東吳大學

資料科學系專題計畫報告書

使用大型語言模型與向量資料庫的 Podcast 智慧搜尋系統

楊媛晴、曾子珊、陳妙音、林依婷

指導教授：邱嘉豪　助理教授

摘要

隨著播客收聽量在近年快速成長，使用者對內容檢索的需求亦日益提升。然而，傳統以關鍵詞為基礎的搜尋方法在處理非結構化音訊資料時受限甚多，多仰賴自動語音辨識（ASR）轉錄文字進行比對，易因轉錄錯誤與語意缺失而降低檢索的精確度與相關性。為解決上述問題，本研究提出一個結合大型語言模型（LLMs）與檢索增強生成（RAG）的 Podcast 智慧搜尋系統，透過語意向量化、語料分群與自我校正機制，強化系統對非結構化語音內容的理解能力。

因此，本研究團隊從資料收集、資料前處理、語意生成到輸出前端的各階段進行系統性設計。整合 OpenAI Whisper 進行音訊轉錄，並採用 Hugging Face 語意向量化模型建構向量資料庫；同時加入語意聚合與動態去冗分群機制，使查詢時能自動縮小搜尋空間、保留最具代表性的語料。在語意生成階段，我們設計三項校正評分（事實性、一致性、切題性），提升 LLM 輸出品質，以確保輸出內容兼具語意深度與正確性。前端部分則透過 Gradio 建立互動介面，提供查詢呈現、節目資訊瀏覽與個人化 Podcast 參考。實驗最後，為驗證本系統效能，本研究以東吳大學資料科學系一年級學生及一般成人為受測者進行量化問卷調查。

經過實驗結果分析，本研究提出的動態去冗分群機制在效能與生成品質上皆帶來明顯改善：在查詢時間方面，加入分群後平均搜尋時間約縮短 19.5%，中位數亦下降約 12.4%，而標準差由 9.07 降至 3.97，顯示系統反應時間更為穩定。而在生成內容品質上，動態去冗搭配自我校正迴圈可顯著提升回覆的事實性，於 t 檢定結果達高度顯著水準（$p < 0.001$），且 Cohen’s d 達 6.17，顯示效果量極大。此外，我們採用七分量表作為基準作問卷調查，共回收了 42 份有效樣本。在三個情境問題中，本系統的回覆品質之評分均明顯優於 Apple Podcast 與 Spotify 等傳統平台（$p < 0.01$），整體滿意度平均介於 4.45～5.05 分之間，顯示使用者普遍肯定本系統在搜尋品質、語意精準度與操作體驗上的表現。

綜合上述，本研究所提出的 Podcast 智慧搜尋系統成功提升非結構化音訊資料的檢索效能與回答品質，不僅解決現行平台在語意理解上的限制，也展示出實際應用價值與未來延伸至多語言、多主題與大規模音訊資料的潛力。

目錄

第一章、 緒論.....1

1.1 背景介紹.....1

1.2 研究動機.....2

1.3 研究目的.....2

第二章、 文獻探討.....3

2.1 傳統播客檢索方法的局限性(目前純複製 hackmd)(婷:我覺得這段不用改) 3

2.2 大型語言模型在語意理解與搜尋的應用(目前純複製 hackmd).....4

2.3 利用向量資料庫增強語意搜尋(目前純複製 hackmd).....5

第三章、 研究方法.....6

3.1 系統整體架構與語意搜尋流程.....6

3.2 動態去冗分群機制.....7

3.3 自我校正迴圈機制.....9

3.4 系統評估與實驗設計.....10

3.4.1 系統參數與模型效能之量化分析.....10

3.4.2 使用者問卷與體驗評估設計.....12

第四章、 實驗結果分析.....13

4.1 動態去冗分群個數之選擇分析.....13

4.2 加入動態去冗分群之秒數分析.....15

4.3 自我校正迴圈閾值之設定分析.....16

4.4 回覆事實性分數分析.....17

4.5 問卷調查分析.....18

4.5.1 系統回覆評分之分析.....18
4.5.2 Podcast 智慧搜尋系統滿意度分析.....20
第五章、 結論與未來研究.....21
5.1 結論.....21
5.2 未來研究.....21
第六章、 參考文獻.....23

圖目錄

圖 三-1 系統流程圖.....6

圖 三-2 動態去冗分群機制流程圖.....8

圖 三-3 自我校正迴圈機制流程圖.....9

圖 四-1 平均值雷達圖.....14

圖 四-2 事實性分數盒鬚圖.....17

表目錄

表 四-1 不同分群設定比較表 （紅色為最終選擇之分群設定）.....13

表 四-2 加入動態去冗前後之秒數統計分析.....15

表 四-3 加入動態去冗分群機制前後之各題執行秒數比較表.....15

表 四-4 迴圈中三項評分閾值門檻設定分析.....17

表 四-5 事實性分數統計分析.....17

表 四-6 系統回覆評分之各項指標比較結果.....19

表 四-7 系統滿意度之平均數與標準差比較結果.....20

第一章、 緒論

1.1 背景介紹

隨著數位音訊內容的快速發展，Podcast 已成為資訊獲取、娛樂與學習的重要媒介。根據多份報告指出，全球 Podcast 使用者人數逐年增加，對內容多樣性與個人化體驗的需求也日益提升。相較於文字或影片內容，Podcast 屬於非結構化的音訊資料，使得其內容搜尋、分類與語意理解更具挑戰。因此，建置一套能夠理解語意並提供個人化推薦的智慧系統，對於提升使用者體驗與資訊可近性具有重大意義(Besser, Larson et al. 2010)。

儘管現有搜尋引擎（如 Google ）已能透過關鍵字或語意搜尋協助使用者定位資訊，但其運作仍仰賴使用者明確表達需求，並自行篩選搜尋結果。此種被動式搜尋模式對於 Podcast 的探索情境而言並不完全合適，因為使用者往往缺乏明確的查詢意圖，而是希望根據個人興趣獲得主動、啟發式的內容推薦。為了彌補這樣的落差，本研究提出一個結合大型語言模型（ LLMs ）與向量資料庫的智慧搜尋系統，能主動理解內容語意與使用者偏好，並在無需明確查詢指令的情況下，提供具情境感知（ Context-aware ）與對話式互動的推薦體驗(Jing, Su et al. 2025)。

