研究進度展示

探索 RAG 中關鍵字匹配與向量語意搜尋結合的最佳權重配置

目次

背景

最終目標：
為業界提供一個量化的 alpha 設定方法，無論是在資料上線前的領域分類，還是使用者問題進來時的動態預處理，都能透過本研究結果給出最佳的 hybrid search 配比。

- 許多現有方法直接使用固定的alpha比率（如6:4、5:5），未能基於資料屬性和使用者問題進行調整。

參考文獻

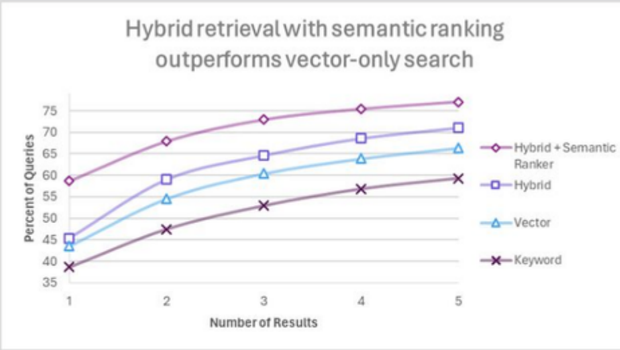
IBM :
Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers

Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with...

Retrieval-Augmented Generation (RAG) is a prevalent approach to infuse a private knowledge...

ar5iv

Microsoft :
Azure AI Search: Outperforming vector search with hybrid retrieval and ranking capabilities



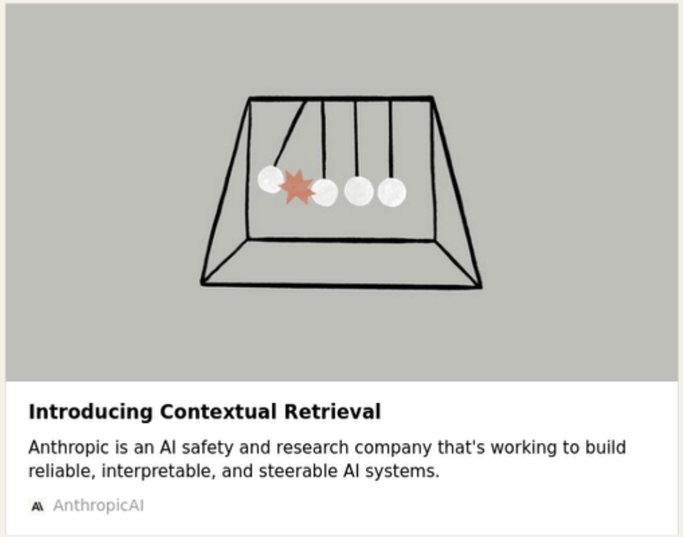
Azure AI Search: Outperforming vector search with hybrid retrieval and ranking capabilities

How do you find the best content to feed your generative AI models? In this blog post, we show you how to use Azure AI Search's hybrid retrieval and..

TECHCOMMUNITY.MICROSOFT.COM

參考文獻

Anthropic :
Introducing Contextual Retrieval



Advancing the Evaluation of Traditional Chinese Language Models: Towards a Comprehensive Benchmark Suite

研究方法-技術使用

向量搜尋

- Embedding 模型選擇：text-embedding-3-large

關鍵字搜尋

- 繁體中文斷詞 (For BM25F)：CKIP Transformers
  - Weaviate, LlamaIndex 等關鍵字搜索工具都未提供中文 tokenizer
  - 也未查到繁體中文圈有相關的文章，所以需自己探索處理

https://www.53ai.com/news/RAG/2024080632561.html

研究方法

分為兩層：

1. 第一層：資料集的全面測試
    - 全面的測試，測試範圍為 10:0 到 0:10，並統計各比例下的 top-1 準確率
    - 結果：已經完成了這部分的測試，並針對不同問題類型（多關鍵字、少關鍵字、長語句、短語句）統整了準確率和圖表

研究方法

分為兩層：

1. 第二層：兩個應用分支
    - (1) 問題種類預處理：當使用者的問題進來時，可以透過預處理來判斷問題的類型，並依據我們的實驗結果來給出最佳的 alpha 值。例如，長語句可能適合較高比例的向量搜尋，而多關鍵字問題則適合較高比例的關鍵字搜尋。

