Semantic Retrieval in Health Education: A RAG-Based QA System from Healthcare YouTube Videos

Author：

一作與通訊 - YONG-XIANG CHANG, MING-HSENG TSENG

剩餘順序：TSAI-WEN LI, CHEN CHOU, YI-JING HO, 卓軒瑋

Abstract(還未寫)

I. Introduction

隨著數位媒體的快速發展，**YouTube 已成為全球健康教育資訊最突出的來源之一**。它不僅受到公眾的廣泛關注，還被醫療機構和公共衛生機構廣泛用於健康促進和知識傳播[1,2]。大量研究表明，**醫療專業人員基於影片的健康交流顯著增強了觀眾對疾病的理解以及他們接受與健康相關的行為改變的意願**[3]。然而，YouTube 上健康教育影片的數量龐大、主題多樣性和製作質量參差不齊，在資訊檢索和理解方面對觀眾構成了重大挑戰。**目前仍然缺乏從這些影片中提取關鍵資訊和核心內容的有效機制**[4,5]。此外，當前的平臺沒有為**自動摘要、語義理解、智慧問答和結構化內容聚合**提供足夠的技術支援。因此，觀眾經常花費大量時間手動搜索和比較資訊，這損害了健康教育提供的有效性和效率[6,7]。另一方面，人工智慧 （AI） 技術**的快速發展**，如**自動語音辨識 （ASR）、自然語言處理（NLP）** 和**深度學習**，以及自動轉錄、語義摘要和影片內容問題生成**等功能變得越來越複雜**[8,9]。特別是在醫療保健領域，**NLP 已成功應用於 YouTube 健康教育影片，用於內容分類、品質評估和語義交互**[10,11]。在此背景下，本研究提出開發一種融合 ASR 和 NLP 技術的**健康教育影片智慧問答系統**。該系統旨在自動轉錄 YouTube 健康影片中的音訊、提取關鍵點並生成結構化摘要。此外，它還包含一個自然語言交互模組，旨在提高使用者檢索、理解和應用健康資訊的效率[12]。

本研究旨在設計和實施一個智慧問答系統，該系統集成了自動語音辨識（ASR）和檢索增強生成技術（RAG），以支援檢索和與 YouTube 上的健康教育影片即時交互，主要關注來自 *icare 的健康內容*管道。擬議的系統自動提取、組織和回應使用者查詢，從而促進對健康相關信息的高效訪問。首先，該系統使用 ASR 將 YouTube 健康影片的口語內容準確地轉錄為文本，隨後，將應用 RAG 技術從轉錄的文本中提取突出的醫學概念和主題資訊，最終將其存儲於結構化知識庫中，以支援後續的查詢和理解[6]。我們開發了一個基於語義嵌入的問答介面，使系統能夠處理使用者提出的自然語言問題，其中該模組從健康影片資料庫中檢

索相關內容並生成簡潔、上下文適當的回答[1]。它專為提高答案準確性和上下文理解而設計，從而改善使用者交互體驗。為了提高可訪問性和參與度，這些功能被集成到一個基於 Web 或移動的直觀互動式平臺中。用戶可以通過文本方式輸入問題，系統根據 YouTube 健康教育影片的內容提供即時回復，並提供相關補充資源的連結。該設計旨在促進健康知識的傳播和可及性[13]。最後，將使用定量指標評估系統的有效性，以評估使用者對影片中健康資訊的理解、回憶和應用的改進 [2]。將與傳統的非互動式影片觀看進行比較分析，以驗證該系統對提高健康素養的切實影響。

II. Related Works

i. LLMs 在健康衛教資訊

隨著大型語言模型 （LLM） 的日益普及和成熟，越來越多的研究探索了它們在健康教育領域的潛在應用。Kasneci et al. （2023） 強調了生成式 AI 在促進教育創新和個人化學習體驗方面的變革潛力[14]。同樣，Mohammed 等人（2025 年）強調了生成式 AI 在支援個人化健康管理和患者教育方面的前景，同時也指出了醫學話語中的關鍵挑戰，例如幻覺效應和情境理解不足 [15]。為了解決這些限制，Mohammed 及其同事提出了「DiabetIQ」系統，該系統將 LLM 與檢索增強生成 （RAG） 集成在一起，以將回應錨定到精心策劃的靜態醫學知識庫，從而提高事實準確性並最大限度地降低風險。但是，這種方法受到其數據源靜態性質的限制，限制了對新出現的健康資訊和即時更新的適應性。為了應對這些限制，本研究建議開發一個基於 Open LLM 架構的健康資訊支持系統，旨在動態攝取和處理健康教育內容。該系統旨在通過為不同的公共衛生場景提供持續的數據集成、互動式查詢和上下文感知回應生成，來反映健康通信的真實資訊需求。

ii. RAG 在健康教育系統中的潛在應用

檢索增強生成 （RAG） 由 Lewis 等人 （2020） 首次提出，將語義檢索技術與生成建模相結合，允許系統在回應生成過程中訪問外部知識庫 [16]。這種混合架構大大提高了事實可靠性和知識覆蓋率，尤其是在醫療保健和公共衛生教育等動態知識環境中。Liu et al. （2023） 指出了 LLM 在準確表示長尾知識的能力方面的一個關鍵局限性，尤其是在內容包括低頻術語、專業術語和快速發展的概念的醫學領域 [17]。RAG 的查詢驅動增強機制通過在推理時提供相關的外部知識來減輕這些限制，使其特別適合於解決特定疾病教育、治療副作用或生活方式建議等主題的健康教育問答系統。Gargari 和 Habibi （2025） 通過全面的敘述性文獻綜述得出結論，RAG 可以顯著降低幻覺風險，可用於患者教育、臨

