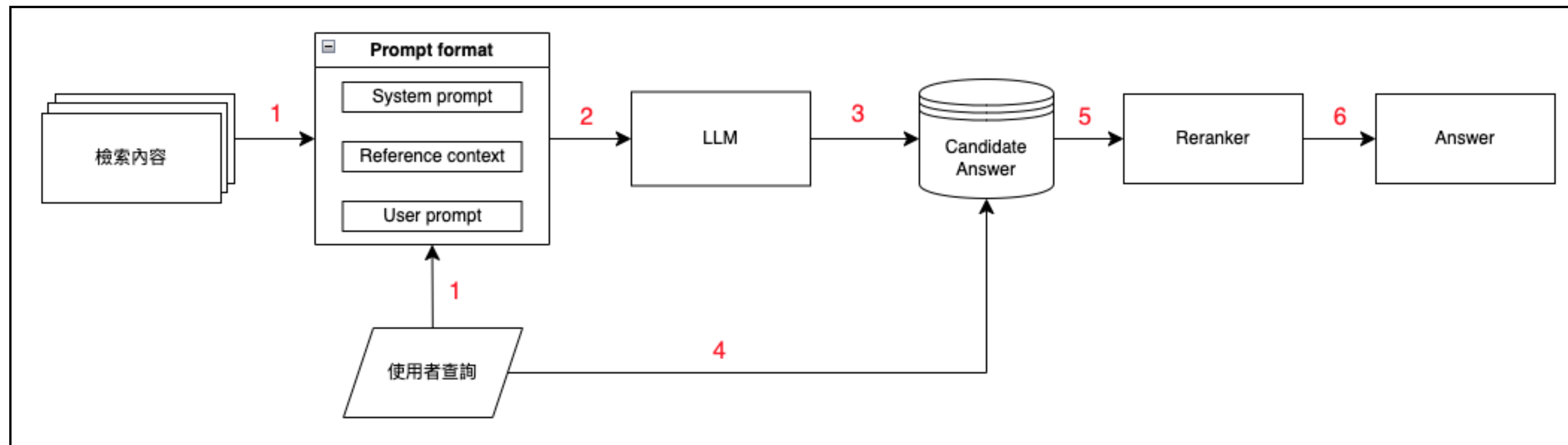


文本生成實驗

生成策略

- 採樣策略 (簡稱"**Sampling (3)**")

生成多組候選答案 (本研究中設定3組)，再利用檢索器選出最符合用戶查詢的對應回應。

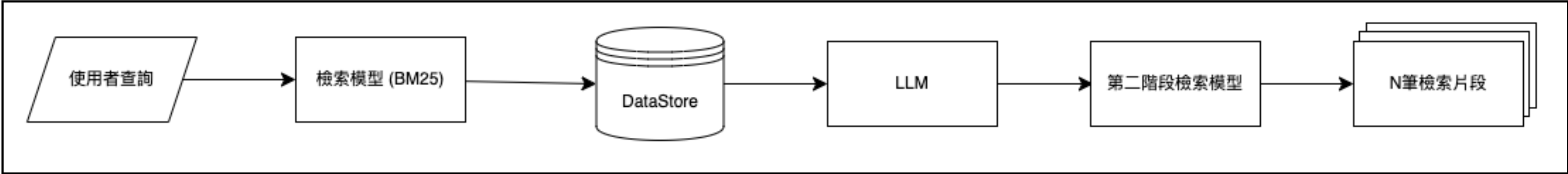


實驗設計

- 文本檢索
- 文本生成
- 延伸實驗

延伸實驗

基於語言模型的檢索優化實驗



- 利用TAME-8B將第一階段檢索案例進行案件的重點整理
- 將重點整理作為輸入，進行第二階段檢索

整理摘要提示詞

<|im_start|>system
You are an AI assistant called Twllm, created by TAME (TAiwan Mixture of Expert) project.
回答時，請遵循以下原則：
1. 僅從提供的context中的內容進行回答
2. 優先提取關鍵字符合度高的段落
3. 排除重複或冗長的描述<|im_end|>
<|im_start|>context
[主要問題]
{摘要}
{客戶提報內容}
[過往解決流程]
{回覆內容}<|im_end|>
<|im_start|>user
請分析context內容，並僅列出與「{使用者查詢}」直接相關的核心資訊，只需要提取關鍵字符合度高的段落<|im_end|>
<|im_start|>assistant

延伸實驗

引導詞實驗

- **TAME-8B在回應某些查詢時，會呈現出一些回應模式**

「根據提供的context內容,與「{使用者查詢}」直接相關的核心資訊如下:」

- **引導詞假設**

加入引導詞對文本生成性能是否可以帶來正面影響，從而導引至理想的輸出內容

實驗結果與分析

- 文本檢索
- 文本生成
- 延伸實驗

文本檢索實驗

第一階段檢索: 案件搜索

- 文本長度與檢索性能呈正相關
 - 所有檢索模型均呈現在整合相關文本特徵後，進一步提高檢索準確度
- BM25性能最為突出
 - 詞袋模型的Token匹配特性
 - 文本整合後提高關鍵詞的匹配機率
- 簡繁轉換的性能改善有限，且結果沒有一致性提升

Feature Set	Model	Accuracy	Avg Query Time (s)
摘要	BM25	52.69%	0.23
	Chuxin(Traditional)	54.84%	0.26
	Chuxin(Simplified)	55.91%	0.33
	Qwen(Traditional)	62.37%	3.28
	Qwen(Simplified)	61.29%	4.37
客戶提報內容	BM25	86.02%	1.23
	Chuxin(Traditional)	83.87%	2.34
	Chuxin(Simplified)	83.87%	2.46
	Qwen(Traditional)	84.95%	15.06
	Qwen(Simplified)	86.02%	16.29
回覆內容	BM25	93.55%	7.47
	Chuxin(Traditional)	82.80%	30.23
	Chuxin(Simplified)	84.95%	30.20
	Qwen(Traditional)	86.02%	70.53
	Qwen(Simplified)	87.10%	70.41
	BM25	100.00%	8.42
摘要 + 客戶提報內容 + 回覆內容	Chuxin(Traditional)	91.40%	30.36
	Chuxin(Simplified)	91.40%	30.43
	Qwen(Traditional)	94.62%	71.03
	Qwen(Simplified)	97.85%	70.80

文本檢索實驗

第二階段檢索: 案件片段提取

- 檢索性能趨勢
 - Precision隨Top N增加而下降； Recall則於Top 8逐步收斂
 - 反映檢索範圍與候選片段數量的關係
- 模型表現比較
 - BM25: 問答對的Token匹配不一致，導致性能較低
 - 神經網路模型: 深層語義理解，捕捉問答之間的語義關聯模式，突破詞彙匹配的限制
- 實務應用選擇
 - Recall閥值設定(95%)
 - Precision & Compression Rate 輔助分析
- 符合閥值的檢索設定與模型
 - BM25: Top 8
 - Chuxin & Qwen: Top 7
- Qwen(Traditional) 維持較高的Compression Rate，同時保持了相對優異的Precision

Top N	Model	Precision	Recall	F1-Score	Compression Rate
1	BM25	32.26%	29.21%	30.06%	88.15%
	Chuxin(Traditional)	41.94%	38.89%	39.73%	87.45%
	Chuxin(Simplified)	40.86%	38.35%	39.02%	87.48%
	Qwen(Traditional)	45.16%	42.11%	42.96%	85.69%
	Qwen(Simplified)	48.39%	44.80%	45.83%	86.44%
2	BM25	25.27%	45.88%	31.99%	75.92%
	Chuxin(Traditional)	32.80%	60.93%	42.03%	74.23%
	Chuxin(Simplified)	31.18%	57.71%	39.87%	74.74%
	Qwen(Traditional)	32.26%	60.39%	41.49%	72.49%
	Qwen(Simplified)	32.26%	59.86%	41.31%	72.91%
3	BM25	25.99%	69.75%	37.21%	62.75%
	Chuxin(Traditional)	27.78%	74.23%	39.60%	62.17%
	Chuxin(Simplified)	27.06%	72.08%	38.52%	62.94%
	Qwen(Traditional)	28.85%	76.92%	41.11%	58.91%
	Qwen(Simplified)	28.14%	74.77%	40.03%	59.91%
4	BM25	23.92%	78.03%	35.78%	52.22%
	Chuxin(Traditional)	24.73%	80.36%	36.86%	53.45%
	Chuxin(Simplified)	24.73%	80.36%	36.86%	53.27%
	Qwen(Traditional)	25.27%	82.51%	37.72%	48.76%
	Qwen(Simplified)	25.27%	82.51%	37.72%	49.71%
5	BM25	23.17%	85.23%	35.39%	43.84%
	Chuxin(Traditional)	23.82%	88.10%	36.35%	45.68%
	Chuxin(Simplified)	23.17%	84.87%	35.28%	45.10%
	Qwen(Traditional)	24.03%	87.74%	36.50%	41.68%
	Qwen(Simplified)	24.03%	89.18%	36.71%	41.87%
6	BM25	22.71%	91.36%	35.12%	36.72%
	Chuxin(Traditional)	23.42%	94.77%	36.22%	37.70%
	Chuxin(Simplified)	23.42%	95.09%	36.30%	37.51%
	Qwen(Traditional)	23.24%	92.80%	35.79%	35.56%
	Qwen(Simplified)	23.24%	93.33%	35.83%	35.07%
7	BM25	22.09%	94.23%	34.39%	31.04%
	Chuxin(Traditional)	22.25%	95.30%	34.66%	31.90%
	Chuxin(Simplified)	22.25%	95.30%	34.66%	31.77%
	Qwen(Traditional)	22.55%	96.56%	35.09%	29.68%
	Qwen(Simplified)	22.55%	96.56%	35.09%	29.46%
8	BM25	21.52%	95.48%	33.61%	27.03%
	Chuxin(Traditional)	21.65%	96.59%	33.86%	27.41%
	Chuxin(Simplified)	21.65%	96.59%	33.86%	27.49%
	Qwen(Traditional)	21.65%	96.56%	33.85%	25.78%
	Qwen(Simplified)	21.78%	97.63%	34.09%	25.35%
9	BM25	21.17%	97.63%	33.23%	23.73%
	Chuxin(Traditional)	21.29%	97.85%	33.38%	23.79%
	Chuxin(Simplified)	21.17%	97.67%	33.24%	24.69%
	Qwen(Traditional)	21.17%	97.63%	33.23%	22.51%
	Qwen(Simplified)	21.17%	97.63%	33.23%	22.43%
10	BM25	20.89%	97.85%	32.82%	20.58%
	Chuxin(Traditional)	20.89%	97.85%	32.82%	21.49%
	Chuxin(Simplified)	20.78%	97.67%	32.68%	21.96%
	Qwen(Traditional)	20.89%	98.17%	32.85%	20.04%
	Qwen(Simplified)	20.89%	98.17%	32.85%	20.06%

文本檢索實驗

第二階段檢索: 案件片段提取 (混合檢索)

Chuxin和Qwen在不同閾值下呈現交錯的性能優勢，表明兩個模型在檢索範圍上具有互補性

Top N	Model	Precision	Recall	F1-Score	Compression Rate
1	BM25	32.26%	29.21%	30.06%	88.15%
	Chuxin(Traditional)	41.94%	38.89%	39.73%	87.45%
	Chuxin(Simplified)	40.86%	38.35%	39.02%	87.48%
	Qwen(Traditional)	45.16%	42.11%	42.96%	85.69%
	Qwen(Simplified)	48.39%	44.80%	45.83%	86.44%
2	BM25	25.27%	45.88%	31.99%	75.92%
	Chuxin(Traditional)	32.80%	60.93%	42.03%	74.23%
	Chuxin(Simplified)	31.18%	57.71%	39.87%	74.74%
	Qwen(Traditional)	32.26%	60.39%	41.49%	72.49%
	Qwen(Simplified)	32.26%	59.86%	41.31%	72.91%
3	BM25	25.99%	69.75%	37.21%	62.75%
	Chuxin(Traditional)	27.78%	74.23%	39.60%	62.17%
	Chuxin(Simplified)	27.06%	72.08%	38.52%	62.94%
	Qwen(Traditional)	28.85%	76.92%	41.11%	58.91%
	Qwen(Simplified)	28.14%	74.77%	40.03%	59.91%
4	BM25	23.92%	78.03%	35.78%	52.22%
	Chuxin(Traditional)	24.73%	80.36%	36.86%	53.45%
	Chuxin(Simplified)	24.73%	80.36%	36.86%	53.27%
	Qwen(Traditional)	25.27%	82.51%	37.72%	48.76%
	Qwen(Simplified)	25.27%	82.51%	37.72%	49.71%
5	BM25	23.17%	85.23%	35.39%	43.84%
	Chuxin(Traditional)	23.82%	88.10%	36.35%	45.68%
	Chuxin(Simplified)	23.17%	84.87%	35.28%	45.10%
	Qwen(Traditional)	24.03%	87.74%	36.50%	41.68%
	Qwen(Simplified)	24.03%	89.18%	36.71%	41.87%
6	BM25	22.71%	91.36%	35.12%	36.72%
	Chuxin(Traditional)	23.42%	94.77%	36.22%	37.70%
	Chuxin(Simplified)	23.42%	95.09%	36.30%	37.51%
	Qwen(Traditional)	23.24%	92.80%	35.79%	35.56%
	Qwen(Simplified)	23.24%	93.33%	35.83%	35.07%
7	BM25	22.09%	94.23%	34.39%	31.04%
	Chuxin(Traditional)	22.25%	95.30%	34.66%	31.90%
	Chuxin(Simplified)	22.25%	95.30%	34.66%	31.77%
	Qwen(Traditional)	22.55%	96.56%	35.09%	29.68%
	Qwen(Simplified)	22.55%	96.56%	35.09%	29.46%
8	BM25	21.52%	95.48%	33.61%	27.03%
	Chuxin(Traditional)	21.65%	96.59%	33.86%	27.41%
	Chuxin(Simplified)	21.65%	96.59%	33.86%	27.49%
	Qwen(Traditional)	21.65%	96.56%	33.85%	25.78%
	Qwen(Simplified)	21.78%	97.63%	34.09%	25.35%
9	BM25	21.17%	97.63%	33.23%	23.73%
	Chuxin(Traditional)	21.29%	97.85%	33.38%	23.79%
	Chuxin(Simplified)	21.17%	97.67%	33.24%	24.69%
	Qwen(Traditional)	21.17%	97.63%	33.23%	22.51%
	Qwen(Simplified)	21.17%	97.63%	33.23%	22.43%
10	BM25	20.89%	97.85%	32.82%	20.58%
	Chuxin(Traditional)	20.89%	97.85%	32.82%	21.49%
	Chuxin(Simplified)	20.78%	97.67%	32.68%	21.96%
	Qwen(Traditional)	20.89%	98.17%	32.85%	20.04%
	Qwen(Simplified)	20.89%	98.17%	32.85%	20.06%

文本檢索實驗

第二階段檢索: 案件片段提取 (混合檢索)

- 以BM25 (25%), Chuxin (35%), Qwen (40%)的權重配置在Top 3、4、5和10的檢索範圍內實現了最佳的檢索效果
- 基於目前測試集的大小，尚不能證明其最佳的參數配置與其有效性

Top N	Precision	Recall	F1-Score	Compression Rate
1	43.01%	39.43%	40.45%	86.63%
2	31.18%	57.71%	39.87%	72.94%
3	29.21%	78.89%	41.94%	60.76%
4	25.54%	83.58%	38.15%	50.95%
5	24.25%	90.25%	37.07%	42.67%
6	23.24%	92.80%	35.79%	35.66%
7	22.55%	95.70%	35.00%	30.46%
8	21.78%	96.77%	34.01%	26.07%
9	21.29%	97.31%	33.36%	22.45%
10	20.99%	98.39%	32.99%	19.96%