研究方法

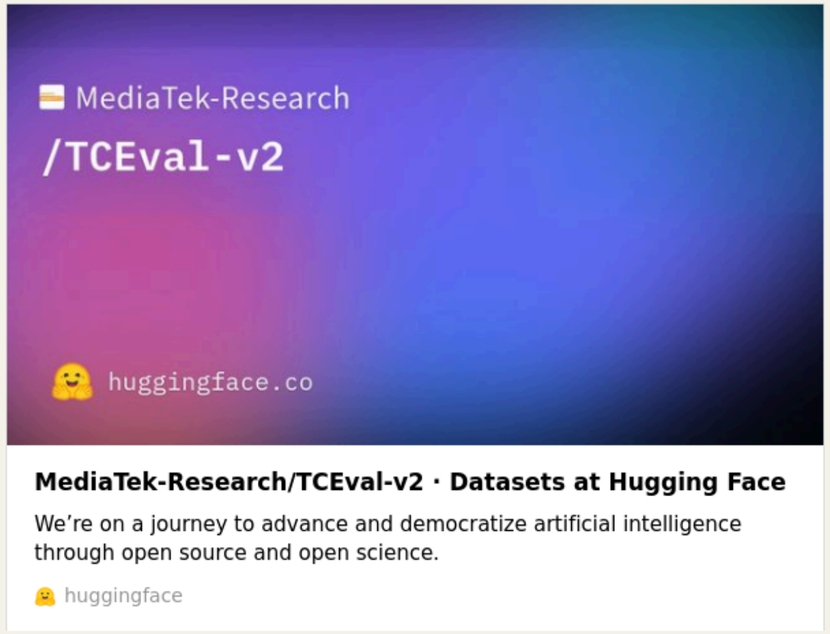
分為兩層：

- 第二層：兩個應用分支 (未完成)
  - (2) 資料集領域分類：針對資料集的不同領域（如音樂、程式碼、數學、歷史等），我們也可以在 RAG 上線前對其進行分類，並根據這些領域特徵來提供最適合的 alpha 設定。這樣的應用讓我們在不同知識屬性下，能夠根據我的實驗結果來量化 alpha 的權重

https://ihower.tw/blog/archives/12167

資料說明

- MediaTek-Research/TCEval-v2
- Subset: drcd (台達閱讀理解資料集)



