基於檢索增強生成的精準資訊擷取系統

Answer Briefly and Precisely: A Segment-based RAG Framework with Specialized Preprocessing Pipeline

指導教授：蔣榮先教授

專題成員：黃亮晨

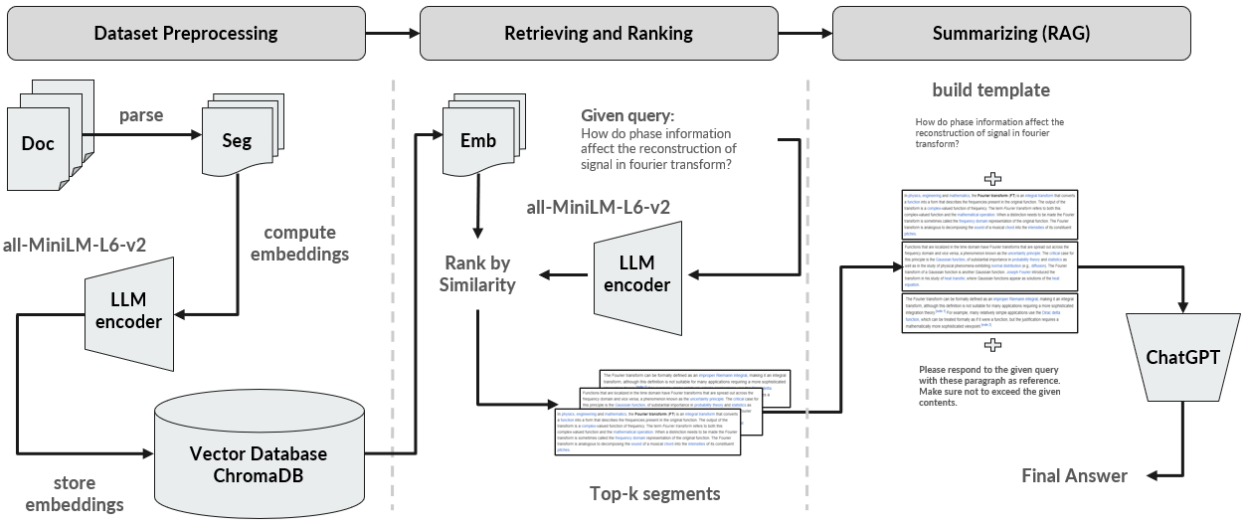
開發工具：Python3

測試環境：Windows 10

一、 簡介：

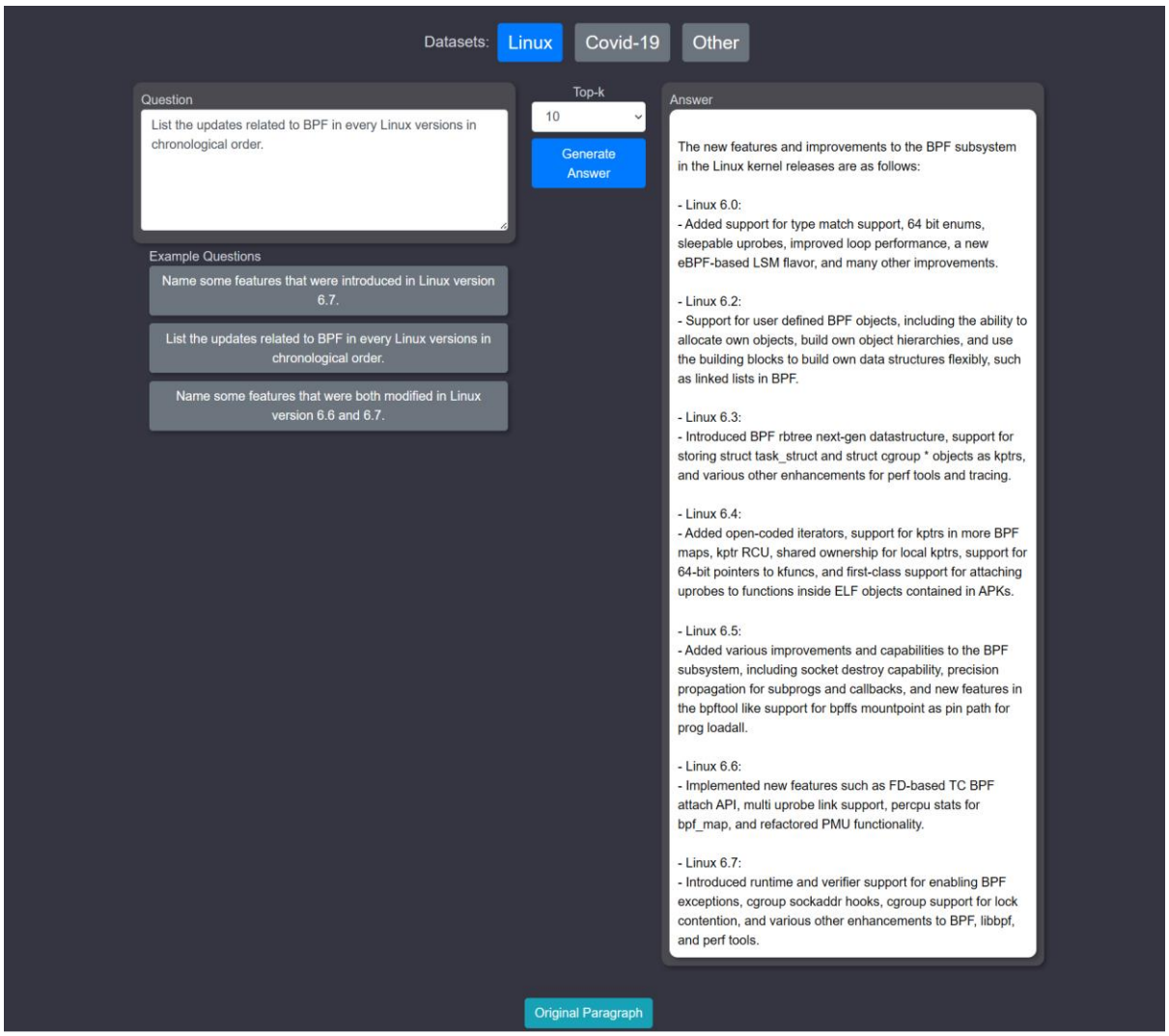
本研究改良傳統搜尋引擎資訊檢索方法（Information Retrieval）及大型語言模型（Large Language Model, LLM），設計出一個系統讓使用者只需輸入問題，不必親自檢視大量資料就能直接取得想要的回答，並利用檢索增強生成(Retrieval Augmented Generation，RAG)框架避免 LLM 的幻覺問題（hallucination），使其回答具有較高的可信度。具體來說，本研究的成果如下：

1. 透過新的前處理框架，改善過長參考文本所衍伸出的問題
2. 經由使用 RAG 框架，避免出現 LLM 的幻覺問題
3. 設計前後端介面，降低使用者的操作門檻



第一階段為資料收集與前處理，透過將長文件切分成簡短段落後，交由 LLM （all-MiniLM-L6-v2）算出每一個段落的特徵並儲存於資料庫中；第二階段為檢索與排序，輸入的問題會經過相同的 LLM（all-MiniLM-L6-v2） 得出一個特徵，再透過比較與排序找出資料庫中與問題最接近的幾組特徵與其對應的文本；第三階段為總結，將問題與檢索出的段落放進設計過的模板中，交由 ChatGPT 進行最終問答。

二、 測試結果：



以上是使用者介面。首先，最上方 Datasets 可以選擇你想要使用的資料集。左邊 Question 欄位使用者可以自行輸入問題，或是直接選擇點選 Example Questions 中的預設問題。接下來，選擇 Top-k 的數值後按下 Generate Answer，右邊就會顯示生成的答案。

最下面的 Original Paragraph 按鈕則是讓使用者在對答案有疑慮或是想看原文的時候，可以點選用來生成 Answer 的 Top-k 個原本的文章。

在這個測試中，我們的問題是請他照時間順序舉出與 BPF 相關的更新條例，可以看到答案有確實按照版本序排序列舉每個版本的更新。