Top N	Model	Precision	Recall	F1-Score	Compression Rate
1	BM25	32.26%	29.21%	30.06%	88.15%
	Chuxin(Traditional)	41.94%	38.89%	39.73%	87.45%
	Chuxin(Simplified)	40.86%	38.35%	39.02%	87.48%
	Qwen(Traditional)	45.16%	42.11%	42.96%	85.69%
	Qwen(Simplified)	48.39%	44.80%	45.83%	86.44%
2	BM25	25.27%	45.88%	31.99%	75.92%
	Chuxin(Traditional)	32.80%	60.93%	42.03%	74.23%
	Chuxin(Simplified)	31.18%	57.71%	39.87%	74.74%
	Qwen(Traditional)	32.26%	60.39%	41.49%	72.49%
	Qwen(Simplified)	32.26%	59.86%	41.31%	72.91%
3	BM25	25.99%	69.75%	37.21%	62.75%
	Chuxin(Traditional)	27.78%	74.23%	39.60%	62.17%
	Chuxin(Simplified)	27.06%	72.08%	38.52%	62.94%
	Qwen(Traditional)	28.85%	76.92%	41.11%	58.91%
	Qwen(Simplified)	28.14%	74.77%	40.03%	59.91%
4	BM25	23.92%	78.03%	35.78%	52.22%
	Chuxin(Traditional)	24.73%	80.36%	36.86%	53.45%
	Chuxin(Simplified)	24.73%	80.36%	36.86%	53.27%
	Qwen(Traditional)	25.27%	82.51%	37.72%	48.76%
	Qwen(Simplified)	25.27%	82.51%	37.72%	49.71%
5	BM25	23.17%	85.23%	35.39%	43.84%
	Chuxin(Traditional)	23.82%	88.10%	36.35%	45.68%
	Chuxin(Simplified)	23.17%	84.87%	35.28%	45.10%
	Qwen(Traditional)	24.03%	87.74%	36.50%	41.68%
	Qwen(Simplified)	24.03%	89.18%	36.71%	41.87%
6	BM25	22.71%	91.36%	35.12%	36.72%
	Chuxin(Traditional)	23.42%	94.77%	36.22%	37.70%
	Chuxin(Simplified)	23.42%	95.09%	36.30%	37.51%
	Qwen(Traditional)	23.24%	92.80%	35.79%	35.56%
	Qwen(Simplified)	23.24%	93.33%	35.83%	35.07%
7	BM25	22.09%	94.23%	34.39%	31.04%
	Chuxin(Traditional)	22.25%	95.30%	34.66%	31.90%
	Chuxin(Simplified)	22.25%	95.30%	34.66%	31.77%
	Qwen(Traditional)	22.55%	96.56%	35.09%	29.68%
	Qwen(Simplified)	22.55%	96.56%	35.09%	29.46%
8	BM25	21.52%	95.48%	33.61%	27.03%
	Chuxin(Traditional)	21.65%	96.59%	33.86%	27.41%
	Chuxin(Simplified)	21.65%	96.59%	33.86%	27.49%
	Qwen(Traditional)	21.65%	96.56%	33.85%	25.78%
	Qwen(Simplified)	21.78%	97.63%	34.09%	25.35%
9	BM25	21.17%	97.63%	33.23%	23.73%
	Chuxin(Traditional)	21.29%	97.85%	33.38%	23.79%
	Chuxin(Simplified)	21.17%	97.67%	33.24%	24.69%
	Qwen(Traditional)	21.17%	97.63%	33.23%	22.51%
	Qwen(Simplified)	21.17%	97.63%	33.23%	22.43%
10	BM25	20.89%	97.85%	32.82%	20.58%
	Chuxin(Traditional)	20.89%	97.85%	32.82%	21.49%
	Chuxin(Simplified)	20.78%	97.67%	32.68%	21.96%
	Qwen(Traditional)	20.89%	98.17%	32.85%	20.04%
	Qwen(Simplified)	20.89%	98.17%	32.85%	20.06%

實驗結果與分析

- 文本檢索
- 文本生成
- 延伸實驗

文本生成實驗

實驗方法

系統提示

- TAME 預設配置
- 明確角色能力
- 無系統提示

檢索內容

- 基於案件檢索結果
- 基於案件片段檢索結果
- 基於時序排序的案件片段檢索結果

生成策略

- 基礎生成策略
- 推理引導策略
- 採樣策略

文本生成

系統提示策略分析

System Prompt	BERT Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
TAME	0.80	0.25	0.63	0.55	0.80	19.35
Without System Prompt	0.81	0.28	0.66	0.60	0.75	26.45
Expert	0.81	0.25	0.64	0.57	0.84	20.21

- Without System Prompt
 - BLEU與ROUGE指標上表現最佳，意味著用詞相似程度較高
 - Inference Time明顯較長，原因是其生成內容冗長，經常出現無意義的重複序列
 - 檢索知識庫中的「回覆內容」中包含大量Q&A對，產生類似 Few-Shot Prompting 的效果，使其在部分測試集上表現優異，但部分則表現很差
 - 缺乏系統提示導致模型輸出缺乏任務焦點和結構性，僅能從檢索資料判別輸出模式
- TAME & Expert Prompt
 - 各項性能指標接近，無特別優異
 - 指出具體任務需求的系統提示可使TAME-8B回覆內容更為穩定

System Prompt	Retrieval Type	Skills	BERT Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
TAME	Time Sorted	Base	0.81 (0.11)	0.30 (0.29)	0.63 (0.32)	0.56 (0.34)	0.82	5.22 (6.14)
		Reasoning	0.79 (0.10)	0.17 (0.17)	0.64 (0.29)	0.56 (0.31)	0.77	33.91 (15.7)
		Sampling (3)	0.82 (0.11)	0.30 (0.29)	0.65 (0.31)	0.58 (0.33)	0.83	16.88 (16.78)
	Passage	Base	0.83 (0.11)	0.34 (0.31)	0.67 (0.30)	0.60 (0.34)	0.85	5.53 (6.31)
		Reasoning	0.79 (0.09)	0.17 (0.17)	0.61 (0.29)	0.52 (0.30)	0.78	33.33 (15.31)
		Sampling (3)	0.82 (0.11)	0.30 (0.28)	0.66 (0.31)	0.59 (0.34)	0.85	17.42 (21.07)
	Document	Base	0.80 (0.12)	0.26 (0.27)	0.58 (0.32)	0.50 (0.34)	0.76	6.45 (8.48)
		Reasoning	0.77 (0.10)	0.15 (0.15)	0.61 (0.30)	0.52 (0.31)	0.72	36.02 (17.50)
		Sampling (3)	0.81 (0.11)	0.27 (0.26)	0.62 (0.31)	0.55 (0.34)	0.78	19.35 (23.30)
Without System Prompt	Time Sorted	Base	0.82 (0.12)	0.30 (0.31)	0.70 (0.32)	0.65 (0.34)	0.77	11.93 (13.84)
		Reasoning	0.79 (0.11)	0.23 (0.24)	0.58 (0.31)	0.51 (0.33)	0.73	35.89 (14.88)
		Sampling (3)	0.82 (0.12)	0.29 (0.31)	0.69 (0.31)	0.64 (0.33)	0.74	34.36 (36.40)
	Passage	Base	0.82 (0.13)	0.34 (0.32)	0.73 (0.29)	0.66 (0.32)	0.76	10.20 (11.85)
		Reasoning	0.80 (0.11)	0.22 (0.25)	0.57 (0.32)	0.50 (0.33)	0.75	34.47 (15.34)
		Sampling (3)	0.82 (0.12)	0.31 (0.31)	0.71 (0.30)	0.65 (0.33)	0.81	30.29 (32.74)
	Document	Base	0.81 (0.12)	0.29 (0.30)	0.69 (0.32)	0.64 (0.35)	0.72	12.06 (13.61)
		Reasoning	0.79 (0.10)	0.20 (0.23)	0.56 (0.31)	0.48 (0.33)	0.73	35.11 (15.94)
		Sampling (3)	0.81 (0.12)	0.29 (0.30)	0.68 (0.31)	0.63 (0.35)	0.75	33.70 (34.37)
Expert	Time Sorted	Base	0.82 (0.11)	0.29 (0.27)	0.66 (0.31)	0.59 (0.33)	0.83	6.21 (5.73)
		Reasoning	0.78 (0.09)	0.16 (0.17)	0.61 (0.28)	0.52 (0.29)	0.84	34.32 (13.59)
		Sampling (3)	0.81 (0.11)	0.28 (0.27)	0.65 (0.30)	0.58 (0.34)	0.86	18.79 (17.50)
	Passage	Base	0.82 (0.11)	0.31 (0.28)	0.66 (0.29)	0.58 (0.32)	0.87	5.68 (4.17)
		Reasoning	0.79 (0.10)	0.19 (0.19)	0.63 (0.30)	0.55 (0.31)	0.84	34.30 (15.49)
		Sampling (3)	0.82 (0.11)	0.31 (0.28)	0.68 (0.30)	0.60 (0.33)	0.87	18.31 (14.64)
	Document	Base	0.82 (0.10)	0.27 (0.25)	0.66 (0.30)	0.58 (0.33)	0.85	7.33 (7.65)
		Reasoning	0.78 (0.09)	0.15 (0.16)	0.58 (0.29)	0.50 (0.31)	0.78	34.31 (15.37)
		Sampling (3)	0.82 (0.10)	0.27 (0.25)	0.67 (0.29)	0.59 (0.32)	0.82	22.66 (23.58)

題目:

透過介面可查訊單張報表申報日，若要整批查詢所有報表設定的申報日該如何作業，請提供操作步驟或語法

預期答案:

可於於「報表資訊\報表資訊\報表資訊管理」單張查看或於「首頁行事曆」查看所有報表申報資訊

輸入提示 (Without System Prompt + Document + Base):

<|im_start|>context

[主要問題]

[{客戶公司名稱}]整批查詢報表申報日資訊

【聯絡資訊】 {客戶姓名} - {客戶電話} - {客戶e-mail} 【問題類別】 報表問題 【問題項目】 申報日 【問題說明】 整批蒐集所有報表申報日資訊 目前透過介面可查訊單張報表申報日，若要整批查詢所有報表設定的申報日該如何作業，請提供操作步驟或語法，謝謝。

[過往解決流程]

{回覆內容}<|im_end|>

<|im_start|>user

{使用者查詢}<|im_end|>

<|im_start|>assistant

TAME-8B輸出內容

【回覆說明】

1.想請問是否欲查看每張報表的「申報頻率」以及「申報日期限」？

{客戶公司名稱}回覆：是的

2.想確認是否有遇到什麼問題，因此需特別查詢？

{客戶公司名稱}回覆：若遇到事件時，可先確認近期即將申報的報表是否受影響，也可判斷還有多少時間可供事件處理，預計將此資訊收錄到本行報表清單文件內，新增欄位紀錄申報截止日，若無法整批取得將需要花時間逐張整理

【回覆內容】

1.請提供每張報表的「申報頻率」以及「申報日期限」查詢語法

2.若無法整批取得，是否有其他方式可快速取得報表資訊

3.若無法整批取得，是否有其他方式可快速取得報表資訊

4.若無法整批取得，是否有其他方式可快速取得報表資訊

---以下省略至第42次重複後觸碰到1024 tokens的最大長度即停止生成回應---

TAME-8B輸出內容

【回覆說明】

1.想請問是否欲查看每張報表的「申報頻率」以及「申報日期限」？

{客戶公司名稱}回覆：是的

2.想確認是否有遇到什麼問題，因此需特別查詢？

{客戶公司名稱}回覆：若遇到事件時，可先確認近期即將申報的報表是否受影響，也可判斷還有多少時間可供事件處理，預計將此資訊收錄到本行報表清單文件內，新增欄位紀錄申報截止日，若無法整批取得將需要花時間逐張整理

【回覆內容】

1.請提供每張報表的「申報頻率」以及「申報日期限」查詢語法

2.若無法整批取得，是否有其他方式可快速取得報表資訊

3.若無法整批取得，是否有其他方式可快速取得報表資訊

4.若無法整批取得，是否有其他方式可快速取得報表資訊

---以下省略至第42次重複後觸碰到1024 tokens的最大長度即停止生成回應---

【案件狀態】 內部分析 【案件類型】 平台諮詢 【問題分析】 透過介面可查訊單張報表「申報日 每月第X日」，若要整批查詢所有報表設定的申報日該如何作業，請提供操作步驟或語法 【建議方案】 建議將申報日顯示於平台頁面上 【處理說明】 客戶欲查詢每張報表的「申報日 每月第X日」，但從平台上無法查詢相關資訊，匯出excel也無此資訊，再請產品團隊確認是否納入優化，謝謝。 也請確認暫解方案是否提供語法「{SQL語法} 查看Rpt_Freq、Rpt_Day」 【預計交期】 無

【案件狀態】 對外發佈 【案件類型】 平台諮詢 【問題分析】 透過介面可查訊單張報表「申報日 每月第X日」，若要整批查詢所有報表設定的申報日該如何作業，請提供操作步驟或語法 【建議方案】 建議將申報日顯示於平台頁面上 【處理說明】 建議可以透過首頁行事曆瀏覽各張報表申報日資訊

【案件狀態】 對外發佈 【案件類型】 平台諮詢 【信件時間】 2024/1/9 (週二) 下午 05:05 【信件人員】 {公司客服人員姓名} <{公司客服人員之email}> 【問題分析】 透過介面可查訊單張報表申報日，若要整批查詢所有報表設定的申報日該如何作業，請提供操作步驟或語法 【建議方案】 請問需於平台查看申報日的目的 【處理說明】 想請問是否欲查看每張報表的「申報頻率」以及「申報日期限」？想確認是否有遇到什麼問題，因此需特別查詢？ 目前可於於「報表資訊\報表資訊\報表資訊管理」單張查看 或於「首頁行事曆」查看所有報表申報資訊 再請您確認回覆，感謝。 【預計交期】 暫無 【內部分析】 因客戶想要的是整批查詢，提出的為資訊單位人員，若於首頁行事曆報表過多不易檢視，因此再請產品團隊確認以下建議： >1.客戶欲查詢每張報表的「申報日 每月第X日」，但從平台上無法查詢相關資訊，匯出excel也無此資訊，若於平台畫面不顯示「月第X日」，是否Excel匯出結果檔可將「月第X日」帶出，以利客戶確認全部報表的「申報日」資訊，再請產品團隊確認是否納入優化，謝謝。

【案件狀態】 客戶回覆 【案件類型】 諮詢 【信件時間】 2024/1/9 (週二) 下午 05:24 【信件人員】 {公司客服人員姓名} <{公司客服人員之email}> 【回覆說明】 1.想請問是否欲查看每張報表的「申報頻率」以及「申報日期限」？ {客戶公司名稱}回覆：是的 2.想確認是否有遇到什麼問題，因此需特別查詢？ {客戶公司名稱}回覆：若遇到事件時，可先確認近期即將申報的報表是否受影響，也可判斷還有多少時間可供事件處理，預計將此資訊收錄到本行報表清單文件內，新增欄位紀錄申報截止日，若無法整批取得將需要花時間逐張整理

【案件狀態】 對外發佈 【案件類型】 客戶服務 【信件時間】 2024/1/10 (週三) 上午 09:50 【信件人員】 {客服人員姓名} <{客戶人員之email}> 【問題分析】 欲查看每張報表的「申報頻率」以及「申報日期限」，若遇到事件時，可先確認近期即將申報的報表是否受影響，也可判斷還有多少時間可供事件處理 【建議方案】 參閱附件檔，但後續請業管單位確認是否收到主管機關最新更新檔 【處理說明】 平台上僅能單一查詢，請參閱附件資訊，此為同業提供給我們資訊，應從主管機關網站下載，建議後續與業管單位確認是否有收到主管機關更新通知，感謝。 【預計交期】 2024/01/10

【案件狀態】 客戶回覆 【案件類型】 客戶服務 【信件時間】 2024/1/10 (週三) 下午 01:21 【信件人員】 {公司客服人員姓名} <{公司客服人員之email}> 【信件內容】 收到，感謝協助，後續我們也會請行內業管確認正確性，此議題可結案。 【記錄】 最終處理提供客戶主管機關文件，並請客戶可至主管機關網站取得最新版本 ==> 諮詢服務 前期提供客戶語法 ==> 客戶服務

文本生成

系統提示策略分析

System Prompt	BERT Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
TAME	0.80	0.25	0.63	0.55	0.80	19.35
Without System Prompt	0.81	0.28	0.66	0.60	0.75	26.45
Expert	0.81	0.25	0.64	0.57	0.84	20.21

- Without System Prompt
 - BLEU與ROUGE指標上表現最佳，意味著用詞相似程度較高
 - Inference Time明顯較長，原因是其生成內容冗長，經常出現無意義的重複序列
 - 檢索知識庫中的「回覆內容」中包含大量Q&A對，產生類似 Few-Shot Prompting 的效果，使其在部分測試集上表現優異，但部分則表現很差
 - 缺乏系統提示導致模型輸出缺乏任務焦點和結構性，僅能從檢索資料判別輸出模式
- TAME & Expert Prompt
 - 各項性能指標接近，無特別優異
 - 指出具體任務需求的系統提示可使TAME-8B回覆內容更為穩定