床決策支持和醫學知識傳遞系統 [18]。他們的發現加強了 DiabetIQ 框架的經驗觀察，並強調了基於 RAG 的架構在醫療環境中的跨領域潛力，為本研究的設計選擇提供了強大的理論和實證基礎。

iii. RAG 架構中的向量資料庫

在 RAG 架構中，向量資料庫是實現高效語義檢索的基礎元件。這些資料庫存儲文字嵌入（表示文本語義內容的高維向量），並使用相似性搜索機制在生成過程中識別上下文相關的文檔。FAISS、Pinecone 和 ChromaDB 等知名系統已經實現了數百萬個此類嵌入的可擴充存儲和快速檢索。與醫療保健應用程式特別相關的是 ChromaDB，這是一個開源的、可在本地部署的向量資料庫，它優先考慮數據隱私，這是健康資訊系統中的關鍵要求。通過支援本地知識庫部署，ChromaDB 可以在保持即時更新功能的同時，對敏感的健康數據進行精細控制。向量資料庫不僅增強了多輪對話系統和關鍵字驅動查詢的語義精度，還支援動態知識庫擴展，解決了 LLM 依賴靜態、固定訓練語料庫的固有限制。Chen et al.（2023） 證明，在知識密集型應用中，將向量資料庫與嵌入模型相結合可以顯著提高檢索品質、回應一致性和系統可擴展性 [19]。基於這些發現，本研究採用 ChromaDB 作為中間存儲層，用於支援互動式健康教育的隱私感知、特定領域的知識檢索系統。

iv. 用於語意表示的嵌入模型

嵌入模型對於 RAG 系統中的語義檢索層至關重要，它將自然語言文本轉換為保留底層語義關係的密集向量表示。嵌入模型的性能直接影響檢索精度和整體響應品質，尤其是在處理特定領域語言和長尾知識時。通用模型（如 all-MiniLM-L6-v2 和 Sentence-BERT）在多語言和跨域搜索任務中表現出強大的性能。然而，在健康和生物醫學領域，BioBERT（Lee et al.，2020）和 BlueBERT（Peng et al.，2019）等特定領域嵌入模型在捕獲醫學術語、癥狀描述和臨床語義結構方面表現出卓越的能力[20,21]。

在這項研究中，我們在為健康資訊服務定製的問答平臺中實施了檢索增強生成（RAG）架構。RAG 框架作為支援大型語言模型（LLM）生成的動態響應的技術基礎，增強了系統為使用者與健康相關的查詢提供準確和上下文感知答案的能力。通過集成語義檢索機制，基於 RAG 的系統增強了語言模型理解和回應罕見或特定領域的健康教育內容的能力，這些內容可能在模型的預訓練參數中沒有得到充分表示。這種混合方法（將神經生成與外部知識檢索相結合）不僅減輕了靜態 LLM 知識截止的局限性，還確保了更可靠、最新和基於證據的輸出，以回

應用戶查詢。增強的檢索功能使系統能夠動態查詢精選健康資源的索引語料庫，包括來自 YouTube 健康教育影片的轉錄內容、結構化醫療指南和經過驗證的健康知識庫。因此，該平臺在健康通信和教育諮詢場景中都變得更靈活，更能回應更廣泛的使用者需求。 **圖 1.** 提供了為本研究改編和配置的 RAG 模型架構的示意圖。該圖概述了關鍵元件，包括 retriever 模組、編碼器-解碼器架構和融合層。

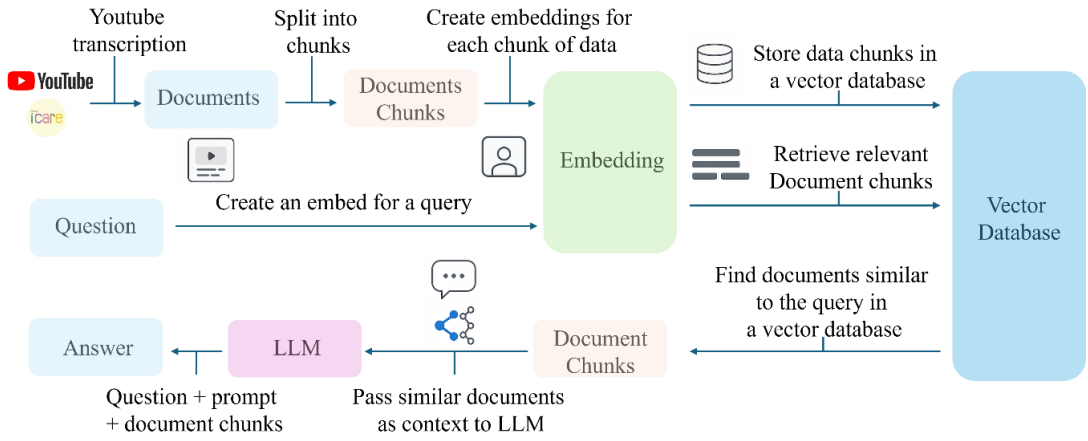


圖 1. 所提出的 RAG-icare 系統的架構允許針對互動式健康資訊服務進行文本生成優化

通過採用 RAG，該系統旨在顯著提高用戶參與度、回應的可信度以及 AI 驅動的健康通信工具的實際效用，尤其是在精度和領域敏感性至關重要的情況下。

III. Materials and Methods

i. 系統架構和應用場景設計

本研究提出了一種用於健康教育的閉環、本地生成式 AI 助手的開發，該助手使用回應式 Web 技術構建，並集成了包括大型語言模型 （LLM）、向量資料庫、嵌入模型和檢索增強生成 （RAG） 在內的關鍵元件。該系統設計為互動式問答平臺，可為使用者運行狀況查詢提供即時的語義相關回應。應用場景包括門診等候區、出院教育和老年人家庭護理，滿足對可訪問和情境化健康資訊的實際需求。 該系統利用了 *icare 愛健康* YouTube 頻道的內容，該頻道提供有關高血壓、糖尿病、失智護理、疫苗接種和營養等主題的全面中文健康教育內容。如圖 **2** 所示，使用自動化管道定期從頻道中提取新發佈的影片。