近年來，大型語言模型的快速發展，使得機器能更準確地理解語意結構與使用者意圖。隨著語音辨識與自然語言處理技術的進步，音訊內容可經由自動轉錄轉換為文字，並透過語意嵌入技術將文本表示為可計算的高維向量，使非結構化的音訊資料能以結構化方式被分析與檢索(Wu, Zheng et al. 2024)。本研究以此為基礎，發展出一套結合語音轉文字、語意向量化與語意分群的內容處理流程，藉以建立具語意理解能力的 Podcast 向量資料庫。

以上，本研究旨在設計並實作一套整合大型語言模型與向量資料庫的 Podcast 智慧搜尋系統，具體目標如下：

(1) 探討如何將 Podcast 音訊內容轉換為具語意代表性的文本與向量，作為語意檢索與推薦依據。

(2) 建構結合大型語言模型理解能力與向量資料庫檢索效能的系統架構。

(3) 評估本系統在檢索品質與使用者滿意度上的表現，並與傳統搜尋方式進行比較

1.2 研究動機

隨著數位音訊內容快速普及，Podcast 已成為人們獲取資訊、娛樂休閒與自主學習的重要媒介，其使用者人數逐年增加，內容主題涵蓋科技、商業、戲劇、心理學與文化等多元領域(Besser, Larson et al. 2010)。Podcast 兼具可攜性與沉浸式體驗，使使用者能在通勤、運動與休息等情境中輕鬆吸收內容。然而，Podcast 屬於非結構化音訊，其內容相較於文字或影像媒介更難以被有效搜尋、分類與分析，導致資訊可檢索性與利用率受到限制。

現有 Podcast 平台多依賴節目標題、主持人名稱或簡短摘要進行關鍵字式搜尋（如 TF-IDF、 BM25)，難以理解自然語言查詢背後的語意脈絡，也無法處理模糊式、情境式或主題導向的查詢需求。例如使用者可能想尋找「近期熱門影集的討論」或「有關科幻電影的觀點分享」，但傳統搜尋往往無法精準匹配內容，搜尋體驗因此受限。尤其在戲劇類 Podcast 中，節目標題與實際討論內容常無法直接對應，使使用者必須逐集收聽才能找到所需資訊，加劇資訊過載問題。

基於上述挑戰，本研究以戲劇類 Podcast 為研究範例，建構一套結合大型語言模型（LLMs）與向量資料庫的語意型智慧搜尋系統。透過語意理解、內容嵌入與主題匹配能力，系統能協助使用者以更直覺且高效率的方式探索 Podcast 內容，突破傳統關鍵字搜尋的侷限，進而提升內容檢索精準度與整體使用體驗(Wu, Zheng et al. 2024)。

本研究的核心觀察如下：

(1) Podcast 普及快速，但內容多為非結構化音訊，搜尋與分析較為困難，造成內容利用率受限。
(2) 傳統關鍵字搜尋依賴標題與摘要，無法理解自然語言語意，對主題式或模糊式查詢支援不足。
(3) Podcast 數量快速成長，尤其戲劇類內容主題多層次、標題語意不足，使用者搜尋效率低、易受資訊過載影響。
(4) 本研究提出結合 LLMs 與向量資料庫的語意型智慧搜尋系統，以提升 Podcast 內容檢索的精準度與探索效率。

1.3 研究目的

(1) 建置一套高準確度、高效率、可用性佳的 Podcast 智慧搜尋系統。

整合語意檢索、動態分群與自我校正生成，使系統能在非結構化音訊資料上提供更穩定、更可靠的搜尋結果，做出在 Podcast 搜尋領域中，一個具可行性、

可擴充且能改善現行技術限制的完整解決方案。

(2) 確保搜尋與生成結果能被清楚、有效地呈現給使用者。

建置直覺互動的前端介面，提升系統的操作體驗與資訊可讀性，讓使用者快速理解回覆內容，增加操作的流暢性與整體使用感受。

(3) 提升 Podcast 內容的語意理解與搜尋準確度。

導入大型語言模型（LLMs）與向量資料庫，改善傳統關鍵字檢索無法理解語意的問題，使系統能理解自然語言查詢背後的語意，而非僅做逐字比對，提高對內容脈絡的掌握度(Wu, Zheng et al. 2024)。

(4) 改善 LLM 回覆易產生幻覺與語意偏差的問題。

導入自我校正機制，以提升生成內容的事實性、一致性與切題性，讓系統的最終回覆更精準、更可信，避免因 LLM 幻覺特性而產生不當推論、錯誤資訊或偏離問題主旨。

(5) 降低冗餘資料對檢索效能及正確性的負面影響。

採用動態去冗與 K-Means 分群機制，以縮小搜尋空間、提升資料代表性，避免搜尋時因為向量資料庫過大，造成查詢延遲及語意稀釋，確保系統回傳的語料更加聚焦且具備語意覆蓋度。

第二章、 文獻探討

2.1 傳統播客檢索方法的局限性

傳統 Podcast 檢索方法的限制 近年來，Podcast 消費量大幅成長內容涵蓋新聞、教育、娛樂與個人成長等多種領域。不同於文字、圖片或影片等資料形式，Podcast 主要以非結構化音訊的形式存在，缺乏標準化的資料結構，使得傳統的索引與檢索系統難以有效處理，對使用者搜尋與探索內容造成重大挑戰。

目前主流的 Podcast 搜尋引擎，例如 Spotify 與 Apple Podcasts，大多仰賴節目標題、描述以及發布日期等 metadata（中繼資料）。儘管這些 metadata 提供了基本的上下文，但它們通常由創作者手動撰寫，內容往往無法精確反映節目的細節或語意主題。已有研究指出，單靠 metadata 難以滿足使用者的資訊需求，當使用者試圖搜尋更具體的主題或語意細節時，會遭遇內容發現的瓶頸(Carterette, Jones et al. 2021)。