System Prompt	Retrieval Type	Skills	BERT Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
TAME	Time Sorted	Base	0.81 (0.11)	0.30 (0.29)	0.63 (0.32)	0.56 (0.34)	0.82	5.22 (6.14)
		Reasoning	0.79 (0.10)	0.17 (0.17)	0.64 (0.29)	0.56 (0.31)	0.77	33.91 (15.7)
		Sampling (3)	0.82 (0.11)	0.30 (0.29)	0.65 (0.31)	0.58 (0.33)	0.83	16.88 (16.78)
	Passage	Base	0.83 (0.11)	0.34 (0.31)	0.67 (0.30)	0.60 (0.34)	0.85	5.53 (6.31)
		Reasoning	0.79 (0.09)	0.17 (0.17)	0.61 (0.29)	0.52 (0.30)	0.78	33.33 (15.31)
		Sampling (3)	0.82 (0.11)	0.30 (0.28)	0.66 (0.31)	0.59 (0.34)	0.85	17.42 (21.07)
	Document	Base	0.80 (0.12)	0.26 (0.27)	0.58 (0.32)	0.50 (0.34)	0.76	6.45 (8.48)
		Reasoning	0.77 (0.10)	0.15 (0.15)	0.61 (0.30)	0.52 (0.31)	0.72	36.02 (17.50)
		Sampling (3)	0.81 (0.11)	0.27 (0.26)	0.62 (0.31)	0.55 (0.34)	0.78	19.35 (23.30)
Without System Prompt	Time Sorted	Base	0.82 (0.12)	0.30 (0.31)	0.70 (0.32)	0.65 (0.34)	0.77	11.93 (13.84)
		Reasoning	0.79 (0.11)	0.23 (0.24)	0.58 (0.31)	0.51 (0.33)	0.73	35.89 (14.88)
		Sampling (3)	0.82 (0.12)	0.29 (0.31)	0.69 (0.31)	0.64 (0.33)	0.74	34.36 (36.40)
	Passage	Base	0.82 (0.13)	0.34 (0.32)	0.73 (0.29)	0.66 (0.32)	0.76	10.20 (11.85)
		Reasoning	0.80 (0.11)	0.22 (0.25)	0.57 (0.32)	0.50 (0.33)	0.75	34.47 (15.34)
		Sampling (3)	0.82 (0.12)	0.31 (0.31)	0.71 (0.30)	0.65 (0.33)	0.81	30.29 (32.74)
	Document	Base	0.81 (0.12)	0.29 (0.30)	0.69 (0.32)	0.64 (0.35)	0.72	12.06 (13.61)
		Reasoning	0.79 (0.10)	0.20 (0.23)	0.56 (0.31)	0.48 (0.33)	0.73	35.11 (15.94)
		Sampling (3)	0.81 (0.12)	0.29 (0.30)	0.68 (0.31)	0.63 (0.35)	0.75	33.70 (34.37)
Expert	Time Sorted	Base	0.82 (0.11)	0.29 (0.27)	0.66 (0.31)	0.59 (0.33)	0.83	6.21 (5.73)
		Reasoning	0.78 (0.09)	0.16 (0.17)	0.61 (0.28)	0.52 (0.29)	0.84	34.32 (13.59)
		Sampling (3)	0.81 (0.11)	0.28 (0.27)	0.65 (0.30)	0.58 (0.34)	0.86	18.79 (17.50)
	Passage	Base	0.82 (0.11)	0.31 (0.28)	0.66 (0.29)	0.58 (0.32)	0.87	5.68 (4.17)
		Reasoning	0.79 (0.10)	0.19 (0.19)	0.63 (0.30)	0.55 (0.31)	0.84	34.30 (15.49)
		Sampling (3)	0.82 (0.11)	0.31 (0.28)	0.68 (0.30)	0.60 (0.33)	0.87	18.31 (14.64)
	Document	Base	0.82 (0.10)	0.27 (0.25)	0.66 (0.30)	0.58 (0.33)	0.85	7.33 (7.65)
		Reasoning	0.78 (0.09)	0.15 (0.16)	0.58 (0.29)	0.50 (0.31)	0.78	34.31 (15.37)
		Sampling (3)	0.82 (0.10)	0.27 (0.25)	0.67 (0.29)	0.59 (0.32)	0.82	22.66 (23.58)

文本生成實驗

實驗方法

系統提示

- TAME 預設配置
- 明確角色能力
- 無系統提示

檢索內容

- 基於案件檢索結果
- 基於案件片段檢索結果
- 基於時序排序的案件片段檢索結果

生成策略

- 基礎生成策略
- 推理引導策略
- 採樣策略

文本生成

檢索方法

Retrieval Type	BERT Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
Time Sorted Average	0.81	0.26	0.65	0.57	0.80	21.95
Passage Average	0.81	0.28	0.66	0.58	0.82	21.06
Document Average	0.80	0.24	0.63	0.56	0.77	23.00

- Passage
 - 過濾冗餘資訊顯著提升生成品質並降低推理時間。
- Time Sorted
 - 雖過濾冗餘資訊的同時保留時序資訊，但性能提升有限。
- Document
 - 保留完整上下文，但過多冗餘資訊干擾模型對重要資訊的判斷，導致性能最低且同時推理時間因Token數而最長。
- Passage 將最相關的回覆內容優先排序，使模型更容易掌握問答對應模式(Q&A對匹配)。

Liu 等人的研究[5]也證實將關鍵資訊放在文本開頭能顯著提升模型表現，支持本研究結果。

System Prompt	Retrieval Type	Skills	BERT Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
TAME	Time Sorted	Base	0.81 (0.11)	0.30 (0.29)	0.63 (0.32)	0.56 (0.34)	0.82	5.22 (6.14)
		Reasoning	0.79 (0.10)	0.17 (0.17)	0.64 (0.29)	0.56 (0.31)	0.77	33.91 (15.7)
		Sampling (3)	0.82 (0.11)	0.30 (0.29)	0.65 (0.31)	0.58 (0.33)	0.83	16.88 (16.78)
	Passage	Base	0.83 (0.11)	0.34 (0.31)	0.67 (0.30)	0.60 (0.34)	0.85	5.53 (6.31)
		Reasoning	0.79 (0.09)	0.17 (0.17)	0.61 (0.29)	0.52 (0.30)	0.78	33.33 (15.31)
		Sampling (3)	0.82 (0.11)	0.30 (0.28)	0.66 (0.31)	0.59 (0.34)	0.85	17.42 (21.07)
	Document	Base	0.80 (0.12)	0.26 (0.27)	0.58 (0.32)	0.50 (0.34)	0.76	6.45 (8.48)
		Reasoning	0.77 (0.10)	0.15 (0.15)	0.61 (0.30)	0.52 (0.31)	0.72	36.02 (17.50)
		Sampling (3)	0.81 (0.11)	0.27 (0.26)	0.62 (0.31)	0.55 (0.34)	0.78	19.35 (23.30)
Without System Prompt	Time Sorted	Base	0.82 (0.12)	0.30 (0.31)	0.70 (0.32)	0.65 (0.34)	0.77	11.93 (13.84)
		Reasoning	0.79 (0.11)	0.23 (0.24)	0.58 (0.31)	0.51 (0.33)	0.73	35.89 (14.88)
		Sampling (3)	0.82 (0.12)	0.29 (0.31)	0.69 (0.31)	0.64 (0.33)	0.74	34.36 (36.40)
	Passage	Base	0.82 (0.13)	0.34 (0.32)	0.73 (0.29)	0.66 (0.32)	0.76	10.20 (11.85)
		Reasoning	0.80 (0.11)	0.22 (0.25)	0.57 (0.32)	0.50 (0.33)	0.75	34.47 (15.34)
		Sampling (3)	0.82 (0.12)	0.31 (0.31)	0.71 (0.30)	0.65 (0.33)	0.81	30.29 (32.74)
	Document	Base	0.81 (0.12)	0.29 (0.30)	0.69 (0.32)	0.64 (0.35)	0.72	12.06 (13.61)
		Reasoning	0.79 (0.10)	0.20 (0.23)	0.56 (0.31)	0.48 (0.33)	0.73	35.11 (15.94)
		Sampling (3)	0.81 (0.12)	0.29 (0.30)	0.68 (0.31)	0.63 (0.35)	0.75	33.70 (34.37)
Expert	Time Sorted	Base	0.82 (0.11)	0.29 (0.27)	0.66 (0.31)	0.59 (0.33)	0.83	6.21 (5.73)
		Reasoning	0.78 (0.09)	0.16 (0.17)	0.61 (0.28)	0.52 (0.29)	0.84	34.32 (13.59)
		Sampling (3)	0.81 (0.11)	0.28 (0.27)	0.65 (0.30)	0.58 (0.34)	0.86	18.79 (17.50)
	Passage	Base	0.82 (0.11)	0.31 (0.28)	0.66 (0.29)	0.58 (0.32)	0.87	5.68 (4.17)
		Reasoning	0.79 (0.10)	0.19 (0.19)	0.63 (0.30)	0.55 (0.31)	0.84	34.30 (15.49)
		Sampling (3)	0.82 (0.11)	0.31 (0.28)	0.68 (0.30)	0.60 (0.33)	0.87	18.31 (14.64)
	Document	Base	0.82 (0.10)	0.27 (0.25)	0.66 (0.30)	0.58 (0.33)	0.85	7.33 (7.65)
		Reasoning	0.78 (0.09)	0.15 (0.16)	0.58 (0.29)	0.50 (0.31)	0.78	34.31 (15.37)
		Sampling (3)	0.82 (0.10)	0.27 (0.25)	0.67 (0.29)	0.59 (0.32)	0.82	22.66 (23.58)

[5] Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157-173.

文本生成

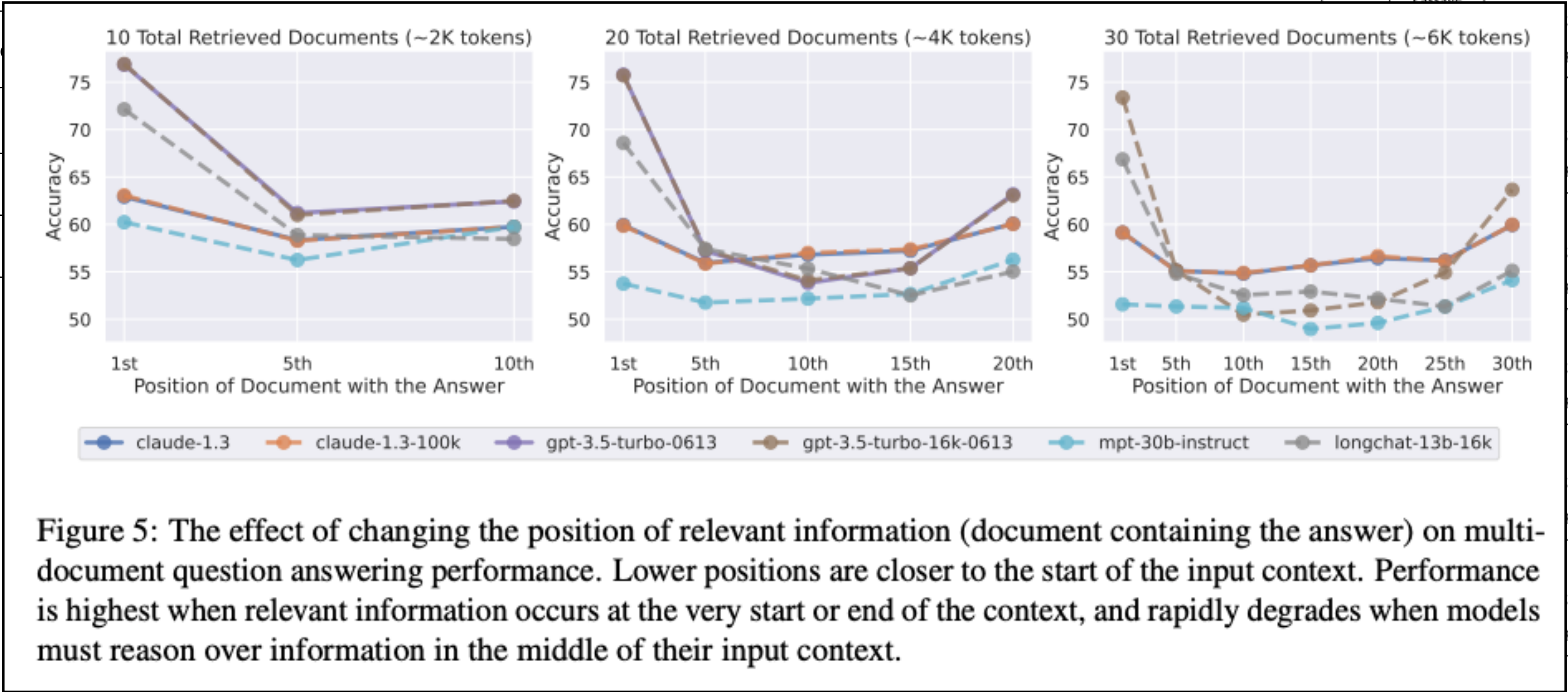
檢索方法

Retrieval Type	BERT Score
Time Sorted Average	0.81
Passage Average	0.81
Document Average	0.80

- Passage
 - 過濾冗餘資訊顯著
- Time Sorted
 - 雖過濾冗餘資訊的
- Document
 - 保留完整上下文，但過多冗餘資訊干擾模型對重要資訊的判斷，導致性能最低且同時推理時間因Token數而最長。

- Passage 將最相關的回覆內容優先排序，使模型更容易掌握問答對應模式(Q&A對匹配)。

Liu 等人的研究[5]也證實將關鍵資訊放在文本開頭能顯著提升模型表現，支持本研究結果。



System Prompt	Retrieval Type	Skills	BERT Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
TAME	Time Sorted	Base	0.81 (0.11)	0.30 (0.29)	0.63 (0.32)	0.56 (0.34)	0.82	5.22 (6.14)
		Reasoning	0.79 (0.10)	0.17 (0.17)	0.64 (0.29)	0.56 (0.31)	0.77	33.91 (15.7)
		Sampling (3)	0.82 (0.11)	0.30 (0.29)	0.65 (0.31)	0.58 (0.33)	0.83	16.88 (16.78)
		Base	0.83 (0.11)	0.34 (0.31)	0.67 (0.30)	0.60 (0.34)	0.85	5.53 (6.31)
		Reasoning	0.79 (0.09)	0.17 (0.17)	0.61 (0.29)	0.52 (0.30)	0.78	33.33 (15.31)
		Sampling (3)	0.82 (0.11)	0.30 (0.28)	0.66 (0.31)	0.59 (0.34)	0.85	17.42 (21.07)
	Passage	Base	0.80 (0.12)	0.26 (0.27)	0.58 (0.32)	0.50 (0.34)	0.76	6.45 (8.48)
		Reasoning	0.77 (0.10)	0.15 (0.15)	0.61 (0.30)	0.52 (0.31)	0.72	36.02 (17.50)
		Sampling (3)	0.81 (0.11)	0.27 (0.26)	0.62 (0.31)	0.55 (0.34)	0.78	19.35 (23.30)
		Base	0.82 (0.12)	0.30 (0.31)	0.70 (0.32)	0.65 (0.34)	0.77	11.93 (13.84)
		Reasoning	0.79 (0.11)	0.23 (0.24)	0.58 (0.31)	0.51 (0.33)	0.73	35.89 (14.88)
		Sampling (3)	0.82 (0.12)	0.29 (0.31)	0.69 (0.31)	0.64 (0.33)	0.74	34.36 (36.40)
	Passage	Base	0.82 (0.13)	0.34 (0.32)	0.73 (0.29)	0.66 (0.32)	0.76	10.20 (11.85)
		Reasoning	0.80 (0.11)	0.22 (0.25)	0.57 (0.32)	0.50 (0.33)	0.75	34.47 (15.34)
		Sampling (3)	0.82 (0.12)	0.31 (0.31)	0.71 (0.30)	0.65 (0.33)	0.81	30.29 (32.74)
		Base	0.81 (0.12)	0.29 (0.30)	0.69 (0.32)	0.64 (0.35)	0.72	12.06 (13.61)
		Reasoning	0.79 (0.10)	0.20 (0.23)	0.56 (0.31)	0.48 (0.33)	0.73	35.11 (15.94)
		Sampling (3)	0.81 (0.12)	0.29 (0.30)	0.68 (0.31)	0.63 (0.35)	0.75	33.70 (34.37)
	Document	Base	0.82 (0.11)	0.29 (0.27)	0.66 (0.31)	0.59 (0.33)	0.83	6.21 (5.73)
		Reasoning	0.78 (0.09)	0.16 (0.17)	0.61 (0.28)	0.52 (0.29)	0.84	34.32 (13.59)
		Sampling (3)	0.81 (0.11)	0.28 (0.27)	0.65 (0.30)	0.58 (0.34)	0.86	18.79 (17.50)
		Base	0.82 (0.11)	0.31 (0.28)	0.66 (0.29)	0.58 (0.32)	0.87	5.68 (4.17)
		Reasoning	0.79 (0.10)	0.19 (0.19)	0.63 (0.30)	0.55 (0.31)	0.84	34.30 (15.49)
		Sampling (3)	0.82 (0.11)	0.31 (0.28)	0.68 (0.30)	0.60 (0.33)	0.87	18.31 (14.64)

[5] Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157-173.