- 其中有不重複文章段落共 1000 段以及對應的 3493 個問題

Dataset Viewer — Auto-converted to Parquet

Subset (76): drcd · 3.5k rows | Split (2): test · 3.49k rows

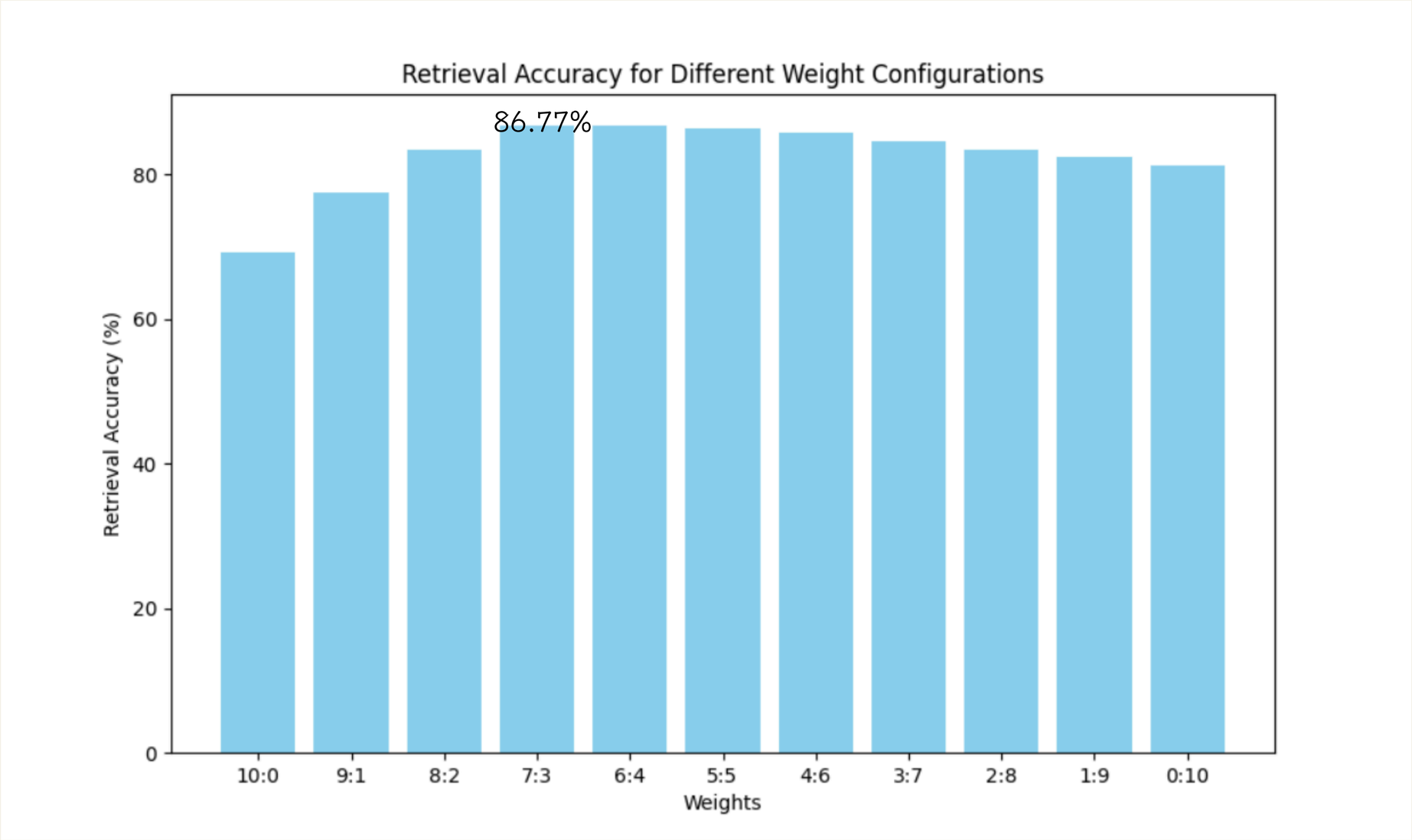
Search this dataset

id string · lengths 6 9	paragraph string · lengths 231 992	question string · lengths 6 59
test-0	要探討從梨俱吠陀到波你尼時代梵語的發展，可以考察印度教其它文本，...	夜柔吠陀與阿闥婆吠陀均可以最為研究哪一門語言的參考？
test-1	要探討從梨俱吠陀到波你尼時代梵語的發展，可以考察印度教其它文本，...	哪一本書規範了梵語的正確語法？
test-2	要探討從梨俱吠陀到波你尼時代梵語的發展，可以考察印度教其它文本，...	中古印度-雅利安語方言的前身與哪一門語言都同時在古印度使用？
test-3	波你尼所定義的梵語是從更早的「吠陀」形式演化出來的。學者經常把吠...	波你尼梵語與哪一門語言非常相似？
test-4	波你尼所定義的梵語是從更早的「吠陀」形式演化出來的。學者經常把吠...	印度教的最早宗教文本以什麼語言撰寫？
test-5	波你尼所定義的梵語是從更早的「吠陀」形式演化出來的。學者經常把吠...	吠陀梵語大約到何時開始演變成古典梵語？