為了克服上述限制，一些研究嘗試導入自動語音辨識技術(ASR)，將 Podcast 的

語音內容轉換為文字，以實現內容導向的搜尋。然而，Podcast 常見的轉錄挑戰包含語音錯誤、口語化表達、語音不連貫及背景雜訊，這些因素會降低 ASR 準確率，進而影響搜尋成效(Spina, Trippas et al. 2017, Clifton, Reddy et al. 2020)。此外，ASR 系統在處理 Podcast 常見的新詞、專有名詞或特定領域術語時，也常面臨挑戰。像是 PodCastle 等系統試圖透過群眾協作詞典擴展與自動詞彙學習來改善這些問題(Goto and Ogata 2011)。

即使 ASR 的轉錄品質不佳，研究發現使用者在摘要與相關性判斷任務上，仍能從轉錄文本中獲得幫助(Spina, Trippas et al. 2017)、(Tsagkias, Larson et al. 2008)。其他方法則會將轉錄內容進行語意切片，並使用詞雲或嵌入式語意表示來提升內容導覽與檢索體驗(Fuller, Tsagkias et al. 2008)、(Eskevich 2014)。然而整體而言，Podcast 的檢索效能仍高度依賴 ASR 的準確性，當辨識錯誤率提高時，檢索效果也會明顯下降(Senay, Linares et al. 2011, Lee, Glass et al. 2015)。

近年來，部分研究開始嘗試跳脫傳統 ASR 管線的限制，探索直接從語音訊號中擷取語意資訊的可能性。舉例來說，有學者透過語音特徵與預訓練語言模型結合，即使轉錄品質不穩，也能辨識出 Podcast 節目的語意段落與結構(Jing, Schneck et al. 2021)。

總結而言，雖然 metadata 與 ASR 提供了 Podcast 索引與搜尋的初步基礎，但在語意層面上的檢索仍顯不足。這促使研究者開始探討整合大型語言模型(LLM)與向量資料庫的進階方法，以更有效捕捉語音內容背後的語意脈絡與使用者意圖。

2.2 大型語言模型在語意理解與搜尋的應用

大型語言模型於語意理解與搜尋的應用 大型語言模型（Large Language Models, LLMs）如 [GPT](Brown, Mann et al. 2020)、[BERT](Devlin, Chang et al. 2019)與[T5](Raffel, Shazeer et al. 2020)的出現，顯著改變了機器對人類語言的理解與處理方式。傳統的搜尋系統大多依賴關鍵字比對與固定規則，往往無法捕捉使用者查詢背後更深層的語意意圖。而 LLM 則具備理解自然語言中語境脈絡與語意關聯的能力，使得資訊檢索更加智慧與靈活。

LLM 的一大創新在於，它們能夠將模糊、口語化或結構鬆散的查詢語句，轉換成具語意結構的向量表示。這表示搜尋系統不再只是根據表面的字詞比對結果，而是能推論使用者意圖，並將其與文件內容的語意進行比對。這種方法特別適用於複雜查詢與語意重構需求高的情境（例如：非關鍵字式的探索性查詢），能顯著提高檢索的相關性與精準度(Khattab and Zaharia 2020)。

此外，LLM 能夠將查詢與文件內容同時轉換成語意向量（vector embeddings），並映射到同一語意空間中進行比對。這使得系統可以透過向量相似度來進行搜尋，比傳統稀疏型的關鍵字檢索更具彈性，能對抗同義詞、重述、語序調整等語言變體的干擾(Reimers and Gurevych 2019)。

在語音內容檢索(spoken content retrieval)方面，LLM 可與語音轉文字(ASR)流程結合，進一步提升語意理解的深度。透過先進的 LLM，轉錄後的文本能生成代表其語意的摘要或向量，就算原始轉錄中存在錯誤或口語雜訊，也能保留語意上的精準性，已被應用於 Podcast 搜尋、主題分段與語意推薦等任務(Lugosch, Ravanelli et al. 2019, Jing, Schneck et al. 2021)。

LLM 也廣泛應用於查詢理解（query understanding）、問答系統（question answering）與語意排序（semantic reranking）任務上，能根據使用者查詢的上下文、語用脈絡與語意層級進行更準確的文件召回與排序(Thakur, Kumar et al. 2022)。

總結來說，LLM 正在重塑資訊檢索的典範，從傳統的關鍵字比對轉向語意導向的搜尋架構。這種轉變對 Podcast 等以非結構化音訊為主的媒體尤其重要，在這類應用中，語言表達高度非正式且 metadata 稀疏，結合語音辨識與 LLM 的語意理解能力，將是建構次世代 Podcast 檢索與推薦系統的關鍵。

2.3 利用向量資料庫增強語意搜尋

雖然大型語言模型（LLMs）能夠將使用者查詢與文件內容轉換為高維度的語意向量，但若要在實務中實現語意搜尋系統，其向量的儲存、索引與查詢效率也同樣關鍵。這促使了向量資料庫（Vector Database, VDB）的興起，其專為處理高維嵌入資料而設計，支援語意相似度檢索。

不同於傳統以倒排索引為核心的關鍵字比對系統，向量資料庫支援基於向量相似度（如 cosine similarity 或歐幾里得距離）的近似最近鄰搜尋（ANN），能在詞彙不重合的情況下，找出語意相關的內容。這對於非結構化、語言表達多樣的資料（例如 Podcast）尤其重要。

目前主流的向量搜尋系統如 FAISS、Pinecone、Weaviate、Milvus 等，皆導入先進的索引技術，例如 HNSW（階層式圖索引）、IVF（倒排檔案索引）與 PQ（產物量化）等演算法，以在精準度與查詢效能之間取得良好平衡(ARADHYA and DIVYA 2024)。

向量資料庫為 LLM 語意搜尋奠定了基礎性架構，讓語意向量不僅能被理解，還

能被有效地查詢與篩選。隨著語意嵌入逐漸成為語言資訊的主要表示方式，LLM + Vector Database 的組合，正快速成為語音搜尋、推薦系統與非結構化知識檢索的核心