文本生成實驗

實驗方法

系統提示

- TAME 預設配置
- 明確角色能力
- 無系統提示

檢索內容

- 基於案件檢索結果
- 基於案件片段檢索結果
- 基於時序排序的案件片段檢索結果

生成策略

- 基礎生成策略
- 推理引導策略
- 採樣策略

文本生成

檢索策略

Skills	BERT Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
Base Average	0.82	0.30	0.66	0.60	0.80	7.85
Reasoning Average	0.79	0.18	0.60	0.52	0.77	34.63
Sampling (3) Average	0.82	0.29	0.67	0.60	0.81	23.53

- Reasoning
 - 「先總結後找答案」的方式導致答案資訊損失，破壞了文本中的Q&A對結構，減弱模仿效果。
- Sampling
 - 與Base性能接近，但Inference Time顯著增加
 - 可能出於候選文本的多樣性不足

System Prompt	Retrieval Type	Skills	BERT Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
TAME	Time Sorted	Base	0.81 (0.11)	0.30 (0.29)	0.63 (0.32)	0.56 (0.34)	0.82	5.22 (6.14)
		Reasoning	0.79 (0.10)	0.17 (0.17)	0.64 (0.29)	0.56 (0.31)	0.77	33.91 (15.7)
		Sampling (3)	0.82 (0.11)	0.30 (0.29)	0.65 (0.31)	0.58 (0.33)	0.83	16.88 (16.78)
	Passage	Base	0.83 (0.11)	0.34 (0.31)	0.67 (0.30)	0.60 (0.34)	0.85	5.53 (6.31)
		Reasoning	0.79 (0.09)	0.17 (0.17)	0.61 (0.29)	0.52 (0.30)	0.78	33.33 (15.31)
		Sampling (3)	0.82 (0.11)	0.30 (0.28)	0.66 (0.31)	0.59 (0.34)	0.85	17.42 (21.07)
	Document	Base	0.80 (0.12)	0.26 (0.27)	0.58 (0.32)	0.50 (0.34)	0.76	6.45 (8.48)
		Reasoning	0.77 (0.10)	0.15 (0.15)	0.61 (0.30)	0.52 (0.31)	0.72	36.02 (17.50)
		Sampling (3)	0.81 (0.11)	0.27 (0.26)	0.62 (0.31)	0.55 (0.34)	0.78	19.35 (23.30)
Without System Prompt	Time Sorted	Base	0.82 (0.12)	0.30 (0.31)	0.70 (0.32)	0.65 (0.34)	0.77	11.93 (13.84)
		Reasoning	0.79 (0.11)	0.23 (0.24)	0.58 (0.31)	0.51 (0.33)	0.73	35.89 (14.88)
		Sampling (3)	0.82 (0.12)	0.29 (0.31)	0.69 (0.31)	0.64 (0.33)	0.74	34.36 (36.40)
	Passage	Base	0.82 (0.13)	0.34 (0.32)	0.73 (0.29)	0.66 (0.32)	0.76	10.20 (11.85)
		Reasoning	0.80 (0.11)	0.22 (0.25)	0.57 (0.32)	0.50 (0.33)	0.75	34.47 (15.34)
		Sampling (3)	0.82 (0.12)	0.31 (0.31)	0.71 (0.30)	0.65 (0.33)	0.81	30.29 (32.74)
	Document	Base	0.81 (0.12)	0.29 (0.30)	0.69 (0.32)	0.64 (0.35)	0.72	12.06 (13.61)
		Reasoning	0.79 (0.10)	0.20 (0.23)	0.56 (0.31)	0.48 (0.33)	0.73	35.11 (15.94)
		Sampling (3)	0.81 (0.12)	0.29 (0.30)	0.68 (0.31)	0.63 (0.35)	0.75	33.70 (34.37)
Expert	Time Sorted	Base	0.82 (0.11)	0.29 (0.27)	0.66 (0.31)	0.59 (0.33)	0.83	6.21 (5.73)
		Reasoning	0.78 (0.09)	0.16 (0.17)	0.61 (0.28)	0.52 (0.29)	0.84	34.32 (13.59)
		Sampling (3)	0.81 (0.11)	0.28 (0.27)	0.65 (0.30)	0.58 (0.34)	0.86	18.79 (17.50)
	Passage	Base	0.82 (0.11)	0.31 (0.28)	0.66 (0.29)	0.58 (0.32)	0.87	5.68 (4.17)
		Reasoning	0.79 (0.10)	0.19 (0.19)	0.63 (0.30)	0.55 (0.31)	0.84	34.30 (15.49)
		Sampling (3)	0.82 (0.11)	0.31 (0.28)	0.68 (0.30)	0.60 (0.33)	0.87	18.31 (14.64)
	Document	Base	0.82 (0.10)	0.27 (0.25)	0.66 (0.30)	0.58 (0.33)	0.85	7.33 (7.65)
		Reasoning	0.78 (0.09)	0.15 (0.16)	0.58 (0.29)	0.50 (0.31)	0.78	34.31 (15.37)
		Sampling (3)	0.82 (0.10)	0.27 (0.25)	0.67 (0.29)	0.59 (0.32)	0.82	22.66 (23.58)

文本生成

將溫度 (temperature) 調整至0.6

- Sampling 策略未達預期
 - 推測生成答案的多樣性不足，將temperature參數從0.1提高至 0.6進行對比實驗
- 指標仍未見顯著且一致的改善，其可能的原因:
 - 檢索模型可能無法有效識別並挑選出更優質的生成結果
 - TAME模型所生成的候選文本可能都與參考文本存在較大差異

System Prompt	Retrieval Type	Skills	BERT Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
TAME	Time Sorted	Base	0.82 (0.11)	0.28 (0.27)	0.62 (0.31)	0.54 (0.34)	0.85	5.65 (6.78)
		Sampling (3)	0.81 (0.10)	0.25 (0.24)	0.63 (0.30)	0.56 (0.33)	0.80	17.37 (14.84)
	Passage	Base	0.83 (0.11)	0.32 (0.29)	0.63 (0.30)	0.56 (0.33)	0.84	5.47 (5.97)
		Sampling (3)	0.80 (0.11)	0.27 (0.27)	0.63 (0.30)	0.56 (0.33)	0.77	16.09 (11.98)
	Document	Base	0.81 (0.12)	0.30 (0.29)	0.63 (0.32)	0.56 (0.34)	0.76	7.20 (10.16)
		Sampling (3)	0.81 (0.11)	0.27 (0.26)	0.65 (0.31)	0.58 (0.33)	0.78	19.46 (20.55)
Without System Prompt	Time Sorted	Base	0.82 (0.12)	0.29 (0.32)	0.68 (0.32)	0.62 (0.35)	0.73	10.12 (11.06)
		Sampling (3)	0.81 (0.11)	0.23 (0.26)	0.68 (0.29)	0.61 (0.32)	0.71	32.49 (25.49)
	Passage	Base	0.81 (0.12)	0.27 (0.29)	0.68 (0.31)	0.63 (0.34)	0.80	11.39 (12.02)
		Sampling (3)	0.80 (0.11)	0.22 (0.27)	0.67 (0.30)	0.61 (0.33)	0.73	31.45 (25.28)
	Document	Base	0.82 (0.12)	0.28 (0.29)	0.70 (0.30)	0.65 (0.33)	0.74	11.70 (12.76)
		Sampling (3)	0.81 (0.11)	0.26 (0.26)	0.71 (0.28)	0.64 (0.31)	0.81	30.36 (28.43)
Expert	Time Sorted	Base	0.82 (0.12)	0.31 (0.30)	0.65 (0.29)	0.58 (0.32)	0.78	5.91 (4.96)
		Sampling (3)	0.81 (0.10)	0.27 (0.26)	0.64 (0.29)	0.57 (0.32)	0.87	18.56 (13.59)
	Passage	Base	0.81 (0.12)	0.29 (0.29)	0.63 (0.32)	0.56 (0.35)	0.82	5.94 (6.42)
		Sampling (3)	0.81 (0.10)	0.27 (0.24)	0.64 (0.29)	0.57 (0.31)	0.86	17.29 (11.03)
	Document	Base	0.81 (0.11)	0.27 (0.27)	0.62 (0.31)	0.54 (0.33)	0.84	7.47 (9.41)
		Sampling (3)	0.81 (0.10)	0.25 (0.25)	0.64 (0.30)	0.56 (0.32)	0.84	19.54 (18.65)

文本生成

將溫度 (temperature) 調整至0.6

- 檢索模型的偏好選擇未能在評估指標上顯現。
- 不過其他指標缺乏一致性，需要進一步比較候選文本與檢索器選擇的內容，以驗證此方法的可行性。

System Prompt	Retrieval Type	Skills	BERT Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
TAME	Time Sorted	Base	0.82 (0.11)	0.28 (0.27)	0.62 (0.31)	0.54 (0.34)	0.85	5.65 (6.78)
		Sampling (3)	0.81 (0.10)	0.25 (0.24)	0.63 (0.30)	0.56 (0.33)	0.80	17.37 (14.84)
	Passage	Base	0.83 (0.11)	0.32 (0.29)	0.63 (0.30)	0.56 (0.33)	0.84	5.47 (5.97)
		Sampling (3)	0.80 (0.11)	0.27 (0.27)	0.63 (0.30)	0.56 (0.33)	0.77	16.09 (11.98)
	Document	Base	0.81 (0.12)	0.30 (0.29)	0.63 (0.32)	0.56 (0.34)	0.76	7.20 (10.16)
		Sampling (3)	0.81 (0.11)	0.27 (0.26)	0.65 (0.31)	0.58 (0.33)	0.78	19.46 (20.55)
Without System Prompt	Time Sorted	Base	0.82 (0.12)	0.29 (0.32)	0.68 (0.32)	0.62 (0.35)	0.73	10.12 (11.06)
		Sampling (3)	0.81 (0.11)	0.23 (0.26)	0.68 (0.29)	0.61 (0.32)	0.71	32.49 (25.49)
	Passage	Base	0.81 (0.12)	0.27 (0.29)	0.68 (0.31)	0.63 (0.34)	0.80	11.39 (12.02)
		Sampling (3)	0.80 (0.11)	0.22 (0.27)	0.67 (0.30)	0.61 (0.33)	0.73	31.45 (25.28)
	Document	Base	0.82 (0.12)	0.28 (0.29)	0.70 (0.30)	0.65 (0.33)	0.74	11.70 (12.76)
		Sampling (3)	0.81 (0.11)	0.26 (0.26)	0.71 (0.28)	0.64 (0.31)	0.81	30.36 (28.43)
Expert	Time Sorted	Base	0.82 (0.12)	0.31 (0.30)	0.65 (0.29)	0.58 (0.32)	0.78	5.91 (4.96)
		Sampling (3)	0.81 (0.10)	0.27 (0.26)	0.64 (0.29)	0.57 (0.32)	0.87	18.56 (13.59)
	Passage	Base	0.81 (0.12)	0.29 (0.29)	0.63 (0.32)	0.56 (0.35)	0.82	5.94 (6.42)
		Sampling (3)	0.81 (0.10)	0.27 (0.24)	0.64 (0.29)	0.57 (0.31)	0.86	17.29 (11.03)
	Document	Base	0.81 (0.11)	0.27 (0.27)	0.62 (0.31)	0.54 (0.33)	0.84	7.47 (9.41)
		Sampling (3)	0.81 (0.10)	0.25 (0.25)	0.64 (0.30)	0.56 (0.32)	0.84	19.54 (18.65)

實驗結果與分析

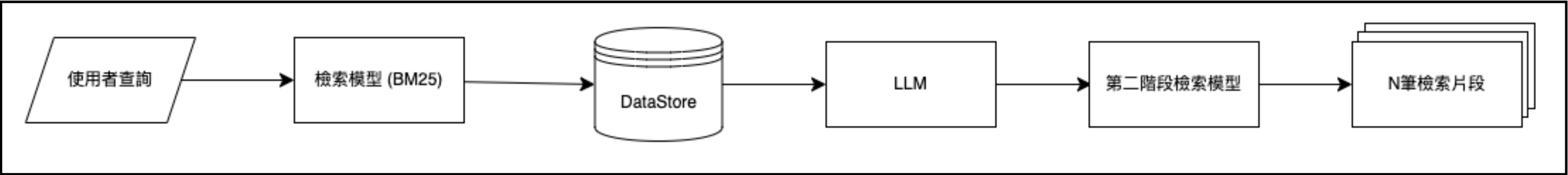
- 文本檢索
- 文本生成
- 延伸實驗

延伸實驗

文本檢索

整理摘要提示詞

<|im_start|>system
You are an AI assistant called Twllm, created by TAME (TAiwan Mixture of Expert) project.
回答時，請遵循以下原則：
1. 僅從提供的context中的內容進行回答
2. 優先提取關鍵字符合度高的段落
3. 排除重複或冗長的描述<|im_end|>
<|im_start|>context
[主要問題]
{摘要}
{客戶提報內容}
[過往解決流程]
{回覆內容}<|im_end|>
<|im_start|>user
請分析context內容，並僅列出與「{使用者查詢}」直接相關的核心資訊，
只需要提取關鍵字符合度高的段落<|im_end|>
<|im_start|>assistant

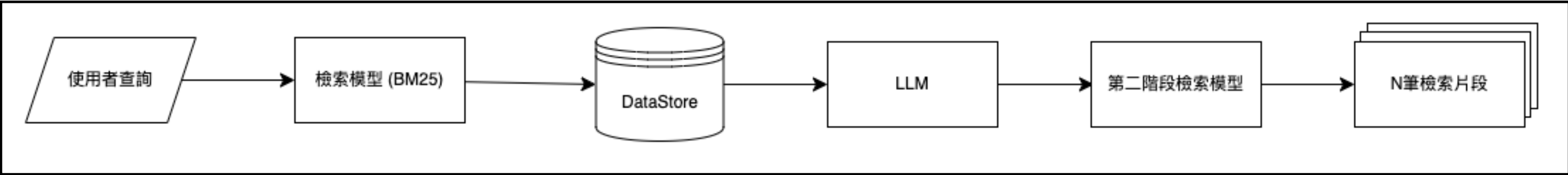


Top N	Model	Precision	Recall	F1-Score	Compression Rate
1	BM25	37.63%	36.02%	36.56%	86.34%
	Chuxin(Traditional)	48.39%	45.70%	46.59%	85.24%
	Chuxin(Simplified)	51.61%	48.92%	49.82%	85.91%
	Qwen(Traditional)	37.63%	35.13%	35.79%	86.41%
2	Qwen(Simplified)	37.63%	34.05%	35.07%	87.03%
	BM25	26.34%	49.10%	33.78%	72.30%
	Chuxin(Traditional)	34.41%	65.05%	44.62%	71.41%
	Chuxin(Simplified)	34.95%	66.13%	45.34%	71.79%
3	Qwen(Traditional)	30.11%	55.23%	38.21%	73.70%
	Qwen(Simplified)	31.72%	58.46%	40.36%	73.05%
	BM25	26.70%	70.82%	38.07%	59.82%
	Chuxin(Traditional)	29.57%	79.96%	42.48%	59.68%
4	Chuxin(Simplified)	29.93%	81.04%	43.02%	60.31%
	Qwen(Traditional)	26.70%	71.90%	38.29%	61.36%
	Qwen(Simplified)	25.63%	68.14%	36.57%	60.93%
	BM25	25.54%	83.41%	38.22%	50.09%
5	Chuxin(Traditional)	27.15%	89.50%	40.66%	49.46%
	Chuxin(Simplified)	27.42%	90.04%	41.02%	49.93%
	Qwen(Traditional)	24.46%	80.18%	36.64%	50.92%
	Qwen(Simplified)	24.73%	80.72%	37.00%	51.10%
6	BM25	23.82%	87.92%	36.41%	42.45%
	Chuxin(Traditional)	25.54%	95.09%	39.06%	41.27%
	Chuxin(Simplified)	25.32%	94.55%	38.76%	41.44%
	Qwen(Traditional)	23.82%	88.46%	36.46%	43.38%
7	Qwen(Simplified)	23.60%	86.85%	36.05%	42.92%
	BM25	22.71%	90.82%	35.09%	36.51%
	Chuxin(Traditional)	24.14%	97.99%	37.38%	34.99%
	Chuxin(Simplified)	23.78%	97.24%	36.91%	34.50%
8	Qwen(Traditional)	23.06%	93.84%	35.81%	36.27%
	Qwen(Simplified)	22.17%	89.00%	34.31%	36.14%
	BM25	21.94%	94.05%	34.22%	30.01%
	Chuxin(Traditional)	22.71%	97.99%	35.43%	29.17%
9	Chuxin(Simplified)	22.71%	97.99%	35.43%	28.96%
	Qwen(Traditional)	22.09%	95.13%	34.49%	30.92%
	Qwen(Simplified)	21.32%	90.29%	33.18%	31.25%
	BM25	21.65%	97.81%	33.99%	25.78%
10	Chuxin(Traditional)	22.05%	99.28%	34.55%	25.58%
	Chuxin(Simplified)	22.05%	99.28%	34.55%	26.05%
	Qwen(Traditional)	21.25%	95.13%	33.30%	27.45%
	Qwen(Simplified)	21.11%	94.59%	33.09%	27.81%
11	BM25	21.17%	98.03%	33.29%	22.52%
	Chuxin(Traditional)	21.41%	99.28%	33.65%	22.12%
	Chuxin(Simplified)	21.41%	99.28%	33.65%	22.50%
	Qwen(Traditional)	20.81%	96.20%	32.72%	24.51%
12	Qwen(Simplified)	20.57%	94.59%	32.31%	25.25%
	BM25	20.78%	98.03%	32.73%	20.12%
	Chuxin(Traditional)	20.99%	99.28%	33.06%	20.32%
	Chuxin(Simplified)	20.99%	99.28%	33.06%	20.41%
13	Qwen(Traditional)	20.46%	96.20%	32.21%	22.23%
	Qwen(Simplified)	20.46%	96.20%	32.21%	22.90%

延伸實驗

文本檢索

- Recall閥值控制在95%的條件下，相較於原先設置，該方法使其F1-Score上升3.97%, Compression Rate則提升了11.59%
- 潛在問題
 - 檢索效果受TAME-8B的回應內容所影響，且沒有帶來一致性的檢索性能提升，其檢索的穩定性與模型的適用範圍仍需進一步驗證。