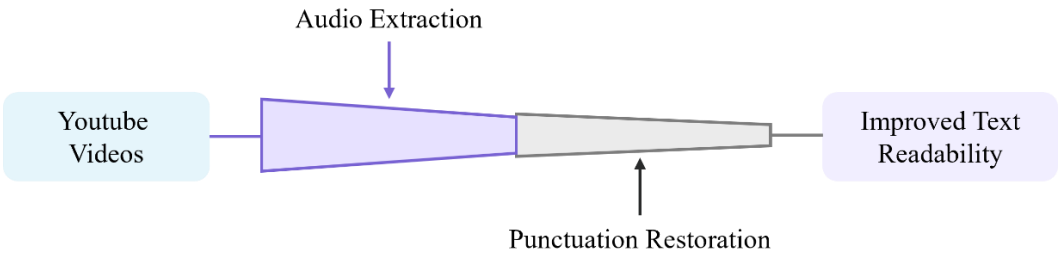


圖 2.處理過程分為上圖階段

令$V = \{v_1, v_2, \dots, v_n\}$為從 YouTube 頻道下載的影片集合，$A = \{a_1, a_2, \dots, a_n\}$，為影片對應的音訊檔案，其中$a_i = FFmpeg(v_i)$。對每個音訊$a_i$，進行語音辨識：

$$\hat{HT}_i = Punct\big(ASR(a_i)\big)$$

使用模型產生通順文本，再從文本中擷取問答對：

$$\mathcal{Q} = \{(q_j, a_j)\}_{j=1}^{m}$$

為擷取出的問答集合，其中$q_j \in Question$，$a_j \in Answer$，嵌入函數 f：$\mathbb{T} \to \mathbb{R}^d$，將文字轉為 d 維向量，對每個$(q_j, a_j)$，生成$v_j = f(a_j)$，儲存於向量資料庫$\mathcal{V} = \{(v_j, q_j)\}_{j=1}^{m}$。使用者輸入問題$q^*$，將轉為向量：$v^* = f(q^*)$透過語意比對找出最相似的 k 筆資料：

$$\mathcal{R} = TopK_k(\mathcal{V}, v^*) = \{(q_{r^1}, a_{r^1}), (q_{r^k}, a_{r^k})\}$$

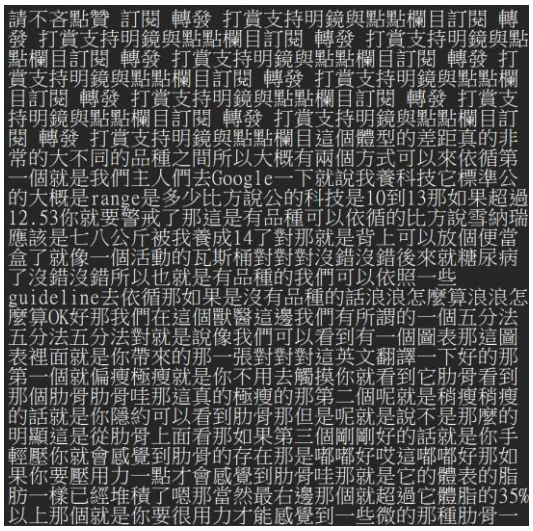
而後使用 LLM 生成回答$\hat{g}a^* = G(g^*, \mathcal{R})$，$G$為 LLMs。

- 音訊轉錄提取

我們將開源工具 yt-dlp 與 FFmpeg 結合使用，以 .m4a 格式從所有影片中下載和提取音訊。容錯設置 （*ignoreerrors=True*） 可確保跨多個影片進行不間斷的處理。

- 問答知識庫的自動化構建

為了支援下游檢索增強生成 （RAG） 應用程式，我們設計了一個自動化管道，旨在將中文語系之音訊轉換為文字，在設計實驗過程中，我們發現 SenseVoiceSmall[X]對於中文語音的理解能力比起 Whisper-large[X]更有效，所需推理的算力較少，時間較快，甚至針對中文斷句處可以自動給予標點符號註記，因此本研究將採用 SenseVoiceSmall 作為主要轉錄模型，Whisper-large 與 SenseVoiceSmall 的轉錄結果比較如 Table X.所示。



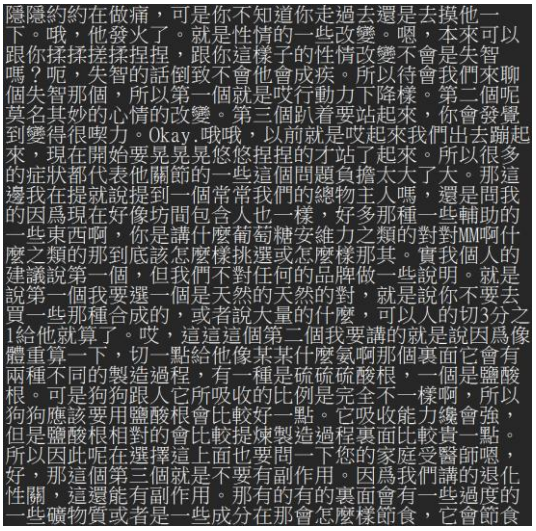


Table X. Whisper-large(左)[X]與 SenseVoiceSmall(右)[X]生成個案結果比較

而後，將轉錄的 .txt 文檔轉換為結構化的 QA 對，並將它們嵌入到本地向量資料庫中。首先，系統使用“？”字元作為問題和答案之間的分隔符，逐行讀取原始.txt 轉錄本。每條記錄都轉換為結構化字典物件，以便進行後續處理。其次，通過 Ollama 運行[23]，使用本地部署的嵌入模型 mxbai-embed-large 為每個答案生成語義向量。相應的問題將存儲為元數據。這些矢量化記錄被封裝為 LangChain Document 物件，並在本地向量資料庫中編入索引以實現高效檢索。

$$\hat{T} = Punct\left(ASR\left(FFmpeg(V)\right)\right) = \mathrm{U}_{i=1}^{n}\hat{T}_i$$