第三章、 研究方法

3.1 系統整體架構與語意搜尋流程

本研究提出一套針對長音訊內容之語意導向的智慧搜尋架構（圖 三-1）。系統設計由三個相互銜接的模組構成，分別對應「音訊語料結構化」、「語意檢索」、「生成式回答」三個層次，旨在提供比傳統關鍵字式演算法更高的語意理解能力與跨片段推理能力。

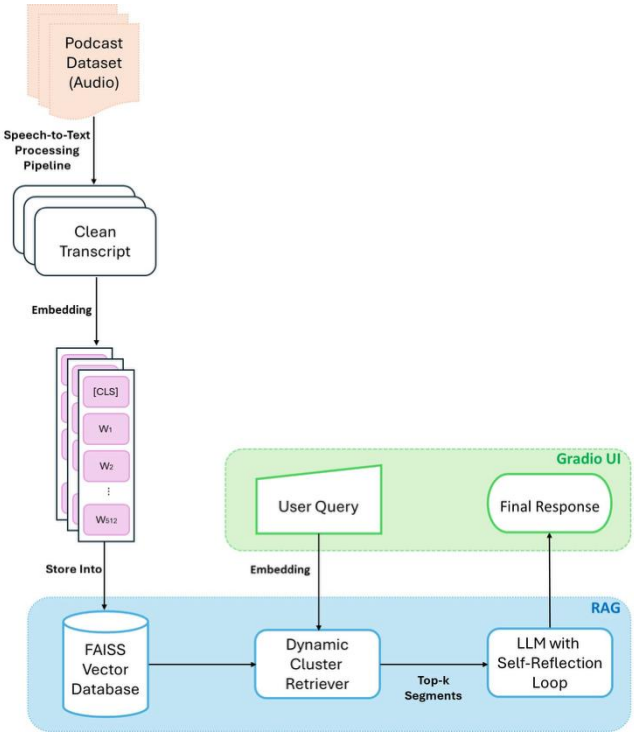

圖 三-1 系統流程圖

根據「音訊語料結構化」、「語意檢索」、「生成式回答」分別用使用開源 api 以及演算法去架設系統。

先對 Podcast 原始音訊進行語音轉文字處理，以取得可靠且具語意完整性的逐字稿。為提升文本品質，使用生成式 AI 進行句法與語意修正，讓逐字稿在斷句、語氣連貫與語意完整度上更適合作為後續語意嵌入的輸入。接續利用大型

語言模型衍生之語意嵌入模型，將修正後的文本片段映射至連續向量空間，形成可進行高維索引的語意表徵。所有語意向量均儲存於 FAISS 向量資料庫中，構成後續語意檢索流程的基礎語料。

緊接著讓查詢語意檢索與語料選取，當使用者提交查詢時，系統將其轉換為查詢向量，並以此向量在資料庫中搜尋具高語意相似度之文本片段。與傳統 TF-IDF 或 BM25 本系統透過向量空間檢索，以語意分布為基礎取得候選片段。我們的系僅依字面重疊進行比對不同，統搭配一套自動化語料聚合策略(K-means)，在語意集合中選取最具代表性的文本子集，確保 LLM 在生成回答時處於一致且高相關性的語意背景下。

為了保障生成式回答與品質控管，系統將精選後的語意片段輸入具自我評估能力的大型語言模型，使其整合跨片段資訊，產生具摘要性與語意連貫性的最終回答。為降低生成式模型的幻覺問題，本架構採用以提示工程為架構的自我校正流程，確保回答內容在事實性、語意一致性與與查詢的切題性上均維持在可接受範圍內。最終結果再透過使用者介面呈現，完成整體查詢回應流程。

本架構透過「向量化語意索引 + 多片段語意聚合 + LLM 回覆生成」三層式設計，使系統能有效突破傳統文字比對式搜尋的限制，在處理長音訊語料、跨句跨段語意推論與單一主題領域中處理跨語境、多樣表述的查詢時，展現更高的精確度與穩健性。

3.2 動態去冗分群機制

為了提升使用者在查詢 Podcast 資訊時的回應速度，所以我們希望避免 LLM 在每次回應時都從完整的向量資料庫進行搜尋，因為此作法不僅增加延遲，也使計算成本顯著提高，所以我們希望能讓 LLM 可以更專注於有代表性的局部資料查詢。基於此需求，我們進一步分析現有文獻中有針對 RAG 進行資料增強的方法，尋找能有效縮減搜尋空間的技術。

本研究的動態去冗分群機制設計架構參見圖 三-2，我們在設計 Podcast 智慧搜尋系統時，參考了 SAKR（Streaming Algorithm + K-Means for Retrieval）方法的核心精神。SAKR 的主要貢獻在於：透過 K-Means 將大型語料預先分群，使查詢時能先定位至最相似的語意群組，再在群內進行向量檢索，藉此大幅降低搜尋空間並提升召回效率。該方法證實了「以群為單位的語意檢索」對於多主題語料具有更好的可擴展性與檢索準確度(Kang, Zhu et al. 2024)。