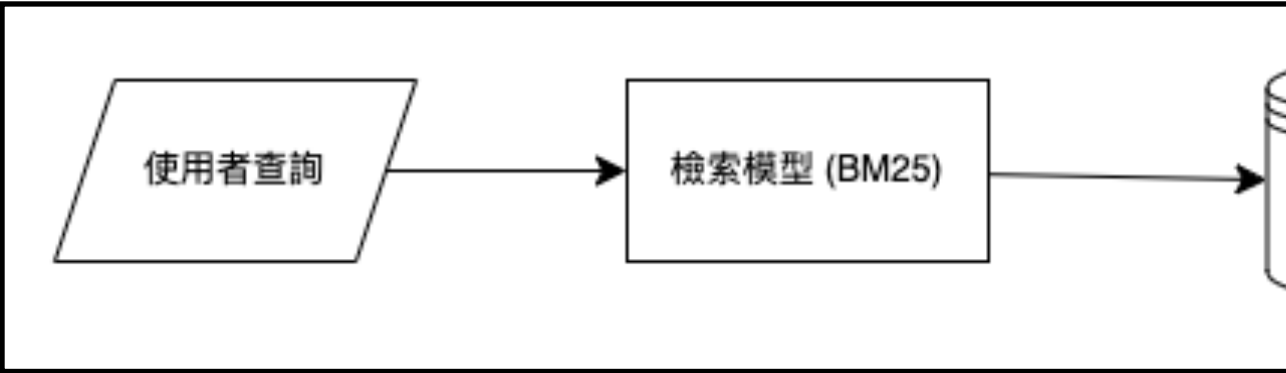


Top N	Model	Precision	Recall	F1-Score	Compression Rate
1	BM25	37.63%	36.02%	36.56%	86.34%
	Chuxin(Traditional)	48.39%	45.70%	46.59%	85.24%
	Chuxin(Simplified)	51.61%	48.92%	49.82%	85.91%
	Qwen(Traditional)	37.63%	35.13%	35.79%	86.41%
2	Qwen(Simplified)	37.63%	34.05%	35.07%	87.03%
	BM25	26.34%	49.10%	33.78%	72.30%
	Chuxin(Traditional)	34.41%	65.05%	44.62%	71.41%
	Chuxin(Simplified)	34.95%	66.13%	45.34%	71.79%
3	Qwen(Traditional)	30.11%	55.23%	38.21%	73.70%
	Qwen(Simplified)	31.72%	58.46%	40.36%	73.05%
	BM25	26.70%	70.82%	38.07%	59.82%
	Chuxin(Traditional)	29.57%	79.96%	42.48%	59.68%
4	Chuxin(Simplified)	29.93%	81.04%	43.02%	60.31%
	Qwen(Traditional)	26.70%	71.90%	38.29%	61.36%
	Qwen(Simplified)	25.63%	68.14%	36.57%	60.93%
	BM25	25.54%	83.41%	38.22%	50.09%
5	Chuxin(Traditional)	27.15%	89.50%	40.66%	49.46%
	Chuxin(Simplified)	27.42%	90.04%	41.02%	49.93%
	Qwen(Traditional)	24.46%	80.18%	36.64%	50.92%
	Qwen(Simplified)	24.73%	80.72%	37.00%	51.10%
6	BM25	23.82%	87.92%	36.41%	42.45%
	Chuxin(Traditional)	25.54%	95.09%	39.06%	41.27%
	Chuxin(Simplified)	25.32%	94.55%	38.76%	41.44%
	Qwen(Traditional)	23.82%	88.46%	36.46%	43.38%
7	Qwen(Simplified)	23.60%	86.85%	36.05%	42.92%
	BM25	22.71%	90.82%	35.09%	36.51%
	Chuxin(Traditional)	24.14%	97.99%	37.38%	34.99%
	Chuxin(Simplified)	23.78%	97.24%	36.91%	34.50%
8	Qwen(Traditional)	23.06%	93.84%	35.81%	36.27%
	Qwen(Simplified)	22.17%	89.00%	34.31%	36.14%
	BM25	21.94%	94.05%	34.22%	30.01%
	Chuxin(Traditional)	22.71%	97.99%	35.43%	29.17%
9	Chuxin(Simplified)	22.71%	97.99%	35.43%	28.96%
	Qwen(Traditional)	22.09%	95.13%	34.49%	30.92%
	Qwen(Simplified)	21.32%	90.29%	33.18%	31.25%
	BM25	21.65%	97.81%	33.99%	25.78%
10	Chuxin(Traditional)	22.05%	99.28%	34.55%	25.58%
	Chuxin(Simplified)	22.05%	99.28%	34.55%	26.05%
	Qwen(Traditional)	21.25%	95.13%	33.30%	27.45%
	Qwen(Simplified)	21.11%	94.59%	33.09%	27.81%
11	BM25	21.17%	98.03%	33.29%	22.52%
	Chuxin(Traditional)	21.41%	99.28%	33.65%	22.12%
	Chuxin(Simplified)	21.41%	99.28%	33.65%	22.50%
	Qwen(Traditional)	20.81%	96.20%	32.72%	24.51%
12	Qwen(Simplified)	20.57%	94.59%	32.31%	25.25%
	BM25	20.78%	98.03%	32.73%	20.12%
	Chuxin(Traditional)	20.99%	99.28%	33.06%	20.32%
	Chuxin(Simplified)	20.99%	99.28%	33.06%	20.41%
13	Qwen(Traditional)	20.46%	96.20%	32.21%	22.23%
	Qwen(Simplified)	20.46%	96.20%	32.21%	22.90%

延伸實驗

文本檢索

- Recall閥值控制在95%的範圍內，F1-Score上升3.97%，Compression Rate提升1.15%
- 潛在問題
 - 檢索效果受TAME-8B的檢索穩定性影響



Top N	Model	Precision	Recall	F1-Score	Compression Rate
1	BM25	32.26%	29.21%	30.06%	88.15%
	Chuxin(Traditional)	41.94%	38.89%	39.73%	87.45%
	Chuxin(Simplified)	40.86%	38.35%	39.02%	87.48%
	Qwen(Traditional)	45.16%	42.11%	42.96%	85.69%
	Qwen(Simplified)	48.39%	44.80%	45.83%	86.44%
2	BM25	25.27%	45.88%	31.99%	75.92%
	Chuxin(Traditional)	32.80%	60.93%	42.03%	74.23%
	Chuxin(Simplified)	31.18%	57.71%	39.87%	74.74%
	Qwen(Traditional)	32.26%	60.39%	41.49%	72.49%
	Qwen(Simplified)	32.26%	59.86%	41.31%	72.91%
3	BM25	25.99%	69.75%	37.21%	62.75%
	Chuxin(Traditional)	27.78%	74.23%	39.60%	62.17%
	Chuxin(Simplified)	27.06%	72.08%	38.52%	62.94%
	Qwen(Traditional)	28.85%	76.92%	41.11%	58.91%
	Qwen(Simplified)	28.14%	74.77%	40.03%	59.91%
4	BM25	23.92%	78.03%	35.78%	52.22%
	Chuxin(Traditional)	24.73%	80.36%	36.86%	53.45%
	Chuxin(Simplified)	24.73%	80.36%	36.86%	53.27%
	Qwen(Traditional)	25.27%	82.51%	37.72%	48.76%
	Qwen(Simplified)	25.27%	82.51%	37.72%	49.71%
5	BM25	23.17%	85.23%	35.39%	43.84%
	Chuxin(Traditional)	23.82%	88.10%	36.35%	45.68%
	Chuxin(Simplified)	23.17%	84.87%	35.28%	45.10%
	Qwen(Traditional)	24.03%	87.74%	36.50%	41.68%
	Qwen(Simplified)	24.03%	89.18%	36.71%	41.87%
6	BM25	22.71%	91.36%	35.12%	36.72%
	Chuxin(Traditional)	23.42%	94.77%	36.22%	37.70%
	Chuxin(Simplified)	23.42%	95.09%	36.30%	37.51%
	Qwen(Traditional)	23.24%	92.80%	35.79%	35.56%
	Qwen(Simplified)	23.24%	93.33%	35.83%	35.07%
7	BM25	22.09%	94.23%	34.39%	31.04%
	Chuxin(Traditional)	22.25%	95.30%	34.66%	31.90%
	Chuxin(Simplified)	22.25%	95.30%	34.66%	31.77%
	Qwen(Traditional)	22.55%	96.56%	35.09%	29.68%
	Qwen(Simplified)	22.55%	96.56%	35.09%	29.46%
8	BM25	21.52%	95.48%	33.61%	27.03%
	Chuxin(Traditional)	21.65%	96.59%	33.86%	27.41%
	Chuxin(Simplified)	21.65%	96.59%	33.86%	27.49%
	Qwen(Traditional)	21.65%	96.56%	33.85%	25.78%
	Qwen(Simplified)	21.78%	97.63%	34.09%	25.35%
9	BM25	21.17%	97.63%	33.23%	23.73%
	Chuxin(Traditional)	21.29%	97.85%	33.38%	23.79%
	Chuxin(Simplified)	21.17%	97.67%	33.24%	24.69%
	Qwen(Traditional)	21.17%	97.63%	33.23%	22.51%
	Qwen(Simplified)	21.17%	97.63%	33.23%	22.43%
10	BM25	20.89%	97.85%	32.82%	20.58%
	Chuxin(Traditional)	20.89%	97.85%	32.82%	21.49%
	Chuxin(Simplified)	20.78%	97.67%	32.68%	21.96%
	Qwen(Traditional)	20.89%	98.17%	32.85%	20.04%
	Qwen(Simplified)	20.89%	98.17%	32.85%	20.06%

去使其

檢索性能

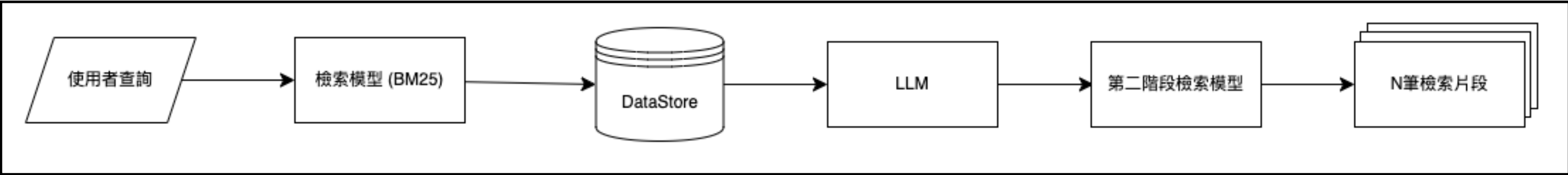


Top N	Model	Precision	Recall	F1-Score	Compression Rate
1	BM25	37.63%	36.02%	36.56%	86.34%
	Chuxin(Traditional)	48.39%	45.70%	46.59%	85.24%
	Chuxin(Simplified)	51.61%	48.92%	49.82%	85.91%
	Qwen(Traditional)	37.63%	35.13%	35.79%	86.41%
	Qwen(Simplified)	37.63%	34.05%	35.07%	87.03%
2	BM25	26.34%	49.10%	33.78%	72.30%
	Chuxin(Traditional)	34.41%	65.05%	44.62%	71.41%
	Chuxin(Simplified)	34.95%	66.13%	45.34%	71.79%
	Qwen(Traditional)	30.11%	55.23%	38.21%	73.70%
	Qwen(Simplified)	31.72%	58.46%	40.36%	73.05%
3	BM25	26.70%	70.82%	38.07%	59.82%
	Chuxin(Traditional)	29.57%	79.96%	42.48%	59.68%
	Chuxin(Simplified)	29.93%	81.04%	43.02%	60.31%
	Qwen(Traditional)	26.70%	71.90%	38.29%	61.36%
	Qwen(Simplified)	25.63%	68.14%	36.57%	60.93%
4	BM25	25.54%	83.41%	38.22%	50.09%
	Chuxin(Traditional)	27.15%	89.50%	40.66%	49.46%
	Chuxin(Simplified)	27.42%	90.04%	41.02%	49.93%
	Qwen(Traditional)	24.46%	80.18%	36.64%	50.92%
	Qwen(Simplified)	24.73%	80.72%	37.00%	51.10%
5	BM25	23.82%	87.92%	36.41%	42.45%
	Chuxin(Traditional)	25.54%	95.09%	39.06%	41.27%
	Chuxin(Simplified)	25.32%	94.55%	38.76%	41.44%
	Qwen(Traditional)	23.82%	88.46%	36.46%	43.38%
	Qwen(Simplified)	23.60%	86.85%	36.05%	42.92%
6	BM25	22.71%	90.82%	35.09%	36.51%
	Chuxin(Traditional)	24.14%	97.99%	37.38%	34.99%
	Chuxin(Simplified)	23.78%	97.24%	36.91%	34.50%
	Qwen(Traditional)	23.06%	93.84%	35.81%	36.27%
	Qwen(Simplified)	22.17%	89.00%	34.31%	36.14%
7	BM25	21.94%	94.05%	34.22%	30.01%
	Chuxin(Traditional)	22.71%	97.99%	35.43%	29.17%
	Chuxin(Simplified)	22.71%	97.99%	35.43%	28.96%
	Qwen(Traditional)	22.09%	95.13%	34.49%	30.92%
	Qwen(Simplified)	21.32%	90.29%	33.18%	31.25%
8	BM25	21.65%	97.81%	33.99%	25.78%
	Chuxin(Traditional)	22.05%	99.28%	34.55%	25.58%
	Chuxin(Simplified)	22.05%	99.28%	34.55%	26.05%
	Qwen(Traditional)	21.25%	95.13%	33.30%	27.45%
	Qwen(Simplified)	21.11%	94.59%	33.09%	27.81%
9	BM25	21.17%	98.03%	33.29%	22.52%
	Chuxin(Traditional)	21.41%	99.28%	33.65%	22.12%
	Chuxin(Simplified)	21.41%	99.28%	33.65%	22.50%
	Qwen(Traditional)	20.81%	96.20%	32.72%	24.51%
	Qwen(Simplified)	20.57%	94.59%	32.31%	25.25%
10	BM25	20.78%	98.03%	32.73%	20.12%
	Chuxin(Traditional)	20.99%	99.28%	33.06%	20.32%
	Chuxin(Simplified)	20.99%	99.28%	33.06%	20.41%
	Qwen(Traditional)	20.46%	96.20%	32.21%	22.23%
	Qwen(Simplified)	20.46%	96.20%	32.21%	22.90%

延伸實驗

文本檢索

- Recall閥值控制在95%的條件下，相較於原先設置，該方法使其F1-Score上升3.97%, Compression Rate則提升了11.59%
- 潛在問題
 - 檢索效果受TAME-8B的回應內容所影響，且沒有帶來一致性的檢索性能提升，其檢索的穩定性與模型的適用範圍仍需進一步驗證。

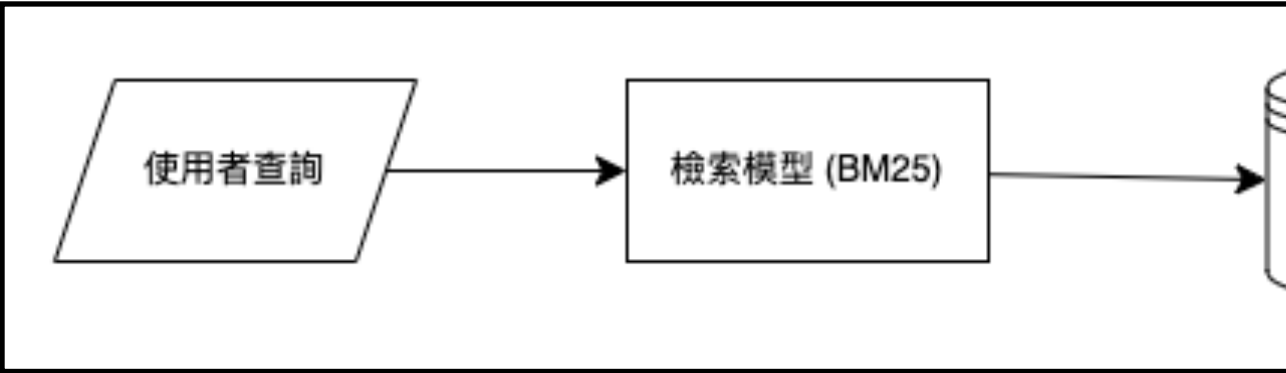


Top N	Model	Precision	Recall	F1-Score	Compression Rate
1	BM25	37.63%	36.02%	36.56%	86.34%
	Chuxin(Traditional)	48.39%	45.70%	46.59%	85.24%
	Chuxin(Simplified)	51.61%	48.92%	49.82%	85.91%
	Qwen(Traditional)	37.63%	35.13%	35.79%	86.41%
2	Qwen(Simplified)	37.63%	34.05%	35.07%	87.03%
	BM25	26.34%	49.10%	33.78%	72.30%
	Chuxin(Traditional)	34.41%	65.05%	44.62%	71.41%
	Chuxin(Simplified)	34.95%	66.13%	45.34%	71.79%
3	Qwen(Traditional)	30.11%	55.23%	38.21%	73.70%
	Qwen(Simplified)	31.72%	58.46%	40.36%	73.05%
	BM25	26.70%	70.82%	38.07%	59.82%
	Chuxin(Traditional)	29.57%	79.96%	42.48%	59.68%
4	Chuxin(Simplified)	29.93%	81.04%	43.02%	60.31%
	Qwen(Traditional)	26.70%	71.90%	38.29%	61.36%
	Qwen(Simplified)	25.63%	68.14%	36.57%	60.93%
	BM25	25.54%	83.41%	38.22%	50.09%
5	Chuxin(Traditional)	27.15%	89.50%	40.66%	49.46%
	Chuxin(Simplified)	27.42%	90.04%	41.02%	49.93%
	Qwen(Traditional)	24.46%	80.18%	36.64%	50.92%
	Qwen(Simplified)	24.73%	80.72%	37.00%	51.10%
6	BM25	23.82%	87.92%	36.41%	42.45%
	Chuxin(Traditional)	25.54%	95.09%	39.06%	41.27%
	Chuxin(Simplified)	25.32%	94.55%	38.76%	41.44%
	Qwen(Traditional)	23.82%	88.46%	36.46%	43.38%
7	Qwen(Simplified)	23.60%	86.85%	36.05%	42.92%
	BM25	22.71%	90.82%	35.09%	36.51%
	Chuxin(Traditional)	24.14%	97.99%	37.38%	34.99%
	Chuxin(Simplified)	23.78%	97.24%	36.91%	34.50%
8	Qwen(Traditional)	23.06%	93.84%	35.81%	36.27%
	Qwen(Simplified)	22.17%	89.00%	34.31%	36.14%
	BM25	21.94%	94.05%	34.22%	30.01%
	Chuxin(Traditional)	22.71%	97.99%	35.43%	29.17%
9	Chuxin(Simplified)	22.71%	97.99%	35.43%	28.96%
	Qwen(Traditional)	22.09%	95.13%	34.49%	30.92%
	Qwen(Simplified)	21.32%	90.29%	33.18%	31.25%
	BM25	21.65%	97.81%	33.99%	25.78%
10	Chuxin(Traditional)	22.05%	99.28%	34.55%	25.58%
	Chuxin(Simplified)	22.05%	99.28%	34.55%	26.05%
	Qwen(Traditional)	21.25%	95.13%	33.30%	27.45%
	Qwen(Simplified)	21.11%	94.59%	33.09%	27.81%
11	BM25	21.17%	98.03%	33.29%	22.52%
	Chuxin(Traditional)	21.41%	99.28%	33.65%	22.12%
	Chuxin(Simplified)	21.41%	99.28%	33.65%	22.50%
	Qwen(Traditional)	20.81%	96.20%	32.72%	24.51%
12	Qwen(Simplified)	20.57%	94.59%	32.31%	25.25%
	BM25	20.78%	98.03%	32.73%	20.12%
	Chuxin(Traditional)	20.99%	99.28%	33.06%	20.32%
	Chuxin(Simplified)	20.99%	99.28%	33.06%	20.41%
13	Qwen(Traditional)	20.46%	96.20%	32.21%	22.23%
	Qwen(Simplified)	20.46%	96.20%	32.21%	22.90%