其中$V = \{v_1, v_2, \ldots, v_n\}$為 YouTube 影片集合。

此方法展示了高度的自動化，允許將大量非結構化文本無縫轉換為可搜索的矢量化 QA 格式。與 LangChain 框架整合，該系統可以構建一個檢索 QA 鏈，使 LLM 能夠從本地向量資料庫檢索相關信息並生成基於上下文的回應。

- 語義嵌入和向量資料庫構建(還沒有加入 qwen)

我們使用開源模型 mxbai-embed-large 將清理后的文本轉錄文本轉換為語義嵌入，該模型在語義搜索任務中表現出強大的性能[22]，而後將生成的文本數據(q_j, a_j)嵌入並索引到**本地向量資料庫$\mathcal{V}$中**，用於下游評估階段的語義檢索和生成任務。

$$\mathcal{V} = \{(f(a_j), q_j) | (q_j, a_j) \in ExtractQA(\hat{T})\}$$

其中$f(\cdot)$為嵌入函數，$\mathcal{V}$為向量資料庫，$\hat{T}$為帶標點之轉錄文本。

使用者可以使用自然語言問題與平台互動，例如「糖尿病患者應該吃什麼」或「高血壓患者可以鍛煉嗎？該系統從嵌入式知識庫中檢索並生成醫學情境化的回應。這種方法通過支援語言變化、動態知識更新和上下文感知生成，

超越了傳統的常見問題解答系統，從而提高了健康素養並促進了主動的健康行為。

ii. 系統元件評估和技術比較

- RAG 框架評估

RAG 管道採用 LangChain。LangChain 表現出卓越的模組化、語義控制和定製能力，使其適合構建可解釋和可擴展的 RAG 系統。

- 大型語言模型（LLMs）的比較(正在新增)

對四個本地可部署的 LLM 進行了基準測試： gemma：7b 、 llama3：8b 、 mistral 和 taide-medicine-qa-tw-q6。

$$\hat{a}^* = G\left(q^*, TopK_k\left(\mathcal{V}, f(q^*)\right)\right)$$

其中q^*為使用者問題，$\hat{a}^*$為系統回應，G為 LLM。

- **gemma:7b** 表現出卓越的語義保真度和流暢性，尤其是在同義詞替換任務中[24]。
- **llama3:8b** 在保持回應完整性和上下文一致性方面表現出穩健的性能[25]。
- **mistral** 和 **taide-medicine** 儘管表面的準確性可以接受，但模型對語義釋義的反應較差[26, 27]。
- 向量資料庫的比較

對 ChromaDB、FAISS、Pinecone 和 Weaviate 進行了比較研究。之所以選擇 ChromaDB，是因為 ChromaDB 具有 本地部署能力、與 LangChain 的相容性以及低延遲和數據隱私保護[28]。 FAISS 雖然在靜態檢索方面性能良好，但缺乏即時更新功能[29]。Pinecone 和 Weaviate 等基於雲的平臺會帶來更高的安全風險和基礎設施成本[30]。

- 嵌入模型的比較

mxbai-embed-large 模型因其在語義任務、本地可部署性（不依賴 API）以及數據隱私和成本效率方面的強大性能而被採用。它與 OpenAI Embeddings、Instructor-XL 和 E5 模型進行了基準測試，這些模型雖然功能強大，但需要外部 API 存取並產生運營成本[31]。

iii. 實驗步驟與數據分析方法

1. 模型評估設計

為了客觀地評估系統生成的答案和預定義的參考答案之間的語義保真度，我們採用了兩個廣泛採用的自動評估指標： ROUGE 和 BERTScore。這些指標相輔相成，分別關注詞彙重疊和語義相似性。

- ROUGE 指標（面向詞彙重疊的評估）

ROUGE （Recall-Oriented Understudy for Gisting Evaluation） 是一種傳統但強大的評估指標，通常用於自然語言生成任務[32]。令a_{ref}為參考答案：

$$\text{ROUGE-L}(a_{ref}, \hat{g}a^*) = LCS(a_{ref}, \hat{g}a^*)/|a_{ref}|$$

- BERTScore (Semantic Similarity-Oriented Evaluation)

由於基於詞法的指標在捕獲語義等效性方面具有局限性，因此我們還加入了 BERTScore 來提高評估深度。BERTScore 利用從預先訓練的語言模型（如 BERT）派生的上下文嵌入來計算生成文本和參考文本的標記級嵌入之間的餘弦相似性[33]。這種方法允許在精確單詞匹配之外對語義對齊進行更細緻的測量。BERTScore F1：將文本嵌入為$X \in \mathbb{R}^{l1\times d}, Y \in \mathbb{R}^{l2\times d}$：

$$\text{Precision} = \left(\frac{1}{l_2}\right)\Sigma_{y\in Y}\max_{x\in X}\cos(x,y) \quad \left| \quad \text{Recall} = \left(\frac{1}{l_2}\right)\Sigma_{y\in X}\max_{y\in Y}\cos(x,y) \right. \quad \left| \quad \text{F1} = \frac{2\cdot Precision\cdot Recall}{Precision+Recall}\right.$$

鑒於 ROUGE 已經從詞彙的角度解決了精度和召回率的各個方面，我們在這項研究中專注於 BERTScore F1 分數，以更上下文敏感的方式評估語義相似性。這在中文 QA 設置中特別有價值，因為釋義、同義詞替換和句法變化很常見。