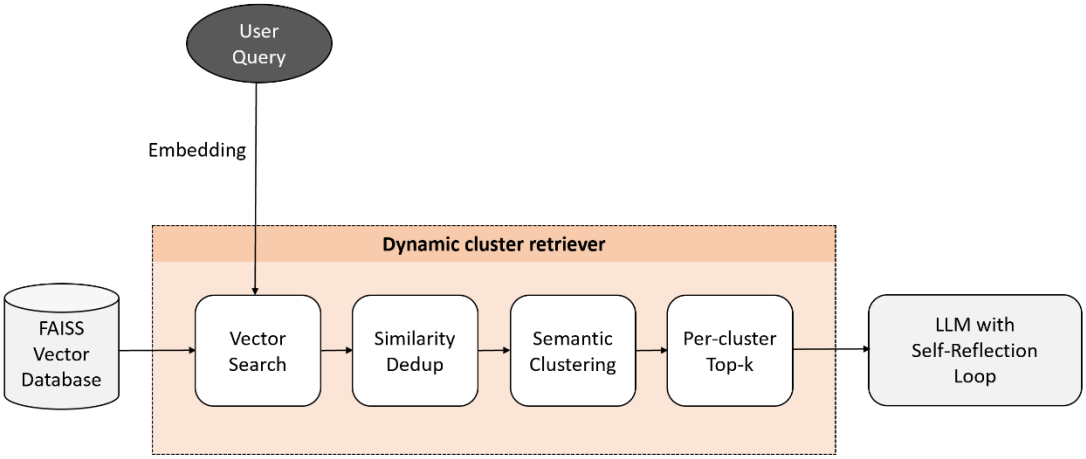

圖 三-2 動態去冗分群機制流程圖

然而，SAKR 的分群方式通常建立於固定 K 值與靜態語料特性，較適用於穩定、批次式更新的資料環境。相較之下，本研究的資料庫為多節目、多主題的戲劇類 Podcast，其語意分布具有高度波動性；不同節目可能針對相同戲劇進行評論，也可能在不同集數討論截然不同的議題。因此，若沿用固定群數的 K-Means，可能無法呈現語料之間的細微差異，也不利於針對使用者提問進行動態調整。

基於此差異，本研究並未直接採用 SAKR 的演算法，而是取其「分群以縮小搜尋空間」的架構概念，並根據 Podcast 語料特性提出改良式流程，形成本系統之動態去冗（Dynamic Pruning）與自適應分群（Adaptive Clustering）機制。

（1）動態去冗機制 Dynamic Vector Pruning

為避免在每次查詢時對完整語料庫進行檢索，本研究會先計算查詢向量與語料中心向量的餘弦相似度、判斷該查詢是否屬於「熱門主題」或「冷門主題」，並使用本研究自行設計的保留比例函式(a)，

$$keep = keep_{min} + (keep_{max} - keep_{min}) \times (x^{\alpha}) \qquad (a)$$

其中x為 prompt 與中心向量距離之正規化值。α為彎曲係數，控制保留量對距離的敏感度。$keep_{min}$與$keep_{max}$為避免極端保留或極端刪減。

此機制可在檢索前先濃縮語料空間，讓後續語意比對更加集中於與查詢高度相關的內容。

（2）自適應分群機制（Adaptive Clustering）

去冗後的語料會再進行 K-Means 分群，但本研究提出比 SAKR 更進一步的做法，會根據查詢語意分散度（熵）動態決定 K，而非使用固定分群數。我們使用公式(b)決定 K：

$$K = \sqrt{N} \times (0.7 + 0.6 \times H_n) \quad \text{(b)}$$

其中N為去冗後向量數量。H_n為查詢與語料相似度分布的熵。

此自適應分群方式，使系統在不同類型提問下能自動調整語意粒度，避免固定 K 造成的偏誤。而此程序是基於多節目 Podcast 特性所發展的機制，後續章節將針對不同 Top-k（如 2×3、3×3、4×3、5×2、6×2）進行量化比較，分析其在內容相關性、完整性、事實性與語意一致度上的差異，並驗證本系統預設值（4 群 × 每群 3 筆）的合理性。

3.3 自我校正迴圈機制

為了優化大型語言模型之生成內容，本研究加入以提示工程為技術架構的自我校正回覆的機制（圖 三-3），基於 Self_Reflection_Medical 模型 (Ji, Yu et al. 2023)進行修改，藉由評估指標、添加新指令及迭代生成，改善大型語言模型的回覆內容。

在 LLM 輸出回覆前，先透過三個指標評分：事實性(fact)、一致性(consistency)及切題性（entailment），評分指標之詳細說明見 3.4.1.3。若首次生成的回覆之三個分數有高於我們設定的閾值，則可以直接輸出；若無，則進入校正迴圈，並添加指令內容，透過迭代的自我修正過程減少幻覺現象。此方法包含兩個迴圈：知識迴圈與回覆迴圈。

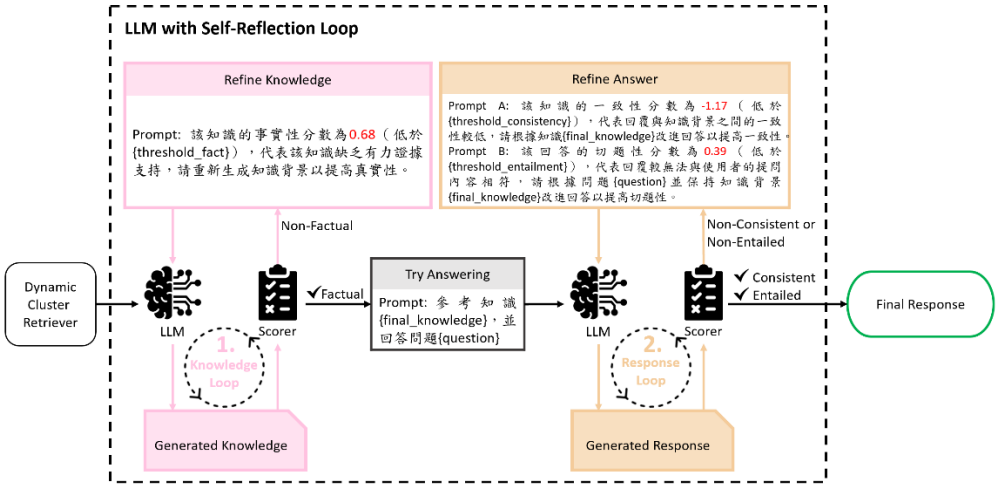

圖 三-3 自我校正迴圈機制流程圖

(1) 知識迴圈

此步驟旨在奠定後續 LLM 文字生成步驟的知識背景基礎，利用 LLM 善於整合相關資訊的特性，重新建構輸出內容主軸。該迴圈會加入以下指令：「該知識的事實性分數為 x，低於閾值$\hat{x}$，代表該知識缺乏有力證據支持，請重新生成知識背景以提高真實性。」引導模型根據 3.2 分群結果優化知識背景的內容，迴圈運行直到事實性分數高於閾值，然而，為避免執行階段太過耗時，本研究設定最大迭代次數為三次。