Previous 1 2 3 ... 35 Next

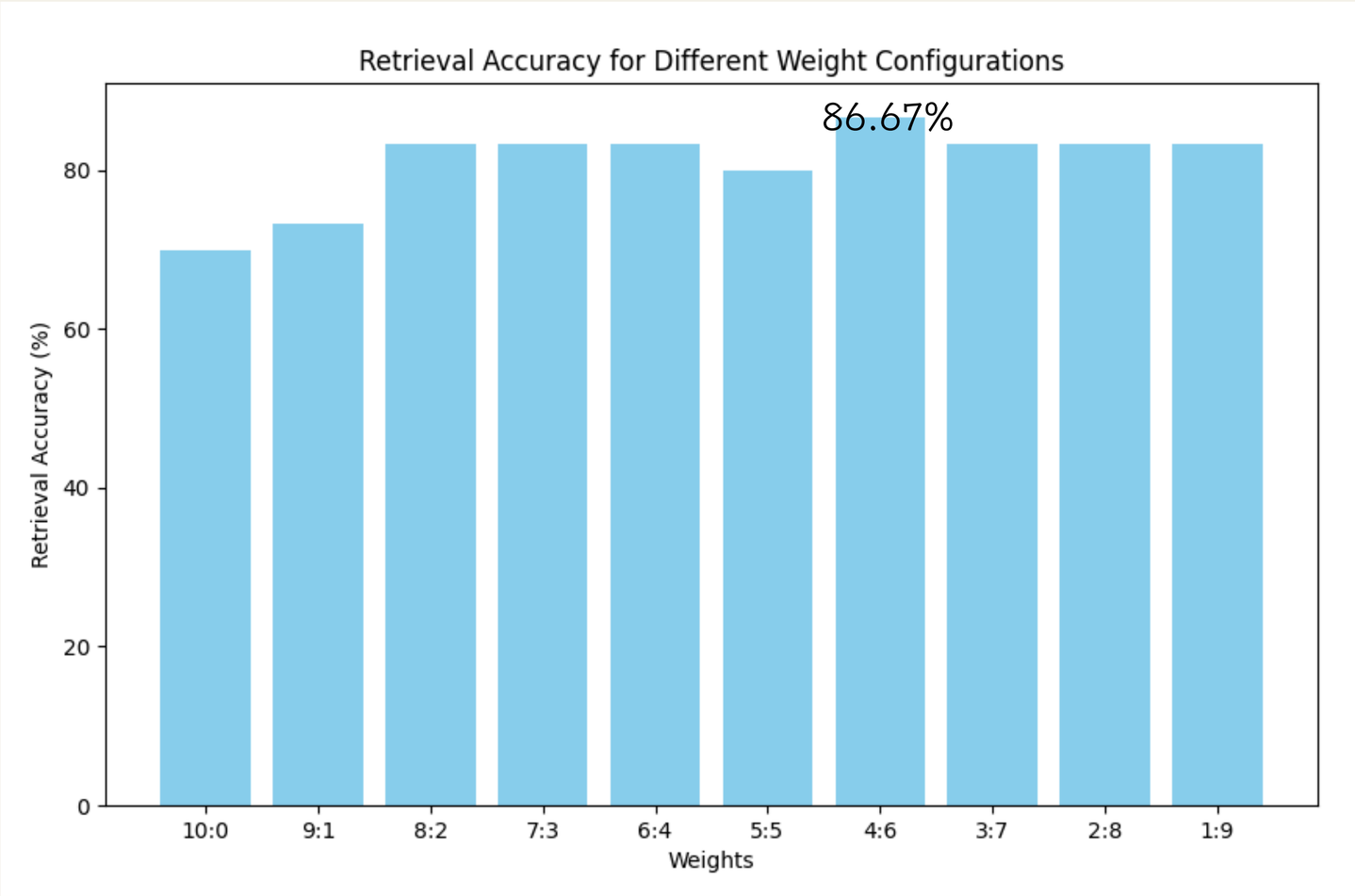
https://github.com/JustinHsu1019/RAG_Search_Weights

研究結果展示 (3493 題全面測試)



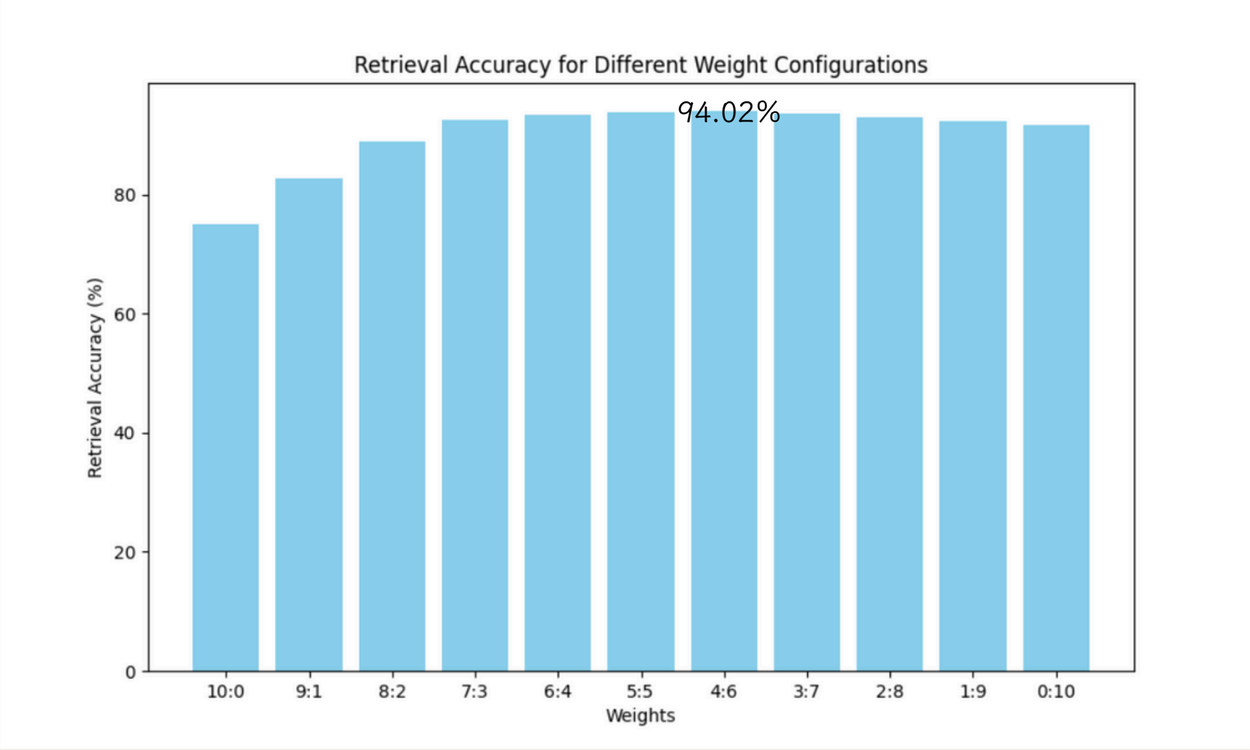
參考比較 (期末專題報告成果)