2. 評估測試集的構建

為確保生成回覆具有精確度指標基準，本研究中使用的評估測試集將從 *icare 愛健康* YouTube 頻道的精選影片自動處理和轉錄，產生具有 GroundTruth 的知識問答平測資料，其內容涵蓋各種健康教育主題（包括慢性病管理、老年人護理、營養和疫苗接種），我們根據題目與答案兩個部分分別進行評估。我們使用 Qwen1.5-7B-Chat 進行自動腳本問答，透過讓模型模擬醫療專家助手，閱讀影片轉譯後的文字內容後，整理並提出多組有意義的問答對，問題將避免口語與贅詞，且每個答案簡明扼要，所產生評估測試集，而後針對生成的問題進行效度評估。評估結果如 **Table 1.**所示。

問題生成的評估指標（n=10）

MATRIC	SCORE
ROGUE-1	96.667%
ROGUE-2	10.00%
ROGUE-L	96.667%
BERTScore	88.896%
Cosine Similarity	71.386%

Table 1.問題生成所得之各項評估分數

而 GroundTruth 則是經由人工標註作為基準，根據評估結果可證實此資料集具有

測試意義，人工標註基準如 **Table 2.**所示。以部分樣本為例，透過健康專業人員以及詞彙專家的評估將得到測試問題是否具有參考精確價值，評估範例如 **Table 3.**所示，評估結果如 **Table 4.**所示。

BENCHMARK（n=10）

n	question	Key words	Predict answer
Q1	地中海貧血名字的由來？ (Where does the name thalassemia come from?)	名稱由來 (the name come from) 地中海 (thalassemia)	最早發現在地中海地區盛行 (The earliest discoveries were prevalent in the Mediterranean region)
Q2	健康檢查發現血球比較小顆時要懷疑什麼？ (What should I suspect when I find a small blood cell during a health checkup?)	血球小 (small blood cell)	懷疑是缺鐵性貧血或地中海型貧血 (Iron deficiency anemia or thalassemia is suspected)
Q3	血球正常 MCV 範圍是多少？ (What is the normal MCV range for blood cells?)	MCV 正常值 (normal MCV)	MCV 正常值約 80 到 100 (The normal MCV is about 80 to 100)
Q4	地中海貧血嚴重程度怎麼分？ (How to determine the severity of thalassemia?)	嚴重程度 (severity)	阿法鏈缺 1–2 個是輕度、缺 3 個 中度、缺 4 個重度； 貝塔鏈缺 1 個輕度，缺 2 個重度 (Alpha chain missing 1–2 is mild, 3 is missing, 4 is severe; The beta chain lacks 1 mild and 2 severe)
Q5	如果缺鐵，會建議怎麼做？ (If deficient in iron, what recommend?)	缺鐵 (deficient in iron)	空腹吃鐵劑，搭配維他命 C 增加吸收 (Take iron on an empty stomach with vitamin C to increase absorption)
Q6	天然食物對地中海貧血有幫助嗎？ (Do whole foods help with thalassemia?)	天然食物 (whole food) 地中海貧血 (thalassemia)	效果有限，動物性蛋白效果較好 (The effect is limited, and the effect of animal protein is better)
Q7	骨髓移植的話，可以根治嗎？ (Can a bone marrow transplant be cured?)	根治 (cured) 骨髓移植 (bone marrow)	理論上可以，但有排斥與併發症風險 (Theoretically, yes, but is a risk of rejection and complications)
Q8	血紅素小於多少可以不用當兵？ (How much heme is less than that don’t have to do military service?)	當兵 (do military service) 小於 (less)	小於 12 可以免役 (Less than 12 can be exempted)

Q9	地中海型貧血要做哪些檢查？ (What tests are done for thalassemia?)	檢查 (test) 地中海貧血 (thalassemia)	抽血檢驗：全血球計數、血紅素、MCV (Blood tests: complete blood count, hemoglobin, MCV)
Q10	補充鐵產生副作用時，該怎麼做？ (What should I do if iron supplementation has side effects?)	副作用 (side effects)	改吃三價鐵或用注射鐵劑替代 (Switch to trivalent iron or replace it with injectable iron)

Table 2. 以 10 題為例，生成回答所參照之人工標註基準

n	label	reason
A1	1	都有回答出地中海貧血是一種遺傳性貧血疾病 (All answers are that thalassemia is a hereditary anemia disease)
A2	1	皆有回答出可能代表貧血 (All answers may represent anemia)
A3	1	回答的數值範圍相同 (The answer has the same range of values)
A4	0	兩者回答不一樣 (The answers are different)
A5	0	模型漏掉原文中提到的需要搭配鐵劑才能有效補充的重點 (The model misses the key points mentioned in the original text that need to be combined with iron agents to effectively supplement)
A6	1	模型雖沒提動物優於植物， 但核心方向接近 (Although the model does not mention that animals are better than plants, the core direction is close)
A7	0	模型回答太廣泛，不是針對地中海型貧血 (The model answers are too broad, not for thalassemia)
A8	1	皆有回答到該疾病不用當兵 (All have answered that the disease does not need to be a soldier)
A9	0	模型漏關鍵檢查方式 (Model missing key inspection method)
A10	1	最後結論一致 (Final conclusions are consistent)

Table 3. 以 10 題為例，生成答案經由人工標記所得之各題項精確度

答案生成的自動評估指標（n=10）

MATRIC	SCORE
ROGUE-1	86.381%
ROGUE-2	20.00%
ROGUE-L	86.381%

BERTScore	66.305%
Cosine Similarity	76.261%
Accuracy	60.00%

Table 4. 答案生成所得之各項評估分數

3. 統計分析策略(待增加模型與更新多組統計數據結果)