(2) 回覆迴圈

此環節同時校驗 LLM 輸出的兩個層面：其一，回覆內容與所依據的資料集知識背景保持一致；其二，回覆內容能切題地回應使用者的提問。

當知識迴圈所產生的知識達到所需品質後，此階段會先基於最終通過評分標準的知識產生回覆，使用指令：「參考知識，並回答問題」，並再次進行一致性評分與切題性評分，若生成之回答仍低於其中一項之閾值，則加入指令：「該回答的一致性分數為 x，低於閾值$\hat{x}$，代表回覆與知識背景之間的一致性較低，請根據知識改進回答以提高一致性。」或「該回答的切題性分數為 x，低於閾值$\hat{x}$，代表回覆較無法與使用者的提問內容相符，請根據問題並保持知識背景改進回答以提高切題性。」，迴圈將重複執行直到兩項分數皆高於閾值，但為了有效控制運行時長，設定三次為迭代上限。

3.4 系統評估與實驗設計

系統評估採取「系統參數與模型效能之量化分析」與「使用者主觀問卷評估」兩大面向，旨在全面驗證本研究導入之動態去冗分群與自我校正迴圈是否能在搜尋效能、回覆品質與整體使用體驗上產生實質提升。

3.4.1 系統參數與模型效能之量化分析

針對系統核心技術進行參數測試，以選定最佳 Top-k 分群組合與適當的校正閾值。本階段以系統行為與模型輸出為分析對象，並採用多項量化指標評估效能與生成品質。

3.4.1.1 動態分群設定之生成品質

動態去冗分群的比較以不同之 Top-k 組合為實驗樣本。生成品質採人工匿

評方式進行，由受測者依五項品質面向進行評分，包括：內容相關性、回覆完整性、事實正確性、可讀性與流暢度，以及語意一致性。

在取得匿名評分數後，本研究再根據以下三項統計指標進行量化比較，以評估各組分群設定之優劣：

- 平均值：用以反映整體生成品質的高低。

$$\text{Mean} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 標準差：用以衡量品質的穩定度與結果一致性。

$$\text{STD} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- 跨題一致性：用以觀察不同題目中是否能持續取得較佳表現。計算方式為:計算五種分群方法在五題各指標的個數。

3.4.1.2 動態分群介入前後之搜尋效能

為評估動態分群機制在「Baseline: 全資料庫向量比對」改成「New: 分群後局部比對」之前後兩者的平均搜尋時間，本研究記錄數據，並根據以下量化指標進行效能比較：

- 平均值：反映整體查詢耗時的高低。

$$\text{Mean} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 中位數：排除極端值後，呈現查詢時間的典型表現。

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)} \qquad \text{Median} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

- 標準差：衡量查詢時間的波動程度與穩定性。

$$\text{STD} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- 改善率：評估每題查詢在加入分群後的時間改善幅度。

$$\text{Improvement} = \frac{\text{Baseline} - \text{New}}{\text{Baseline}}$$

- 加速倍率：表示加入分群後系統效能的相對提升程度。

$$\text{Speedup} = \frac{\text{Baseline}}{\text{New}}$$

透過上述指標，本研究得以檢驗動態分群機制在不同語意範圍與向量密度條件下的實際效能改善情形。

3.4.1.3 自我校正迴圈之評估指標

事實性分數（fact）

事實性分數旨在驗證知識是否忠於檢索資料之上下文，避免提煉階段的資訊扭曲，確保大型語言模型生成的回覆基於可靠來源。使用 multi-qa-MiniLM-L6-cos-v1 模型，並計算檢索資料向量與知識向量的平均內積作為事實性分數。

一致性分數（consistency）

一致性分數旨在評估生成內容是否與上游的知識邏輯連貫，避免引入不在知識當中的新內容。以 CTRL-Eval 模組作為一致性評估方法，該方法建構雙向遮罩（Masking）序列，第一種以知識加遮罩預測生成部分，第二種反之，並使用 google/pegasus-large 模型計算交叉熵損失（Cross Entropy Loss），輔以逆詞頻權重強調稀有詞，最終加權平均得出一致性分數。

切題度分數（entailment）

切題度分數旨在衡量生成回覆是否充分回應使用者查詢的意圖，避免回覆偏題。使用 multi-qa-MiniLM-L6-cos-v1 模型，並計算使用者查詢向量與回覆向量的平均內積作為切題度分數。

3.4.1.4 自我校正迴圈之閾值

自我校正迴圈中採用的三項評分標準（事實性、一致性及切題性）之閾值由本研究團隊內部進行實驗採定，使用 F1-Score 做為評估指標。

- F1-Score：在本實驗中評估回覆分數 x（事實性、一致性或切題性分數）與人工票數 y 在不同閾值下的準確性，取 F1-Score 最高者作為最佳閾值門檻。

3.4.2 使用者問卷與體驗評估設計

驗證新系統在真實互動情境中的使用體驗，本研究進行使用者測試與問卷調查。問卷分為三部分：

- 基本資料收集：包含年齡、性別與平日 Podcast 收聽習慣，用以理解不同背景對系統感受之影響。
- 系統回覆品質評分：受測者針對三項情境題中，舊系統與新系統的回覆進行比較，評估內容的相關性、完整性與準確性。
- 整體系統滿意度：評估系統的操作便利性、介面友善度、回覆清晰度及整體使用體驗。

問卷量化資料後續皆以平均數、標準差、成對樣本 t 檢定與 p 值進行統計分析，客觀驗證新系統在語意品質、操作便利性與整體滿意度等多面向之實際成效。

第四章、 實驗結果分析

4.1 動態去冗分群個數之選擇分析

本研究為選定檢索階段最適合的動態分群規模（Top-k），比較了五組不同的分群設定，分別為 2×3、3×3、5×2、4×3 與 6×2。

實驗針對五題測試問題，以五項生成品質指標進行匿名評分，包括：內容相關性、回覆完整性、事實正確性、可讀性與流暢度，以及語意一致性。透過平均值、標準差與跨題一致性三項量化指標綜合評估各分群設定的優缺點。分析結果如表 四-1 及圖 四-1：

表 四-1 不同分群設定比較表

（紅色為最終選擇之分群設定）

分群設定	2×3	3×3	5×2	4×3	6×2
平均值(MEAN)	15.08	15.92	16.16	15.36	11.20
標準差(STD)	5.57	4.00	4.63	3.84	6.27
跨題一致性(次)	1	6	5	10	3