延伸實驗

文本檢索

- Recall閥值控制在95%的範圍內，F1-Score上升3.97%，Compression Rate提升1.15%
- 潛在問題
 - 檢索效果受TAME-8B的檢索穩定性影響



Top N	Model	Precision	Recall	F1-Score	Compression Rate
1	BM25	32.26%	29.21%	30.06%	88.15%
	Chuxin(Traditional)	41.94%	38.89%	39.73%	87.45%
	Chuxin(Simplified)	40.86%	38.35%	39.02%	87.48%
	Qwen(Traditional)	45.16%	42.11%	42.96%	85.69%
	Qwen(Simplified)	48.39%	44.80%	45.83%	86.44%
2	BM25	25.27%	45.88%	31.99%	75.92%
	Chuxin(Traditional)	32.80%	60.93%	42.03%	74.23%
	Chuxin(Simplified)	31.18%	57.71%	39.87%	74.74%
	Qwen(Traditional)	32.26%	60.39%	41.49%	72.49%
	Qwen(Simplified)	32.26%	59.86%	41.31%	72.91%
3	BM25	25.99%	69.75%	37.21%	62.75%
	Chuxin(Traditional)	27.78%	74.23%	39.60%	62.17%
	Chuxin(Simplified)	27.06%	72.08%	38.52%	62.94%
	Qwen(Traditional)	28.85%	76.92%	41.11%	58.91%
	Qwen(Simplified)	28.14%	74.77%	40.03%	59.91%
4	BM25	23.92%	78.03%	35.78%	52.22%
	Chuxin(Traditional)	24.73%	80.36%	36.86%	53.45%
	Chuxin(Simplified)	24.73%	80.36%	36.86%	53.27%
	Qwen(Traditional)	25.27%	82.51%	37.72%	48.76%
	Qwen(Simplified)	25.27%	82.51%	37.72%	49.71%
5	BM25	23.17%	85.23%	35.39%	43.84%
	Chuxin(Traditional)	23.82%	88.10%	36.35%	45.68%
	Chuxin(Simplified)	23.17%	84.87%	35.28%	45.10%
	Qwen(Traditional)	24.03%	87.74%	36.50%	41.68%
	Qwen(Simplified)	24.03%	89.18%	36.71%	41.87%
6	BM25	22.71%	91.36%	35.12%	36.72%
	Chuxin(Traditional)	23.42%	94.77%	36.22%	37.70%
	Chuxin(Simplified)	23.42%	95.09%	36.30%	37.51%
	Qwen(Traditional)	23.24%	92.80%	35.79%	35.56%
	Qwen(Simplified)	23.24%	93.33%	35.83%	35.07%
7	BM25	22.09%	94.23%	34.39%	31.04%
	Chuxin(Traditional)	22.25%	95.30%	34.66%	31.90%
	Chuxin(Simplified)	22.25%	95.30%	34.66%	31.77%
	Qwen(Traditional)	22.55%	96.56%	35.09%	29.68%
	Qwen(Simplified)	22.55%	96.56%	35.09%	29.46%
8	BM25	21.52%	95.48%	33.61%	27.03%
	Chuxin(Traditional)	21.65%	96.59%	33.86%	27.41%
	Chuxin(Simplified)	21.65%	96.59%	33.86%	27.49%
	Qwen(Traditional)	21.65%	96.56%	33.85%	25.78%
	Qwen(Simplified)	21.78%	97.63%	34.09%	25.35%
9	BM25	21.17%	97.63%	33.23%	23.73%
	Chuxin(Traditional)	21.29%	97.85%	33.38%	23.79%
	Chuxin(Simplified)	21.17%	97.67%	33.24%	24.69%
	Qwen(Traditional)	21.17%	97.63%	33.23%	22.51%
	Qwen(Simplified)	21.17%	97.63%	33.23%	22.43%
10	BM25	20.89%	97.85%	32.82%	20.58%
	Chuxin(Traditional)	20.89%	97.85%	32.82%	21.49%
	Chuxin(Simplified)	20.78%	97.67%	32.68%	21.96%
	Qwen(Traditional)	20.89%	98.17%	32.85%	20.04%
	Qwen(Simplified)	20.89%	98.17%	32.85%	20.06%

去使其

檢索性能



Top N	Model	Precision	Recall	F1-Score	Compression Rate
1	BM25	37.63%	36.02%	36.56%	86.34%
	Chuxin(Traditional)	48.39%	45.70%	46.59%	85.24%
	Chuxin(Simplified)	51.61%	48.92%	49.82%	85.91%
	Qwen(Traditional)	37.63%	35.13%	35.79%	86.41%
	Qwen(Simplified)	37.63%	34.05%	35.07%	87.03%
2	BM25	26.34%	49.10%	33.78%	72.30%
	Chuxin(Traditional)	34.41%	65.05%	44.62%	71.41%
	Chuxin(Simplified)	34.95%	66.13%	45.34%	71.79%
	Qwen(Traditional)	30.11%	55.23%	38.21%	73.70%
	Qwen(Simplified)	31.72%	58.46%	40.36%	73.05%
3	BM25	26.70%	70.82%	38.07%	59.82%
	Chuxin(Traditional)	29.57%	79.96%	42.48%	59.68%
	Chuxin(Simplified)	29.93%	81.04%	43.02%	60.31%
	Qwen(Traditional)	26.70%	71.90%	38.29%	61.36%
	Qwen(Simplified)	25.63%	68.14%	36.57%	60.93%
4	BM25	25.54%	83.41%	38.22%	50.09%
	Chuxin(Traditional)	27.15%	89.50%	40.66%	49.46%
	Chuxin(Simplified)	27.42%	90.04%	41.02%	49.93%
	Qwen(Traditional)	24.46%	80.18%	36.64%	50.92%
	Qwen(Simplified)	24.73%	80.72%	37.00%	51.10%
5	BM25	23.82%	87.92%	36.41%	42.45%
	Chuxin(Traditional)	25.54%	95.09%	39.06%	41.27%
	Chuxin(Simplified)	25.32%	94.55%	38.76%	41.44%
	Qwen(Traditional)	23.82%	88.46%	36.46%	43.38%
	Qwen(Simplified)	23.60%	86.85%	36.05%	42.92%
6	BM25	22.71%	90.82%	35.09%	36.51%
	Chuxin(Traditional)	24.14%	97.99%	37.38%	34.99%
	Chuxin(Simplified)	23.78%	97.24%	36.91%	34.50%
	Qwen(Traditional)	23.06%	93.84%	35.81%	36.27%
	Qwen(Simplified)	22.17%	89.00%	34.31%	36.14%
7	BM25	21.94%	94.05%	34.22%	30.01%
	Chuxin(Traditional)	22.71%	97.99%	35.43%	29.17%
	Chuxin(Simplified)	22.71%	97.99%	35.43%	28.96%
	Qwen(Traditional)	22.09%	95.13%	34.49%	30.92%
	Qwen(Simplified)	21.32%	90.29%	33.18%	31.25%
8	BM25	21.65%	97.81%	33.99%	25.78%
	Chuxin(Traditional)	22.05%	99.28%	34.55%	25.58%
	Chuxin(Simplified)	22.05%	99.28%	34.55%	26.05%
	Qwen(Traditional)	21.25%	95.13%	33.30%	27.45%
	Qwen(Simplified)	21.11%	94.59%	33.09%	27.81%
9	BM25	21.17%	98.03%	33.29%	22.52%
	Chuxin(Traditional)	21.41%	99.28%	33.65%	22.12%
	Chuxin(Simplified)	21.41%	99.28%	33.65%	22.50%
	Qwen(Traditional)	20.81%	96.20%	32.72%	24.51%
	Qwen(Simplified)	20.57%	94.59%	32.31%	25.25%
10	BM25	20.78%	98.03%	32.73%	20.12%
	Chuxin(Traditional)	20.99%	99.28%	33.06%	20.32%
	Chuxin(Simplified)	20.99%	99.28%	33.06%	20.41%
	Qwen(Traditional)	20.46%	96.20%	32.21%	22.23%
	Qwen(Simplified)	20.46%	96.20%	32.21%	22.90%

延伸實驗

文本檢索 (加上引導詞)

- 基於原提示詞尾部加入引導詞
- 相較於未加入引導詞的實驗結果，各模型在檢索性能上並未顯示出特別的性能優勢。

整理摘要提示詞

<|im_start|>system
You are an AI assistant called Twllm, created by TAME (TAiwan Mixture of Expert) project.
回答時，請遵循以下原則:
1. 僅從提供的context中的內容進行回答
2. 優先提取關鍵字符合度高的段落
3. 排除重複或冗長的描述<|im_end|>
<|im_start|>context
[主要問題]
{摘要}
{客戶提報內容}
[過往解決流程]
{回覆內容}<|im_end|>
<|im_start|>user
請分析context內容，並僅列出與「{使用者查詢}」直接相關的核心資訊，只需要提取關鍵字符合度高的段落<|im_end|>
<|im_start|>assistant
根據提供的context內容，與「{使用者查詢}」直接相關的核心資訊如下:

Top N	Model	Precision	Recall	F1-Score	Compression Rate
1	BM25	36.56%	34.95%	35.48%	85.41%
	Chuxin(Traditional)	46.24%	43.55%	44.44%	86.06%
	Chuxin(Simplified)	46.24%	43.55%	44.44%	86.17%
	Qwen(Traditional)	38.71%	36.56%	37.28%	86.10%
	Qwen(Simplified)	37.63%	35.48%	36.20%	86.10%
2	BM25	30.11%	56.99%	39.07%	71.56%
	Chuxin(Traditional)	35.48%	67.20%	46.06%	72.10%
	Chuxin(Simplified)	35.48%	66.67%	45.88%	71.99%
	Qwen(Traditional)	31.18%	57.74%	39.91%	72.67%
	Qwen(Simplified)	32.80%	59.00%	41.26%	72.29%
3	BM25	26.70%	72.80%	38.58%	59.73%
	Chuxin(Traditional)	28.49%	76.56%	40.95%	60.30%
	Chuxin(Simplified)	28.85%	78.17%	41.59%	60.76%
	Qwen(Traditional)	27.06%	72.44%	38.72%	60.67%
	Qwen(Simplified)	27.78%	73.51%	39.58%	60.38%
4	BM25	25.00%	82.69%	37.65%	50.45%
	Chuxin(Traditional)	25.81%	85.38%	38.86%	50.59%
	Chuxin(Simplified)	25.81%	85.91%	38.94%	50.53%
	Qwen(Traditional)	25.00%	81.25%	37.36%	49.75%
	Qwen(Simplified)	24.73%	80.18%	36.93%	49.63%
5	BM25	23.82%	88.82%	36.57%	42.12%
	Chuxin(Traditional)	24.68%	93.12%	38.01%	41.94%
	Chuxin(Simplified)	25.11%	94.37%	38.56%	42.69%
	Qwen(Traditional)	23.17%	84.70%	35.34%	42.44%
	Qwen(Simplified)	23.39%	85.23%	35.64%	41.34%
6	BM25	22.89%	92.44%	35.43%	35.53%
	Chuxin(Traditional)	23.60%	96.20%	36.62%	34.71%
	Chuxin(Simplified)	23.42%	95.99%	36.43%	34.63%
	Qwen(Traditional)	22.17%	88.46%	34.28%	35.53%
	Qwen(Simplified)	22.53%	90.61%	34.89%	35.16%
7	BM25	22.09%	94.80%	34.43%	30.22%
	Chuxin(Traditional)	22.55%	97.49%	35.21%	29.12%
	Chuxin(Simplified)	22.40%	97.28%	35.03%	29.51%
	Qwen(Traditional)	21.94%	94.91%	34.31%	30.10%
	Qwen(Simplified)	21.94%	94.05%	34.22%	30.10%
8	BM25	21.38%	95.88%	33.49%	25.97%
	Chuxin(Traditional)	22.05%	99.28%	34.55%	25.01%
	Chuxin(Simplified)	21.92%	99.10%	34.40%	25.06%
	Qwen(Traditional)	21.38%	96.20%	33.54%	25.65%
	Qwen(Simplified)	21.25%	95.13%	33.30%	25.52%
9	BM25	20.93%	96.95%	32.90%	22.72%
	Chuxin(Traditional)	21.41%	99.28%	33.65%	22.06%
	Chuxin(Simplified)	21.29%	99.10%	33.50%	22.09%
	Qwen(Traditional)	20.93%	97.28%	32.94%	23.24%
	Qwen(Simplified)	20.81%	96.20%	32.72%	22.31%
10	BM25	20.67%	97.49%	32.55%	19.79%
	Chuxin(Traditional)	21.10%	99.46%	33.19%	20.70%
	Chuxin(Simplified)	20.99%	99.28%	33.06%	20.47%
	Qwen(Traditional)	20.56%	97.28%	32.40%	21.41%
	Qwen(Simplified)	20.56%	97.28%	32.40%	20.58%

延伸實驗

文本生成 (加上引導詞)

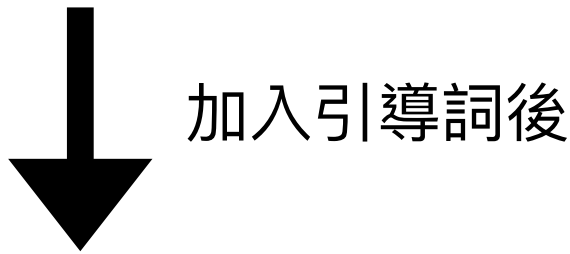
- 將對應的系統提示詞尾部加入「與「{使用者查詢}」高度相關的解決方案如下:」作為引導詞，進行對照實驗
- 具有系統提示的情況下，加入引導詞後，TAME在BERT Score、BLEU Score、ROUGE-1及ROUGE-L等評估指標上均顯著提升。
- 無系統提示 (Without System Prompt)的情況下，加入引導詞反而導致性能全面下降。這可能是干擾TAME-8B從檢索內容中模仿其回應內容的能力，從而導致重點擷取能力下降。
- 這現象突顯了系統提示與引導詞之間的效應:

明確的任務指示配合相應的引導詞，能夠有效提升模型性能。

與Kojima等人(2022)的研究結果相互呼應，僅透過在提示詞尾部加入「Let's think step by step.」的簡單引導，即能顯著提升LLMs的推理能力。

System Prompt	Retrieval Type	Bert Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
TAME	Time Sorted	0.84 (0.12)	0.38 (0.31)	0.75 (0.29)	0.70 (0.33)	0.78	6.55 (6.56)
	Passage	0.84 (0.11)	0.36 (0.29)	0.76 (0.29)	0.70 (0.33)	0.78	7.19 (7.74)
	Document	0.84 (0.13)	0.36 (0.29)	0.74 (0.30)	0.69 (0.33)	0.77	9.08 (11.01)
Without System Prompt	Time Sorted	0.77 (0.11)	0.15 (0.19)	0.69 (0.34)	0.64 (0.35)	0.73	20.05 (14.78)
	Passage	0.76 (0.12)	0.14 (0.19)	0.67 (0.35)	0.63 (0.36)	0.62	21.58 (15.31)
	Document	0.76 (0.12)	0.16 (0.20)	0.67 (0.34)	0.63 (0.35)	0.60	21.40 (15.18)
Expert	Time Sorted	0.84 (0.11)	0.35 (0.28)	0.76 (0.27)	0.70 (0.30)	0.84	7.22 (7.13)
	Passage	0.84 (0.10)	0.33 (0.27)	0.76 (0.28)	0.71 (0.32)	0.77	7.59 (7.46)
	Document	0.83 (0.11)	0.31 (0.27)	0.72 (0.29)	0.67 (0.32)	0.80	10.03 (11.83)

System Prompt	BERT Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
TAME	0.81	0.30	0.62	0.55	0.81	5.74
Without System	0.82	0.31	0.71	0.65	0.75	11.40
Expert	0.82	0.29	0.66	0.58	0.85	6.41