30 題教育領域「課程」測試資料

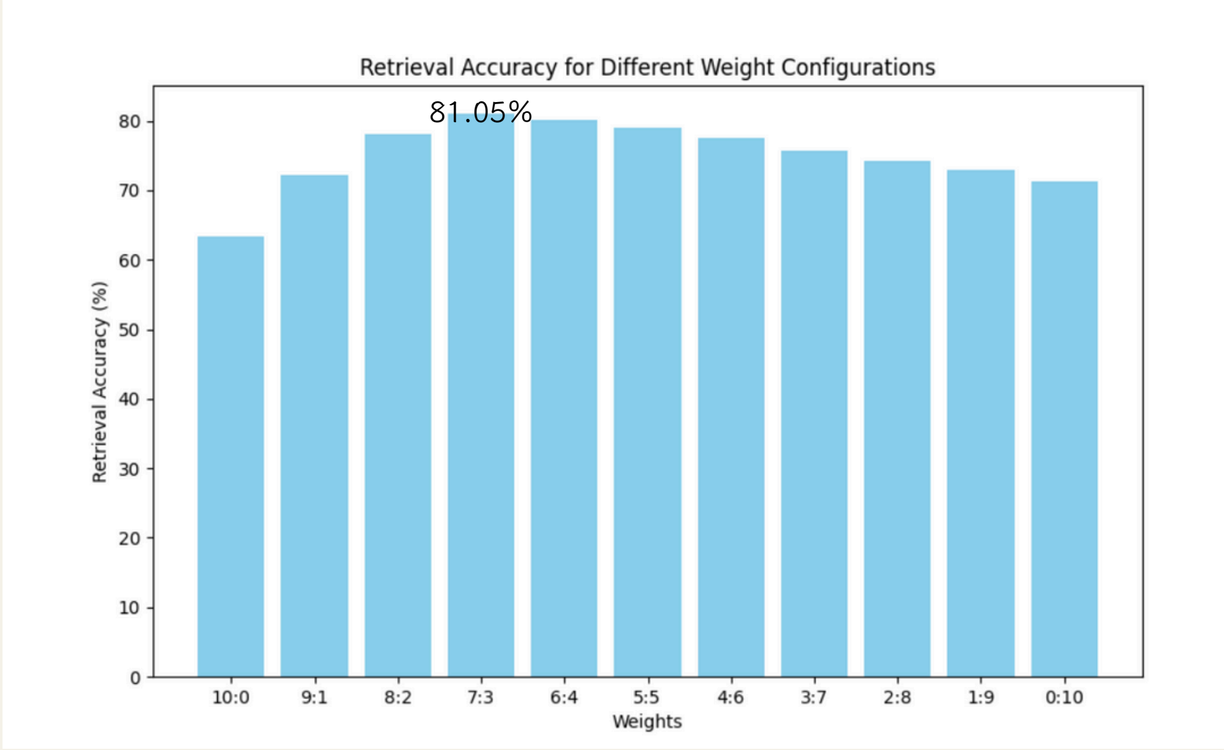


研究結果展示 (題目關鍵字數量)