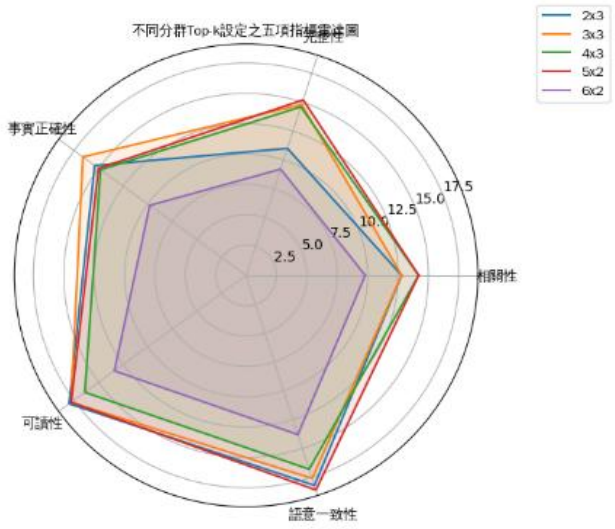

圖 四-1 平均值雷達圖

(1) 平均分數比較

在五組動態分群設定中，5×2 與 3×3 的平均分數表現最佳，相較之下，2×3 與 4×3 的平均分數則處於中間區段，而 6×2 的平均分數最低。經分析結果後得出，較小群數與較多樣本數的組合（如 5×2、3×3）能在語意聚合上提供更佳的支持，使模型能產生品質較高的回覆。

(2) 標準差比較

從數據可見，4×3 的標準差最低，表示其在不同題型間呈現最一致的表現，生成品質波動最小。而 3×3 與 5×2 也具備相對良好的穩定度，相對地，2×3 與 6×2 顯示出較大的波動性，不穩定性更高。此比較結果顯示出 4×3 在不同語意與題型下具備良好的穩健性，較不受資料分布差異影響。

(3) 跨題一致性分析

根據統計，4×3 在 25 項評分指標中有 10 次獲得最高分，為所有組別中最具優勢者。而 2×3 與 6×2 的最高分次數明顯較低，顯示其生成品質的不足。因此，4×3 在多題、多維度上的頻繁領先，反映其不僅能在少數情境中取得好成績，更能在不同語意類型下維持優良的生成品質。

綜合三項指標，4×3 雖然不是平均分數最高的組別，但在穩定性與跨題表現上呈現最完整且均衡的優勢。5×2 雖然擁有最高平均分，但其波動較大；3×3 的表現雖良好，但跨題一致性仍不及 4×3。所以，4×3 在三項評估指標中皆擁有明確優勢，能提供更可靠的生成品質，基於此，本研究決定採用 4×3 作為動態去冗分群的最佳 Top-k 設定。

4.2 加入動態去冗分群之秒數分析

本研究比較加入動態去冗分群機制前後之查詢反應時間，並以平均值（Mean）、中位數（Median）、標準差（Std）、個別題目改善率（Improvement %）及加速倍率（Speedup）作為評估指標。分析結果如表 四-2 及表 四-3：

表 四-2 加入動態去冗前後之秒數統計分析

	加入前(秒)	加入後(秒)
平均值(MEAN)	18.30	14.73
中位數(MEDIAN)	17.01	14.90
標準差(STD)	9.07	3.97

(Kang, Zhu et al. 2024)

表 四-3 加入動態去冗分群機制前後之各題執行秒數比較表

	改善率(%)	加速倍率(%)
Q1	-28.08	-22
Q2	-6.63	-6
Q3	+0.18	+0
Q4	+42.15	+73
Q5	-10.65	-10
Q6	-41.79	-30
Q7	+10.10	+11
Q8	+66.24	+196
Q9	+12.87	+15
Q10	+46.45	+87

(1) 整體效能提升

從整體統計量可觀察到：

- 平均查詢時間（Mean）由 18.30 秒降至 14.73 秒，改善幅度約 19.5%。
- 中位數（Median）由 17.01 秒降至 14.90 秒，表示排除極端值後仍呈現下降。
- 標準差（Std）由 9.07 降至 3.97，變異縮小超過一半，代表反應時間更穩定。

此結果顯示動態去冗分群能有效縮小搜尋空間，使查詢時間更加一致並減少高延遲情況的發生。

(2) 部分題改善率上升

在 10 題測試中，有 6 題呈現正向改善，顯示動態分群能在多數情況下縮短查詢時間。改善率顯著者包括：Q8：+66.24%、Q10：+46.45%、Q4：+42.15%、Q9：+12.87%、Q7：+10.10%。這些題目在加入動態分群後，查詢時間明顯變快，代表其相關資料原本比較多，或是語意比較分散。加入分群後，系統能先把不重要或重複的資料排除，讓要搜尋的資料量變少，因此整體反應時間就縮短了。

(3) 部分題加速倍率顯著提升

加速倍率結果顯示某些查詢可獲得相當顯著的效能提升，其中：Q8:+196%、Q10:+87%、Q4:+73%。此類查詢的共同特徵為語意分散或對應的向量數量較多，故動態分群能有效減少搜尋空間，使查詢時間明顯降低。

(4) 效能不提升之情況分析

部分題目（Q1、Q2、Q5、Q6）出現負改善率，其原因可能包括：原查詢所需的向量數量較少、查詢語意高度集中。總而言之，動態分群更適用於語意涵蓋較廣、向量密度較大的查詢；而在語意範圍窄或資料量小的情況下，改善幅度相對有限或可能出現額外延遲。

綜合上述結果可得知：

- 在多數查詢情境下，動態去冗分群機制能縮短查詢時間、提升整體效能並增加系統穩定性。
- 效益最顯著者多為語意較廣或所需向量庫較大的查詢，表示動態分群的優勢在於縮減搜尋空間並提升檢索效率。
- 少部分查詢因資料量小或語意集中，分群結果可能導致額外的時間負擔。