最終以模型 M 在 RAG 模式與 baseline 模式下的 BERTScore F1 分數集合為 $F^{RAG} = \{f_1{}^{RAG}, \dots, f_n{}^{RAG}\}$、$F^{base} = \{f_1{}^{base}, \dots, f_n{}^{base}\}$進行成對樣本 t 檢定：

$$t = \frac{\bar{d}}{S_d/\sqrt{n}}$$

其中$d_i = f_i^{RAG} - f_i^{base}$，$\bar{d}$為平均差異，$S_d$為標準差。檢定假設$H_0: \mu_d = 0$ vs $H_1: \mu_d \neq 0$，顯著性水準$\alpha = 0.1$。

圖 3. 在 RAG 增強和非 RAG 基線條件下比較了 Gemma：7b 和 LLaMA3：8b 模型之間的語義相似性分數，說明了結合檢索機制在提高語義精度和連貫性方面的切實好處。

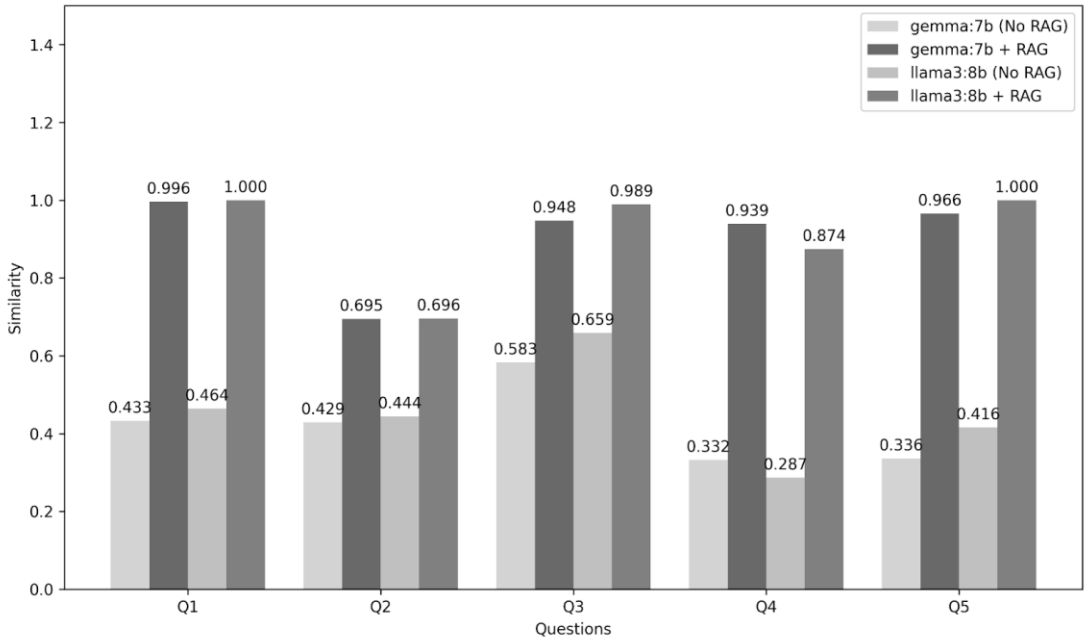


圖 3. gemma：7b 和 llama3：8b 在回答有和沒有 RAG 的問題中的相似性比較

為了評估 RAG 的應用是否顯著影響同一語言模型架構中的模型性能，進行了成對樣本 t 檢定。該測試比較了啟用 RAG 和非 RAG 基線條件之間的語義相似性分數，從而控制了模型之間的可變性，並關注 RAG 集成的模型內效應。所有統計分析均使用 **Python 中的 SciPy 庫**進行，該庫支援成對樣本 t 檢定的

穩健實現。採用**雙尾檢定**來檢測任一方向上的任何統計學顯著差異。鑒於樣本量 n = 5 可能降低統計功效，顯著**性閾值** （α） 設置為 **0.1** 而不是傳統的 0.05。這種調整提高了探索**性分析的敏感性，與 Table 5.**中所示的研究初步調查目標一致。

模型	原生 LLM 平均相似度	使用 RAG 平均相似度	t 值	p 值	是否顯著
gemma:7b	0.429	0.909	-6.915	0.002	是
llama3:8b	0.474	0.912	-6.940	0.002	是

Table 5. 是否有 RAG 的 gemma：7b 和 llama3：8b 的配對樣本 t 檢驗

IV. Results(Table 可能需篩選數據)

本研究使用 **ROUGE** 和 **BERTScore** 指標評估了四種大型語言模型（LLM）的性能——**gemma：7b**、**llama3：8b**、**mistral** 和特定領域的 **taide-medicine-qa-tw-q6**。**Table 6.**顯示了 **ROUGE-1**、**ROUGE-2**、**ROUGE-L** 和 **BERTScore F1** 的計算分數。

LLM	ROUGE-1	ROUGE-2	ROUGE-L	BERT-F1
gemma:7b	0.200	0.000	0.200	0.849
llama3:8b	0.200	0.000	0.200	0.820
mistral	0.200	0.000	0.200	0.811
taide-medicine-qa-tw-q6	0.100	0.000	0.100	0.716

Table 6. 問答評價中 4 種大型語言模型的比較

在三種通用模型（gemma：7b、llama3：8b 和 mistral）中，觀察到 ROUGE-1 和 ROUGE-L 的分數相同 （均為 **0.200**），而所有模型的 ROUGE-2 均記錄為 **0.000**。這表明生成的響應和參考回應之間幾乎完全沒有**二元語法重疊**，這反

映了最小的短語級匹配。相比之下，**領域專業模型** taide-medicine-qa-tw-q6 在所有 ROUGE 指標上的表現不佳， **在 ROUGE-1、ROUGE-2 和 ROUGE-L 上分別獲得** 0.100、**0.000** 和 **0.100** 的分數。這些結果表明，與通用模型相比，該模型在**詞彙重疊**方面表現不佳。在**語義層面，正如** BERTScore F1 **所反映的那樣**，gemma：7b 模型的表現優於其他模型，得分為 **0.849**，其次是 llama3：8b（**0.820**）和 mistral （**0.811**）。taide-medicine-qa-tw-q6 模型獲得了最低的分數 （**0.716**），表明對其生成的輸出與參考答案之間的**語義相似性的**把握相對較弱。