加入引導詞後

System Prompt	BERT Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
TAME	0.84	0.37	0.75	0.70	0.78	7.61
Without System	0.77	0.15	0.68	0.63	0.65	21.01
Expert	0.84	0.33	0.75	0.69	0.80	8.28

延伸實驗

文本生成 (加上引導詞)

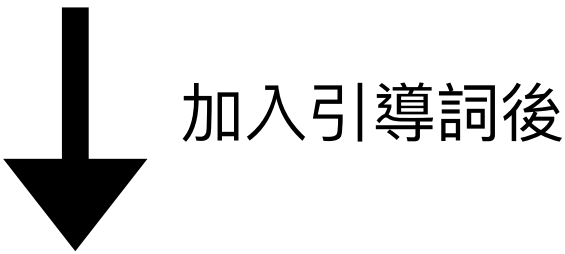
- 將對應的系統提示詞尾部加入「與「{使用者查詢}」高度相關的解決方案如下:」作為引導詞，進行對照實驗
- 具有系統提示的情況下，加入引導詞後，TAME在BERT Score、BLEU Score、ROUGE-1及ROUGE-L等評估指標上均顯著提升。
- 無系統提示 (Without System Prompt)的情況下，加入引導詞反而導致性能全面下降。這可能是干擾TAME-8B從檢索內容中模仿其回應內容的能力，從而導致重點擷取能力下降。
- 這現象突顯了系統提示與引導詞之間的效應:

明確的任務指示配合相應的引導詞，能夠有效提升模型性能。

與Kojima等人(2022)的研究結果相互呼應，僅透過在提示詞尾部加入「Let's think step by step.」的簡單引導，即能顯著提升LLMs的推理能力。

System Prompt	Retrieval Type	Bert Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
TAME	Time Sorted	0.84 (0.12)	0.38 (0.31)	0.75 (0.29)	0.70 (0.33)	0.78	6.55 (6.56)
	Passage	0.84 (0.11)	0.36 (0.29)	0.76 (0.29)	0.70 (0.33)	0.78	7.19 (7.74)
	Document	0.84 (0.13)	0.36 (0.29)	0.74 (0.30)	0.69 (0.33)	0.77	9.08 (11.01)
Without System Prompt	Time Sorted	0.77 (0.11)	0.15 (0.19)	0.69 (0.34)	0.64 (0.35)	0.73	20.05 (14.78)
	Passage	0.76 (0.12)	0.14 (0.19)	0.67 (0.35)	0.63 (0.36)	0.62	21.58 (15.31)
	Document	0.76 (0.12)	0.16 (0.20)	0.67 (0.34)	0.63 (0.35)	0.60	21.40 (15.18)
Expert	Time Sorted	0.84 (0.11)	0.35 (0.28)	0.76 (0.27)	0.70 (0.30)	0.84	7.22 (7.13)
	Passage	0.84 (0.10)	0.33 (0.27)	0.76 (0.28)	0.71 (0.32)	0.77	7.59 (7.46)
	Document	0.83 (0.11)	0.31 (0.27)	0.72 (0.29)	0.67 (0.32)	0.80	10.03 (11.83)

System Prompt	BERT Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
TAME	0.81	0.30	0.62	0.55	0.81	5.74
Without System	0.82	0.31	0.71	0.65	0.75	11.40
Expert	0.82	0.29	0.66	0.58	0.85	6.41



System Prompt	BERT Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
TAME	0.84	0.37	0.75	0.70	0.78	7.61
Without System	0.77	0.15	0.68	0.63	0.65	21.01
Expert	0.84	0.33	0.75	0.69	0.80	8.28

延伸實驗

文本生成 (加上引導詞)

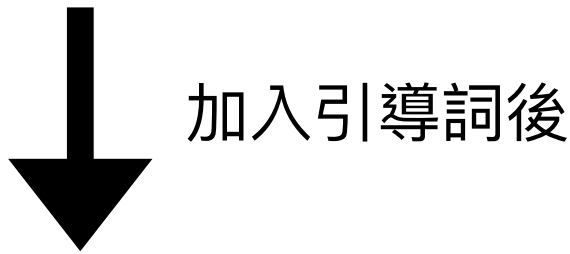
- 將對應的系統提示詞尾部加入「與「{使用者查詢}」高度相關的解決方案如下:」作為引導詞，進行對照實驗
- 具有系統提示的情況下，加入引導詞後，TAME在BERT Score、BLEU Score、ROUGE-1及ROUGE-L等評估指標上均顯著提升。
- 無系統提示 (Without System Prompt)的情況下，加入引導詞反而導致性能全面下降。這可能是干擾TAME-8B從檢索內容中模仿其回應內容的能力，從而導致重點擷取能力下降。
- 這現象突顯了系統提示與引導詞之間的效應:

明確的任務指示配合相應的引導詞，能夠有效提升模型性能。

與Kojima等人(2022)的研究結果相互呼應，僅透過在提示詞尾部加入「Let's think step by step.」的簡單引導，即能顯著提升LLMs的推理能力。

System Prompt	Retrieval Type	Bert Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
TAME	Time Sorted	0.84 (0.12)	0.38 (0.31)	0.75 (0.29)	0.70 (0.33)	0.78	6.55 (6.56)
	Passage	0.84 (0.11)	0.36 (0.29)	0.76 (0.29)	0.70 (0.33)	0.78	7.19 (7.74)
	Document	0.84 (0.13)	0.36 (0.29)	0.74 (0.30)	0.69 (0.33)	0.77	9.08 (11.01)
Without System Prompt	Time Sorted	0.77 (0.11)	0.15 (0.19)	0.69 (0.34)	0.64 (0.35)	0.73	20.05 (14.78)
	Passage	0.76 (0.12)	0.14 (0.19)	0.67 (0.35)	0.63 (0.36)	0.62	21.58 (15.31)
	Document	0.76 (0.12)	0.16 (0.20)	0.67 (0.34)	0.63 (0.35)	0.60	21.40 (15.18)
Expert	Time Sorted	0.84 (0.11)	0.35 (0.28)	0.76 (0.27)	0.70 (0.30)	0.84	7.22 (7.13)
	Passage	0.84 (0.10)	0.33 (0.27)	0.76 (0.28)	0.71 (0.32)	0.77	7.59 (7.46)
	Document	0.83 (0.11)	0.31 (0.27)	0.72 (0.29)	0.67 (0.32)	0.80	10.03 (11.83)

System Prompt	BERT Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
TAME	0.81	0.30	0.62	0.55	0.81	5.74
Without System	0.82	0.31	0.71	0.65	0.75	11.40
Expert	0.82	0.29	0.66	0.58	0.85	6.41



加入引導詞後

System Prompt	BERT Score	BLEU Score	ROUGE-1 (Recall)	ROUGE-L (Recall)	LLM Score	Inference Time
TAME	0.84	0.37	0.75	0.70	0.78	7.61
Without System	0.77	0.15	0.68	0.63	0.65	21.01
Expert	0.84	0.33	0.75	0.69	0.80	8.28

比較加入引導詞之差異

Expert + Passage + Base 方法的比較

- 加入引導詞確實能產生更符合參考文本的生成結果
- LLM Score與人工標註的結果 (Pair-matching)不一致
- 混淆矩陣分析
 - 未加入引導詞，模型傾向預測為相似，導致False Positive過高
 - 加入引導詞，則False Positive & False Negative的比例較為平衡
- LLM Score在評估文本相似度時存在預測傾向不一致問題，從而導致評估結果不僅與人工判斷有所出入，也與其他評估指標的趨勢不一致

Metrics	有加入引導詞	未加入引導詞								
LLM Score	0.77	0.87								
Pair-matching	0.71	0.66								
Confusion Matrix	<table><tr><td>59</td><td>13</td></tr><tr><td>7</td><td>14</td></tr></table>	59	13	7	14	<table><tr><td>57</td><td>24</td></tr><tr><td>4</td><td>8</td></tr></table>	57	24	4	8
59	13									
7	14									
57	24									
4	8									
True Positive (TP)	59	57								
True Negative (TN)	14	8								
False Positive (FP)	13	24								
False Negative (FN)	7	4								
Precision	0.82	0.93								
Recall	0.89	0.70								
F1-Score	0.86	0.80								
Accuracy	0.78	0.70								

比較加入引導詞之差異

Expert + Passage + Base 方法的比較

- 加入引導詞確實能產生更符合參考文本的生成結果
- LLM Score與人工標註的結果 (Pair-matching)不一致
- 混淆矩陣分析
 - 未加入引導詞，模型傾向預測為相似，導致False Positive過高
 - 加入引導詞，則False Positive & False Negative的比例較為平衡
- LLM Score在評估文本相似度時存在預測傾向不一致問題，從而導致評估結果不僅與人工判斷有所出入，也與其他評估指標的趨勢不一致

Metrics	有加入引導詞	未加入引導詞								
LLM Score	0.77	0.87								
Pair-matching	0.71	0.66								
Confusion Matrix	<table><tr><td>59</td><td>13</td></tr><tr><td>7</td><td>14</td></tr></table>	59	13	7	14	<table><tr><td>57</td><td>24</td></tr><tr><td>4</td><td>8</td></tr></table>	57	24	4	8
59	13									
7	14									
57	24									
4	8									
True Positive (TP)	59	57								
True Negative (TN)	14	8								
False Positive (FP)	13	24								
False Negative (FN)	7	4								
Precision	0.82	0.93								
Recall	0.89	0.70								
F1-Score	0.86	0.80								
Accuracy	0.78	0.70								

結論與未來工作

實驗總結

文本檢索

- 特徵整合

- 文本特徵整合後再進行案件檢索，可以提升檢索準確度

- 二階段檢索架構

- 保留重點資訊，同時進一步過濾約30%的冗餘資訊

- 文本總結輔助檢索應用

- 將檢索案件交由TAME-8B整理案件重點後，再進行篩選，可以過濾約40%的冗餘資訊並提升F1-Score約3.97%
- 檢索效果受TAME-8B的回應內容影響，且在實驗中檢索模型並未帶來一致性的檢索性能提升，其檢索的穩定性與模型適用範圍仍需進一步驗證

實驗總結

文本生成

- **系統提示**

缺乏明確任務指示時，TAME-8B傾向尋找檢索內容中的回覆模式模仿，若無可模仿模式則性能明顯下降，生成無意義序列的機率增加。

- **檢索內容**

減少冗餘資訊可以顯著提高文本生成的性能，且重要內容優先排序比時序資訊更有助於模式匹配。

- **生成策略**

- Reasoning: 可能因破壞文本中的Q&A對結構而表現不佳
- Sampling: 可能因檢索模型的偏好選擇未能在評估指標上顯現，需要進一步比較候選文本與檢索器選擇的內容，從而驗證此方法的可行性

- **延伸實驗**

在明確任務指示下，加入相互呼應的引導詞能顯著提升性能；反之則會干擾模型的模式匹配能力。

未來工作

研究限制與待驗證之處

- **測試資料規模**

測試資料集規模有限，可能影響結果的泛化性

- **Sampling方法的有效性**

初步分析可能源於檢索模型的偏好選擇。需要進一步比較候選文本與檢索結果，以驗證此方法的可行性

- **LLM Score的預測傾向與準確性問題**

基於TAME-8B的LLM Score在文本相似度評估上存在預測傾向不一致問題，需要設計更robust的提示模板以提升評估可靠度

未來工作

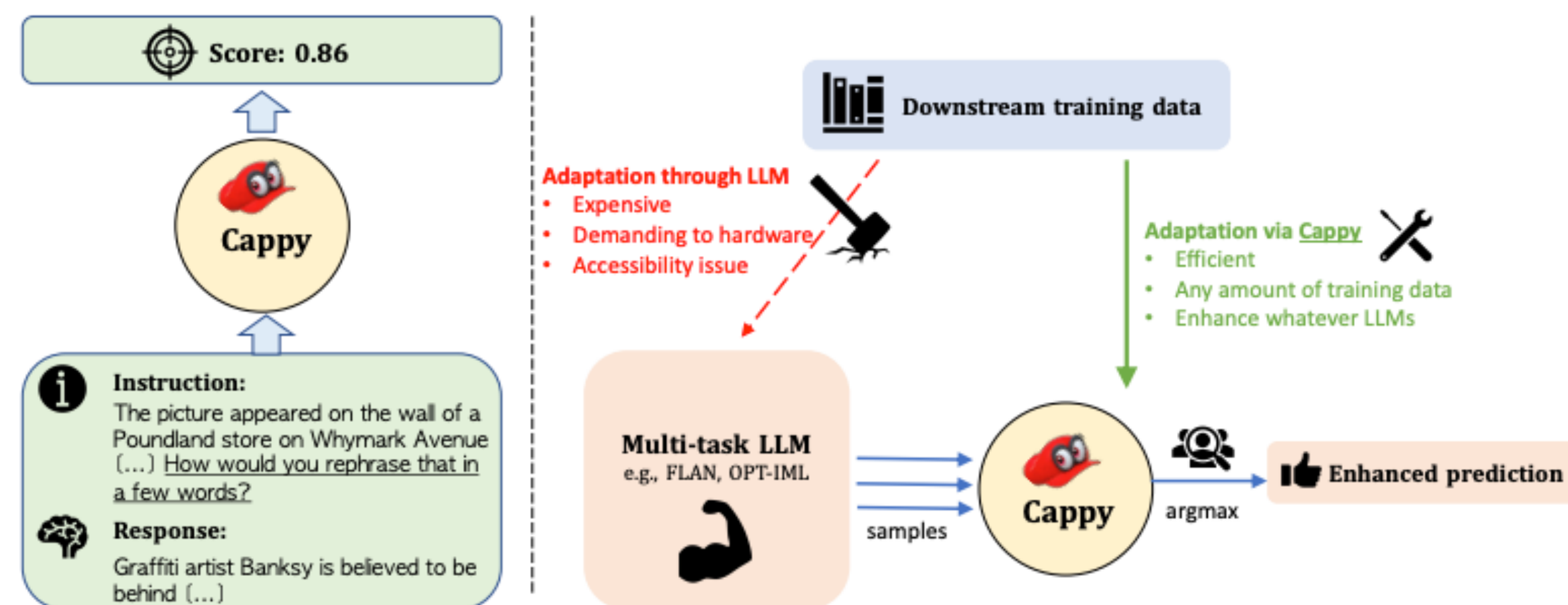
研究方向

- 搜集使用者反饋評分，建制評分模型
- 使用者輸入之隱私資訊檢測機制
 - 基於NER的隱私辨識
 - 基於規則式的隱私資訊處理機制
- 檢索增強生成的採樣策略
- 檢索內容優化

未來工作

搜集使用者反饋評分，建制評分模型

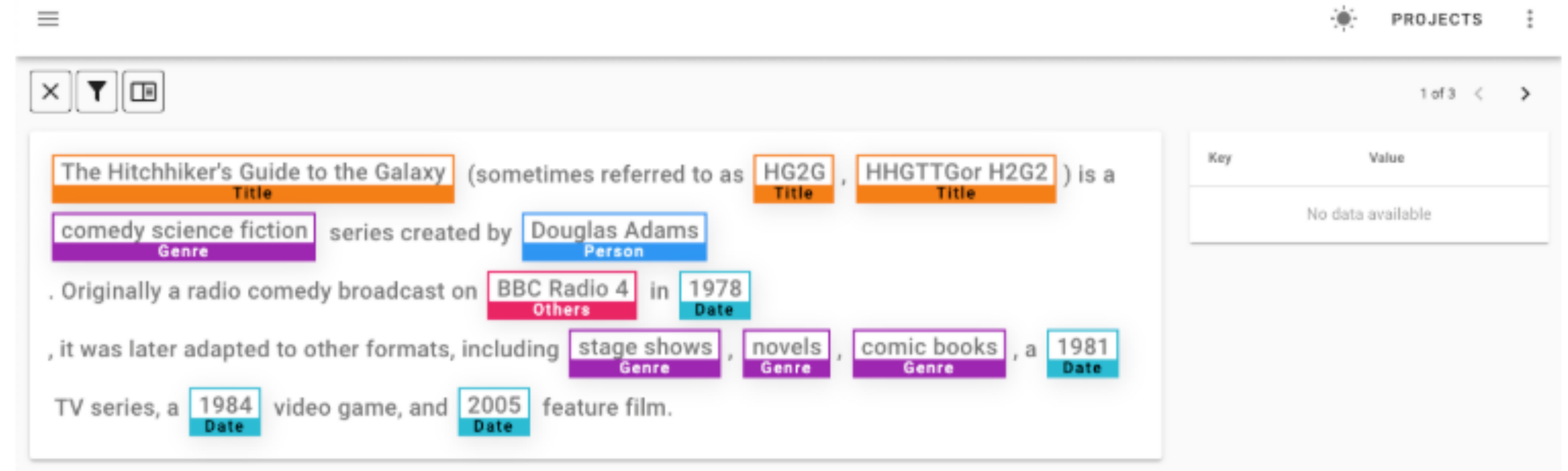
- 訓練資料過濾
- 基於人類反饋的強化學習
- 輸出品質優化
- 自動化評估機制