關鍵字數量大於平均

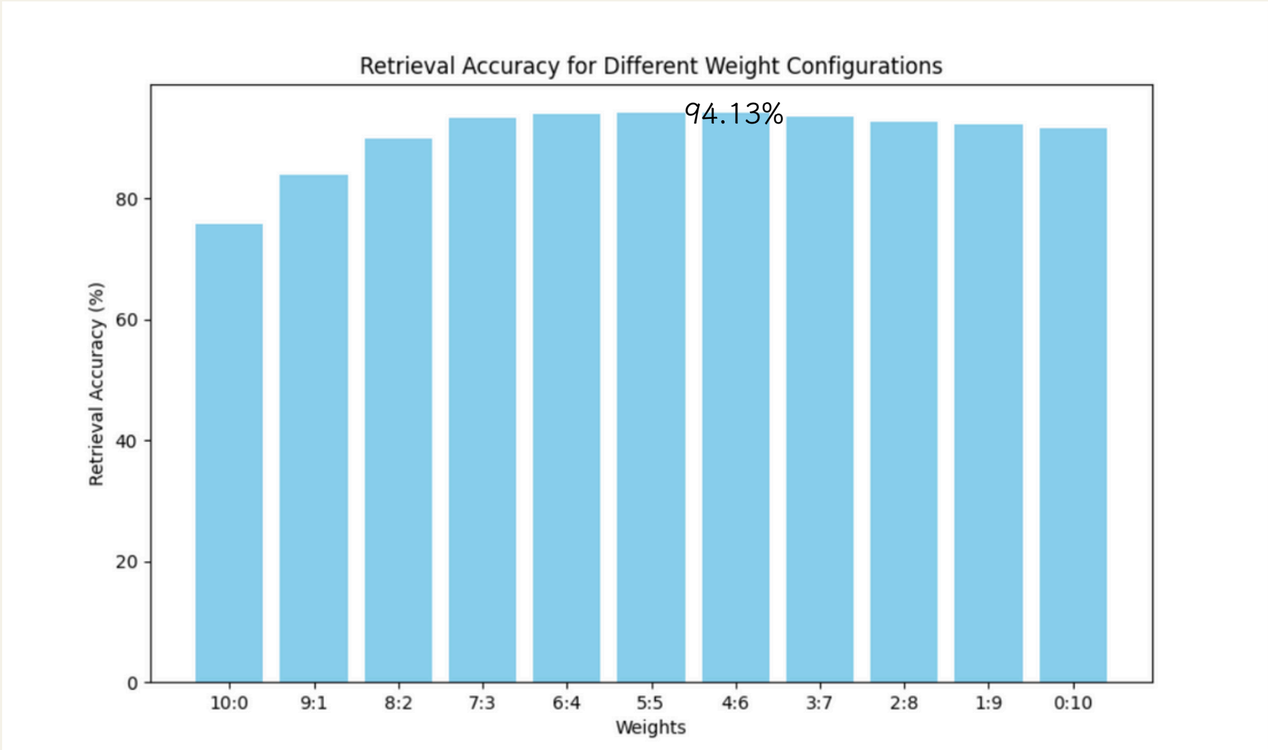


關鍵字數量少於平均

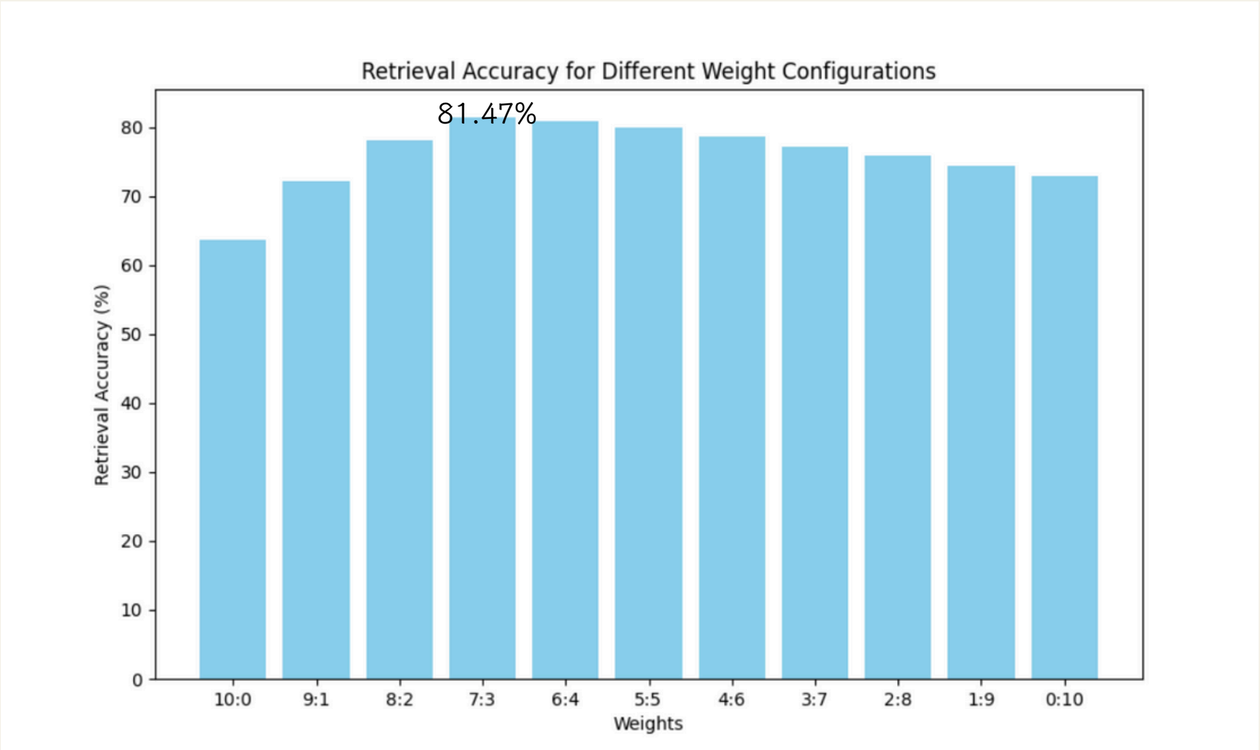


研究結果展示 (題目語句長短)

語句長度大於平均



語句長度少於平均



結論

3493 題全面測試

- 結果顯示過去期末專題成果分佈大致正確，關鍵字與向量搜尋各佔半的左右多是準確率最高的分佈區
- BM25 關鍵字匹配度較高是由於資料集內無同義詞或錯字問題，且題目與文本有較多直接可對應的專有名詞
- 1000 篇文章似乎太過獨立，互相之間很少重疊，故專有名詞較輕易匹配至正確文本

結論

使用者問題種類的差異

- 原先預想
  - 分出四類不同的問題種類：高關鍵字準確率 / 低關鍵字準確率 / 高向量準確率 / 低向量準確率
- 關鍵字多寡與語句長短似乎成相同分佈，關鍵字少 / 語句短只意味著攜帶較少訊息，又因測試資料相互之間獨立的特性，故無論向量搜尋或關鍵字匹配都較不易比多關鍵字準
  - 似乎對於此資料集，關鍵字起了過大的作用
  - 且此類資料集似乎不符合業界常見的資料儲存形式

下一步

- 為文章做分類 / 尋找其他測試資料集
- 重新考慮問題分類基準，想辦法分出四個成果
- 工具開發：開發開源工具，讓開發者可在上線前將資料集丟進來，透過LLM自動生成各類型的問題（例如長語句、短語句、多關鍵字、少關鍵字），並進行測試，以幫助他們找到最適合該資料集的 alpha 值

問題

- 不知道怎麼為資料集用「領域」做分類，要取出什麼樣的「分類基準」才會有價值？
- 要重新尋找資料集嗎？“問題“要自己造嗎？還是用專業的資料集成果會比較受認可？
- 研究大方向上是否可以
- 寫論文的注意事項

問題-分類方式

分類定義：

- 由於現有的 1000 篇文章過於獨立且還未想出好的方法分出明確的領域，我目前正在考慮要不要捨棄這些文章，改用其他的專業資料集；如果不捨棄，則需要為這些文章進行分類，但在尚未清楚文章內容前，無法隨意定義分類標準。
  - 重選資料集：分別提取包含法律、產品手冊、論文、程式碼等不同領域的知識。再自定義 “問題“
  - 疑慮：自定義問題產出的結果可能不被認可