總之，儘管所有模型都表現出**有限的短語級對齊**（由 ROUGE-2 測量），但通用模型通過 BERTScore 表現出**卓越的語義對齊**。這些發現強調了通用 LLM 在捕獲上下文含義方面的穩健性，即使表面詞彙匹配不佳也是如此。相反，**特定領域的模型**雖然專門用於中醫 （TCM） 問答，但似乎需要進一步優化才能在**詞彙**和**語義性能上實現平等**。

V. User Interface and System Feasibility

本研究介紹了閉環、**本地生成式 AI 教學輔助系統的開發**，旨在支援**大型語言模型 （LLM） 的自適應配置**和**嵌入模型**，以滿足不同的教學場景。系統架構允許用戶根據特定的應用需求靈活地選擇模型組合，例如計算約束、域特異性或所需的回應保真度。核心介面具有**互動式問答 （QA） 模組**，使用戶能夠提出自然語言查詢。作為回應，該系統利用**檢索增強生成 （RAG）** 技術從底層資料庫中檢索相關知識並生成語義一致的答案。為了進一步提高上下文準確性，系統提供了 **PDF 文件上傳功能**，允許使用者將課程材料直接導入平臺。這可確保 AI 生成的回應與**課程特定內容保持一致**，從而支援個人化和課程一致的說明。如圖 **4** 所示，QA 介面的左側面板用作檔上傳區域，有助於無縫集成外部教學文檔。

圖 4. 問答系統頁面

VI. Discussion

本研究的重點是開發一個集成了 SenseVoiceSmall 自動語音辨識 （ASR）模型和中文標點符號恢復模型的閉環本地 AI 管道，目的是轉錄來自 YouTube 頻道 *icare 的健康*影片，作為構建結構化問答 （QA） 知識庫的基礎。儘管採用概念設計，但該系統在實施過程中遇到了一些實際限制，如下所述。

i. 來源限制導致數據採集不完整

在影片數據收集過程中，使用了開源工具 yt-dlp，使用頻道主頁 URL（例如 https://www.youtube.com/channel/...）作為主要下載源。但是，此方法只能檢索頻道登陸頁面上直接列出的影片，而無法包含嵌入在嵌套播放清單中的專案。這種疏忽導致了影片語料庫的不完整，因為一些沒有在主頻道介面上展示的影片被無意中遺漏了。預期未來將以頻道的上傳播放清單網址作為目標（例如 https://www.youtube.com/playlist?list=UU...）或提取單個播放清單連結，以確保全面覆蓋。

ii. 不完整的處理工作流程(待補上 youtube cookie 問題)

該技術實現建立在集成 yt-dlp 和 SenseVoiceSmall 的批處理腳本之上。雖然該系統能夠在基本層面處理影片，但出現了幾個關鍵限制 ：沒有過濾從重疊播放清單中下載的重複影片，多語言影片（例如，中英雙語）缺乏語言軌道優先順序，以及檔命名衝突、崩潰恢復和輸入驗證處理不當。這些問題凸顯了改進**容錯**和數據**驗證機制**的必要性，因為不一致的轉錄輸出可能會將錯誤傳播到知識庫

中並損害檢索品質。

iii. 醫療影片語言中的多模態複雜性

健康教育影片的半結構化和多模式性質帶來了額外的挑戰。影片通常包含不完整的語法結構和省略號，具體取決於視覺輔助工具、手勢或幻燈片批注。常見表達方式，如 “this red area shows...”在沒有伴隨的視覺效果的情況下缺乏獨立的意義。先前的研究強調，在這種情況下，純語音模型容易產生語義空白，從而限制了系統保持上下文準確理解和推理的能力[34]。

VII. Conclusion

本研究旨在開發一個能夠根據從目標 **YouTube 健康教育頻道**檢索到的影片內容執行**問答（QA）**的系統。該實施由多階段處理管道組成。

在初始階段，我們使用開源實用程式 **yt-dlp** 通過指定頻道的主頁 URL 來批量下載所有可公開訪問的影片。隨後，**使用** FFmpeg **將下載的影片檔轉換為純音訊 M4A 格式**，從而通過語音辨識模型進行下游處理。考慮到 **YouTube Data API 的配額限制和不穩定性**，我們選擇了非 API 方法，以避免訪問限制造成的中斷，並確保影片採集的完整性。這一決定實現了更可靠的大規模數據採集，而無需依賴基於憑證的限制。在音訊提取之後，我們使用 SenseVoiceSmall 模型將音訊內容轉錄成原始中文文本。通過反覆運算測試和定製，我們建立了一個強大的後處理管道，可以生成**可讀、結構連貫的轉錄本**，適用於知識提取和語義檢索。此階段的產物是一個**龐大的健康教育評測語料庫**，經過結構化和清理，作為 **QA 系統的**知識來源。

為了評估我們系統的有效性，我們實施了檢索**增強生成**（**RAG**）架構，該架構將語義搜索機制與生成語言模型集成在一起。對不同的 LLM 進行了比較實驗，以衡量 RAG 整合對**語義保真度的影響**。使用 **ROGUE 與 BERTScore F1** 作為主要評估指標，結果表明配備 RAG 的模型表現出與參考答案的語義相似度明顯更高，得分接近 **0.9**。這與傳統的僅生成模型形成鮮明對比，後者完全依賴於學習的語言模式，並且經常產生上下文不精確的輸出。

研究結果證實，**即使在**中文環境中，**基於檢索的增強**也可以顯著改善**模型回應中的語義對齊和事實基礎**。這支援了將基於 RAG 的系統應用於**精確健康資訊傳播**的可行性，尤其是在特別要求準確解釋醫學語言的領域。

VIII.Reference

[1] Chung, Joohyun, Sangmin Song, and Heesook Son. "Exploring Natural Language Processing through an Exemplar Using YouTube." *International journal of environmental research and public health* 21.10 (2024): 1357.