未來工作

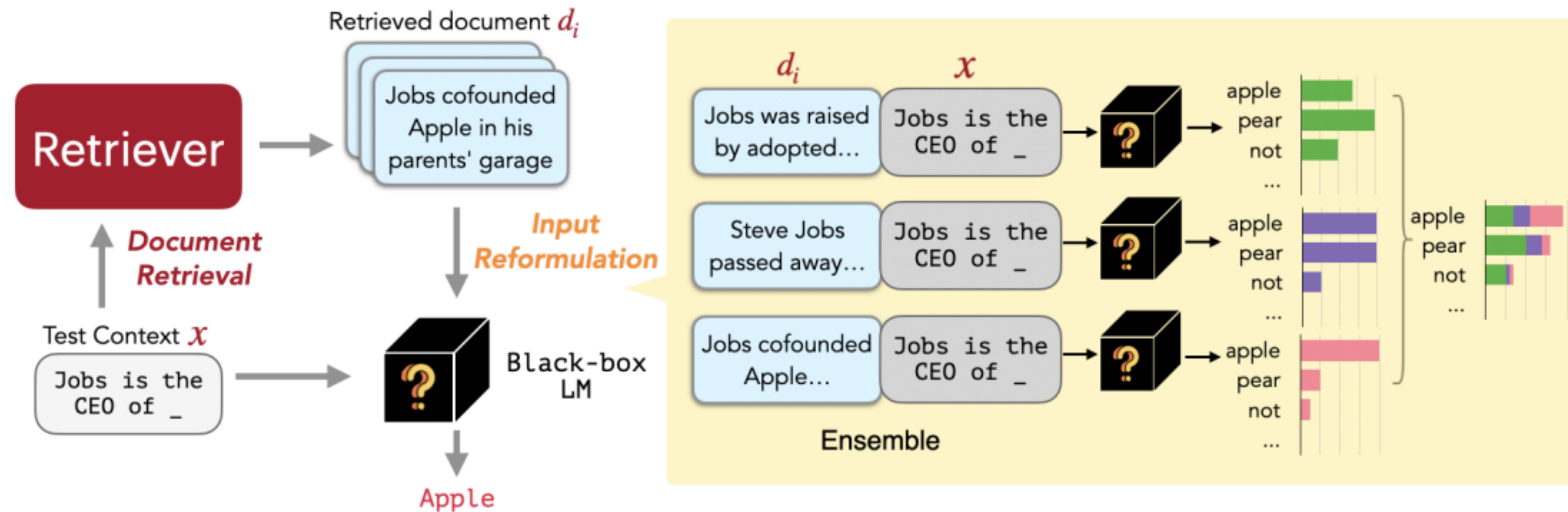
使用者輸入之隱私資訊檢測機制

- 基於NER的隱私辨識
- 基於規則式的隱私資訊處理機制



未來工作

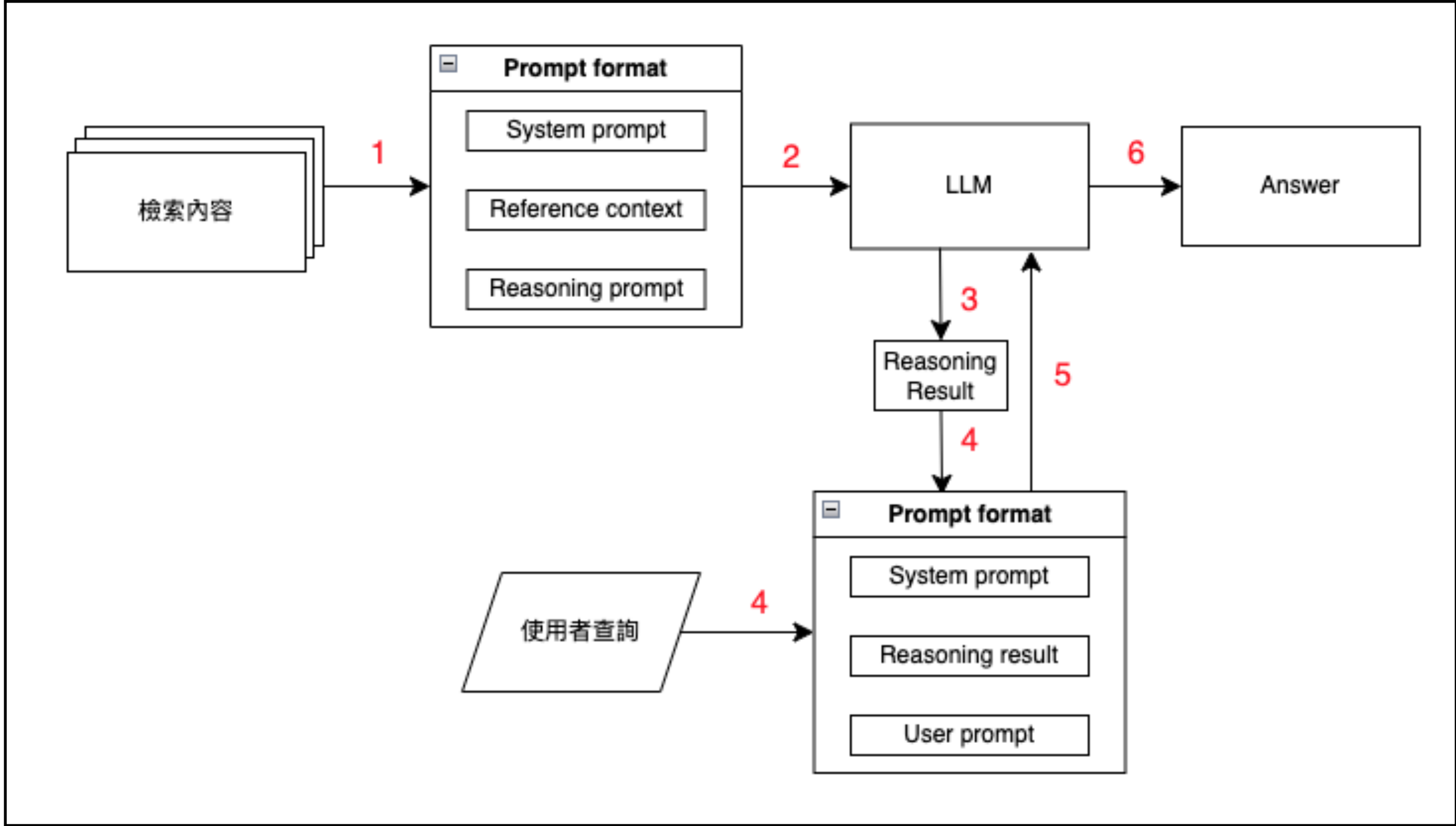
檢索增強生成的採樣策略



- 利用檢索模型獲取多個相關文檔，將各檢索文檔與輸入內容組合生成多組Token採樣分布
- 將檢索內容與輸入內容間的餘弦相似度作為權重，整合採樣分布後進行最終採樣
- 在MetaEdge的應用場景中值得一試，但需要在**性能提升與推理效率**同時驗證

未來工作

檢索內容優化



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)---

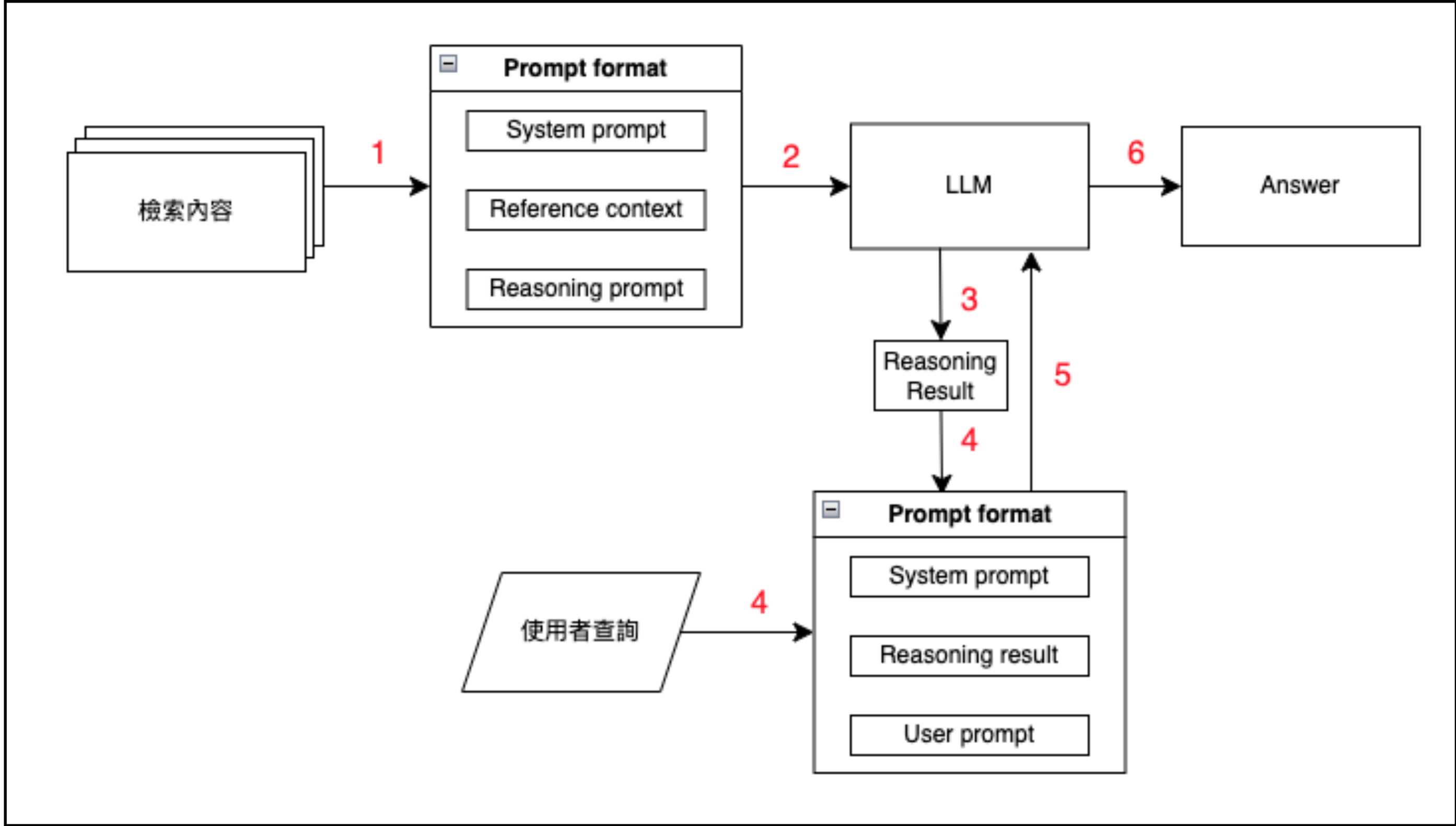
回覆內容(8):

【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

未來工作

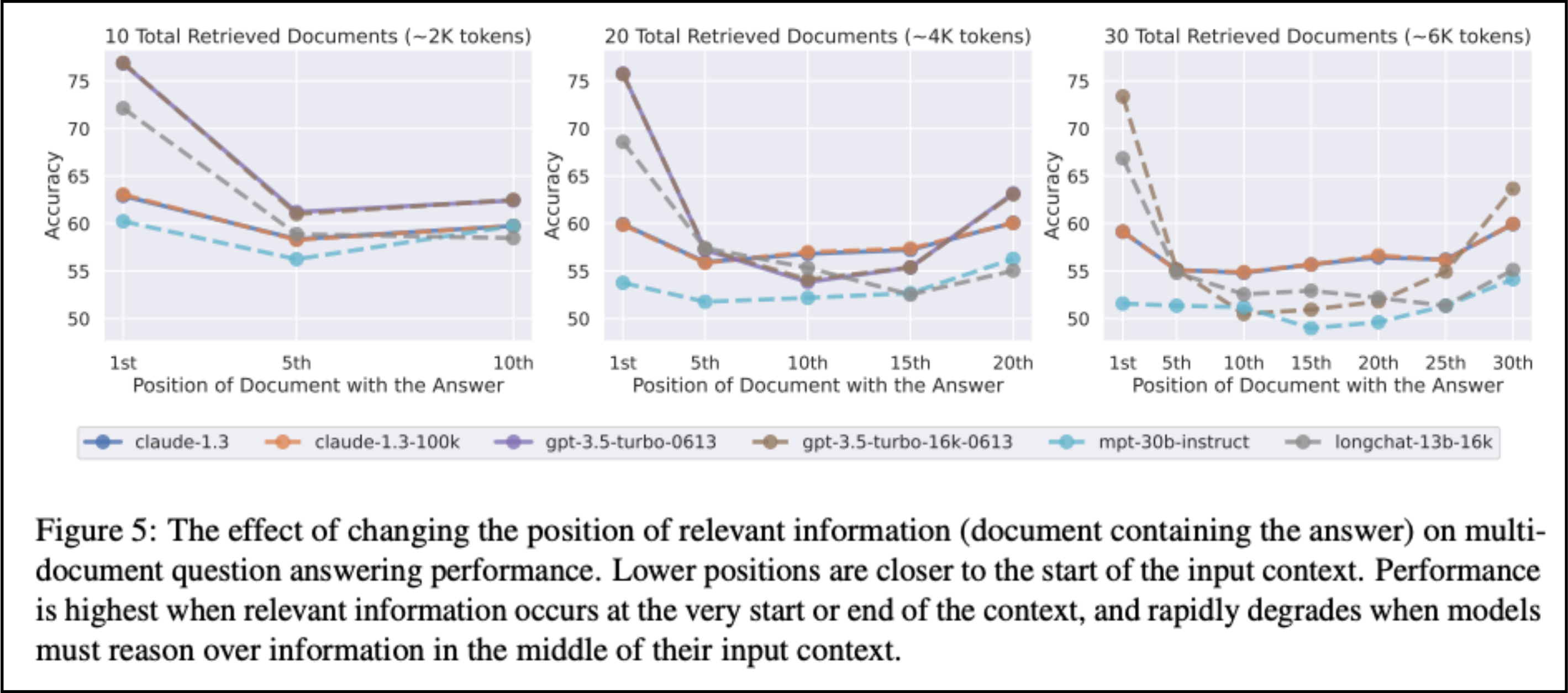
檢索內容優化



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

未來工作

檢索內容優化

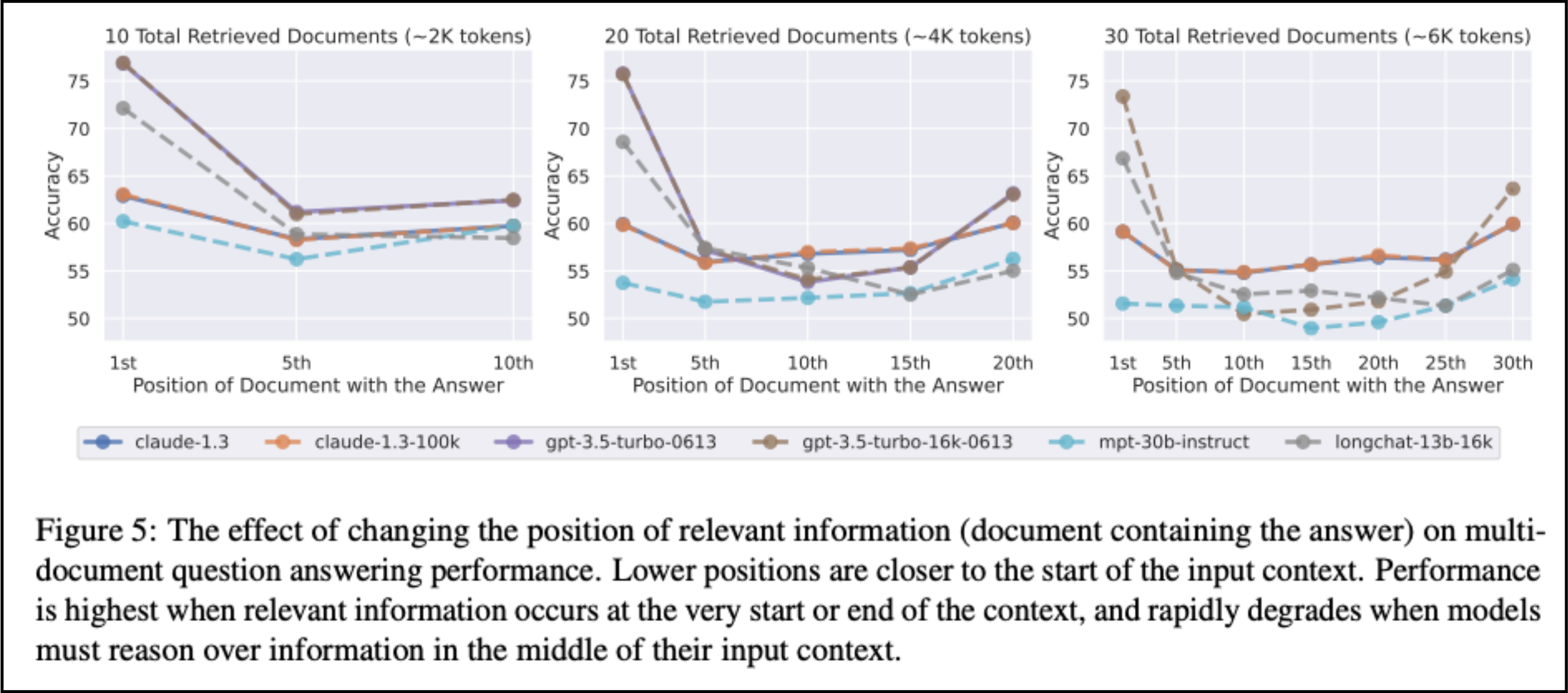


Document 1
Document 2
Document 3
.
.
.
Document N

Top N 搜尋結果的擺放位置優化

未來工作

檢索內容優化



Document 1
Document 3
.
.
.
Document 5
.
.
.
Document 4
Document 2

Top N 搜尋結果的擺放位置優化

Q&A