一、摘要

大型語言模型(Large Language Models)在各種領域中的許多應用上嶄露頭角，逐漸成為普羅大眾日常生活的一部分。然而，LLM 的幻覺問題(Hallucination)卻會造成真假難辨的生成結果，造成使用者對 LLM 產生疑慮，進而影響 LLM 各種應用的持續推廣。傳統的資訊檢索（Information Retrieval）雖然在基於關鍵字的搜尋中表現出色，能夠有效地提供精確的搜尋結果，但其主要依賴用戶自行從大量搜尋結果中篩選和提取相關資訊。不只是如此，使用者仍需手動從搜尋出的文章中定位所需的段落或片段，並進一步對查詢結果進行總結與解釋。本研究旨在透過改良版的檢索增強生成(Retrieval Augmented Generation)，使大型語言模型的回答內容有理有據，提高使用者對大型語言模型回覆內容的信心程度。

二、研究動機與研究問題

(一)、研究動機

自 ChatGPT 問世以來，大型語言模型(Large Language Models，以下簡稱 LLM)迅速地崛起並走進公眾與企業的視野，在聊天對話、問題回應、翻譯等各式自然語言任務上達到傑出的表現，部分經過特別訓練的 LLM 甚至在專家考試、多模態理解與整合等複雜任務中表現出驚人的準確性，在可以預見的未來將對各行各業產生深遠的影響。

然而，在這些令人驚艷的表現背後，有些問題逐漸浮上檯面，其中幻覺問題(Hallucination)是一大隱憂。當 LLM 在生成回覆時，由於並不涉及搜尋與判斷資訊正確性的階段，因此有可能會回答看似語意通順、邏輯清晰，實則子虛烏有、錯誤百出的回應，當使用者無法辨別真假時，可能會造成嚴重的後果。

為了改善幻覺問題，檢索增強生成(Retrieval Augmented Generation，以下簡稱 RAG)被提出。透過給予 LLM 含有正確答案的文本，藉此提高回覆內容的正確性，有效降低了產生幻覺的可能性。然而，獲取文本的過程、添加進 LLM 的文本數量與長度等，也將影響 RAG 的運作效能。若獲取的文本帶有錯誤的資訊，或是因為過長的參考資料造成過多冗餘資訊，都將對 LLM 回答的品質造成負面影響。本研究旨在透過改良RAG框架的運作流程來解決上述問題。

(二)、研究問題

透過理解大型語言模型的運作原理、幻覺問題可能的成因與RAG的優勢後，我們決定藉由改良 RAG的運作流程來進一步提升這個框架在特定問題上的表現。具體而言，本研究的問題如下：

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. 結合預訓練模型與既有的工具，以最小成本建置一個可供驗證的範本
3. 設計前後端介面，降低使用者的操作門檻

五、結果與討論

(一)、實驗結果

以下是使用 all-MiniLM-L6-v2 作為 Embedding model，在不同問題下，更改不同 Top-K 得出的實驗結果。

	Top-K	Precision	AP	NDCG
Covid-19 Wiki Q1	5	1	1	1
Covid-19 Wiki Q1	10	0.9	0.96	0.93
Covid-19 Wiki Q1	20	0.65	0.93	0.75
Covid-19 Wiki Q2	5	0.6	0.92	0.7
Covid-19 Wiki Q2	10	0.3	0.92	0.45
Covid-19 Wiki Q2	20	0.15	0.92	0.3
Covid-19 Wiki Q3	5	0.8	1	0.87
Covid-19 Wiki Q3	10	0.5	0.91	0.63
Covid-19 Wiki Q3	20	0.3	0.81	0.44

Q1 : What measures can people take to prevent COVID-19 while they are outside?
Q2 : What is the name of the virus that causes COVID-19?
Q3 : How can I tell if I have COVID-19?

從實驗數據中可以看出，當 Top-K 設為 5 或 10 時，模型在 Precision 和 NDCG 指標上均表現出較高的分數，顯示出良好的效能。然而，當 K 值增加至 20 時，雖然 AP 指標保持穩定，但 Precision 顯著下降，且 NDCG 顯示回應的實用性有所減少。因此，選擇 Top-5 或 Top-10 的回應已能滿足需求，增加回應數量並未顯著提升回答品質，反而可能會對系統效率產生負面影響。

以下是使用 all-MiniLM-L6-v2 作為 Embedding model ，在不同問題下，加上自動辨識版本後，更改不同 Top-K得出的實驗結果。

	Top-K	Precision	AP	NDCG
Linux Update Q1	5	1	1	1
Linux Update Q1	10	1	1	1
Linux Update Q1	20	1	1	1
Linux Update Q2	5	1	1	1
Linux Update Q2	10	0.8	0.92	0.85
Linux Update Q2	20	0.6	0.85	0.69
Linux Update Q3	5	1	1	1
Linux Update Q3	10	1	1	1
Linux Update Q3	20	1	1	1