[2] Liu, Xiao, Anjana Susarla, and Rema Padman. "Ask your doctor to prescribe a YouTube video: An augmented intelligence approach to assess understandability of YouTube videos for patient education." *Available at SSRN 3711751* (2020).

[3] Liu, Xiao, et al. "Go to YouTube and call me in the morning: Use of social media for chronic conditions." *Liu, X., Zhang, B., Susarla, A., and Padman* (2019): 257-283.

[4] Singh, Yogendra, et al. "Youtube video summarizer using nlp: A review." *International Journal of Performability Engineering* 19.12 (2023): 817.

[5] Guo, Yawen, et al. "YouTube Videos for Public Health Literacy? A Machine Learning Pipeline to Curate Covid-19 Videos." *MEDINFO 2023—The Future Is Accessible*. IOS Press, 2024. 760-764.

[6] Patil, Ratna, et al. "YouTube Video Summarizer Using ASR." *2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*. IEEE, 2024.

[7] Kaulage, Anant, et al. "Edu-lingo: A Unified NLP Video System with Comprehensive Multilingual Subtitles." *2024 Second International Conference on Data Science and Information System (ICDSIS)*. IEEE, 2024.

[8] Zanwar, Siddhi, et al. "Automated Notes and Question Generation." *2024 IEEE 4th International Conference on Software Engineering and Artificial Intelligence (SEAI)*. IEEE, 2024.

[9] Hussain, Nafisa. "Multimodal Language Models for Domain-Specific Procedural Video Summarization." *arXiv preprint arXiv:2407.05419* (2024).

[10] Pothugunta, Krishna, et al. "Assessing inclusion and representativeness on digital platforms for health education: Evidence from YouTube." *Journal of Biomedical Informatics* 157 (2024): 104669.

[11] Guo, Yawen, et al. "YouTube Video Analytics for Patient Engagement: Evidence from Colonoscopy Preparation Videos." *arXiv preprint arXiv:2410.02830* (2024).

[12] Vayadande, Kuldeep, et al. "Efficient content exploration on YouTube: Automatic speech recognition-based video summarization." *2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*.

IEEE, 2023.

[13] Pothugunta, Krishna, et al. "Assessing inclusion and representativeness on digital platforms for health education: Evidence from YouTube." *Journal of Biomedical Informatics* 157 (2024): 104669.

[14] Kasneci, Enkelejda, et al. "ChatGPT for good? On opportunities and challenges of large language models for education." *Learning and individual differences* 103 (2023): 102274.

[15] Mohammed, Saif, et al. "DiabetIQ: An Intelligent Diabetes Management Application with an Integrated LLM-Augmented RAG Chatbot and ML-Based Risk Early Prediction."

[16] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in neural information processing systems* 33 (2020): 9459-9474.

[17] Liu, Nelson F., et al. "Lost in the middle: How language models use long contexts." *arXiv preprint arXiv:2307.03172* (2023).

[18] Gargari, Omid Kohandel, and Gholamreza Habibi. "Enhancing medical AI with retrieval-augmented generation: A mini narrative review." *Digital health* 11 (2025): 20552076251337177.

[19] Liu, Xin, Xuhong Guo, and Qi Liao. "Accurate Estimation of Transport Coefficients Using Model-free Time Correlation Functions in Equilibrium Simulations." *arXiv preprint arXiv:2305.04512* (2023).

[20] Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics* 36.4 (2020): 1234-1240.

[21] Peng, Yifan, Shankai Yan, and Zhiyong Lu. "Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets." *arXiv preprint arXiv:1906.05474* (2019).

[22] Li, Xianming, and Jing Li. "Angle-optimized text embeddings." *arXiv preprint arXiv:2309.12871* (2023).

[23] Marcondes, Francisco S., et al. "Using ollama." *Natural Language Analytics with Generative Large-Language Models: A Practical Approach with Ollama and Open-Source LLMs*. Cham: Springer Nature Switzerland, 2025. 23-35.

[24] Team, Gemma, et al. "Gemma: Open models based on gemini research and technology." *arXiv preprint arXiv:2403.08295* (2024).

[25] Grattafiori, Aaron, et al. "The llama 3 herd of models." *arXiv preprint arXiv:2407.21783* (2024).

[26] Chaplot, Devendra Singh. "Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed." *arXiv preprint arXiv:2310.06825* (2023).

[27] Hsu, Chan-Jan, et al. "Advancing the evaluation of traditional chinese language models: Towards a comprehensive benchmark suite." *arXiv preprint arXiv:2309.08448* (2023).

[28] Jeong, Cheonsu. "A study on the implementation of generative ai services using an enterprise data-based llm application architecture." *arXiv preprint arXiv:2309.01105* (2023).

[29] Pan, James Jie, Jianguo Wang, and Guoliang Li. "Survey of vector database management systems." *The VLDB Journal* 33.5 (2024): 1591-1615.

[30] Perron, Brian E., et al. "A Primer on Word Embeddings: AI Techniques for Text Analysis in Social Work." *arXiv preprint arXiv:2411.07156* (2024).

[31] Muennighoff, Niklas, et al. "MTEB: Massive text embedding benchmark." *arXiv preprint arXiv:2210.07316* (2022).

[32] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." *Text summarization branches out*. 2004.

[33] Zhang, Tianyi, et al. "Bertscore: Evaluating text generation with bert." *arXiv preprint arXiv:1904.09675* (2019).

[34] Srinivasan, K., Li, J., & Marsic, I. (2020). *Multimodal Understanding in Instructional Videos*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.