因此，動態去冗分群機制對於 Podcast 語意搜尋的效能確實具提升效果，特別適用於高維度、高冗餘之語意檢索情境。

4.3 自我校正迴圈閾值之設定分析

本實驗僅保留 3.2 動態去冗分群機制，不進行 3.3 自我校正迴圈機制，實驗準備六題問題向本系統進行詢問並重複操作五次，共產生 30 則回覆，每則回覆皆有三項評分標準之分數，團隊人員評估單一回覆是否有達到事實性、一致性及切題性，並以 F1-Score 最高者為最終閾值設定結果。實驗結果如表 四-4：事實性閾值為 0.83，一致性閾值為-1.28，切題性閾值為 0.47。

表 四-4 迴圈中三項評分閾值門檻設定分析

事實性閾值分析		一致性閾值分析		切題度閾值分析	
Threshold	**F1-Score**	**Threshold**	**F1-Score**	**Threshold**	**F1-Score**
0.80	0.8139	-1.30	0.6666	0.46	0.7586
0.81	0.8139	-1.29	0.6666	0.47	0.8
0.82	0.7905	-1.28	0.6857	0.48	0.7586
0.83	0.8372	-1.27	0.5263	0.49	0.7407
0.84	0.7766	-1.26	0.5373	0.50	0.6923
0.85	0.7455	-1.25	0.5373	0.51	0.64

4.4 回覆事實性分數分析

本研究發現動態去冗分群機制可以使得生成過程的事實性分數提高，故設計實驗透過 t 檢定評估本系統加入動態去冗分析機制前後的事實性分數差異，實驗中模型 A 為僅加入自我校正迴圈，無動態去冗機制、模型 B 為本研究最終模型：動態去冗機制與自我校正迴圈。實驗結果如圖 四-2 及表 四-5 事實性分數統計分析。

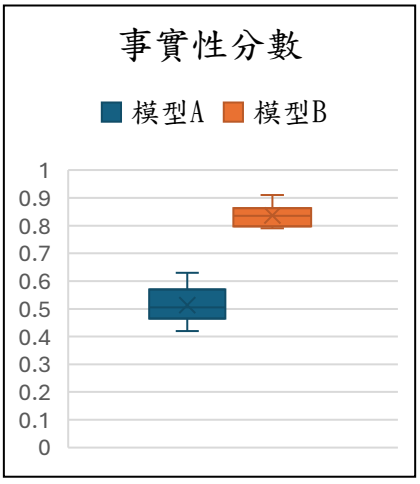

圖 四-2 事實性分數盒鬚圖

表 四-5 事實性分數統計分析

皮爾森相關係數	0.58

t 統計量	-19.66
P(T<=t) 單尾	5.27721E-09
臨界值：單尾	1.83
Cohen'd	6.17

檢定結果顯示，t 統計量為-19.66，遠低於單尾臨界值-1.83，顯示模型 B（有動態去冗分群機制）的事實性分數顯著高於模型 A（無動態去冗分群機制），單尾 P 值僅為 5.27721E-09，高度顯著拒絕了虛無假設（前後平均無差異）。本實驗也計算 Cohen's d 值，這是一種衡量兩組平均數差異的效果量指標（effect size），見公式(c)。

$$\hat{d} = \frac{|\overline{X_1} - \overline{X_2}|}{\sqrt{\frac{(n_1-1){s_1}^2+(n_2-1){s_2}^2}{n_1+n_2-2}}} \quad (c)$$

$\bar{X}$為樣本平均數，s為標準差，n為樣本數，Cohen's d 值為 6.17，大於 0.8 表示兩模型差異甚大，效果量大。

根據上述的統計量分析，證明動態去冗分群機制確實能改善本系統回覆生成的事實性，確立了該機制因檢索更精準的資料，有助於提升大型語言模型的生成品質。

4.5 問卷調查分析

本研究設計一問卷給東吳大學學生填寫，有效問卷 42 份，問卷內容分為兩大題，第一大題為系統回覆評分，目的為比較同一問題在傳統 Podcast 平台（Apple Podcast 及 Spotify）上得到的搜尋結果和本系統生成的回覆結果；第二大題為 Podcast 智慧搜尋系統滿意度調查。兩大題都以 7 分量表作為評分準則(Finstad 2010)，第一大題之評分級距:1 分為完全錯誤，與內容不符；7 分為完全正確，與理想高度一致。第二大題之評分級距:1 分為非常不同意，系統難以使用、不直覺；7 分為非常同意，整體流暢、完全符合使用直覺，捕捉細微的語意差異，突破傳統指標的局限性。

4.5.1 系統回覆評分之分析

此大題共包含三個情境問題，這些問題模擬使用者在收聽 Podcast 中可能的查詢，例如「除了韓劇，還有哪些韓綜備受討論?」，將這些情境問題輸入傳統

Podcast 平台（Apple Podcast 及 Spotify）之搜尋介面，並截圖其搜尋結果，其次，將相同問題輸入本系統中，截圖其生成之搜尋回覆，受訪者需針對兩回覆結果分別評分，本研究將評分結果進行平均數、標準差、成對樣本 t 檢定以及 p 檢定的計算比較，各題分析結果如表 四-6，

表 四-6 系統回覆評分之各項指標比較結果

題目	傳統 Podcast 平台平均數	Podcast 智慧搜尋系統平均數	傳統 Podcast 平台標準差	Podcast 智慧搜尋系統標準差	t 值	p 值	是否顯著
Q1	4.5	5.21	1.153	1.22	3.223	0.00249	顯著
Q2	3.29	5.4	1.701	1.363	6.776	3.4×10^{-8}	顯著
Q3	3.19	5.33	1.798	1.391	5.42	2.9×10^{-6}	顯著

(1) 三項情境題均達顯著差異（$p < 0.05$）：

代表智慧搜尋系統在三種不同使用情境下，整體表現皆顯著優於傳統 Podcast 平台。

(2) 本研究系統平均分數明顯較高:

所有題目中，智慧搜尋系統的回覆品質評分皆高於 Apple Podcast/ Spotify，顯示系統能提供更準確、更貼近需求的回答。

(3) 評分一致性佳:

智慧搜尋系統的標準差普遍小於傳統平台，代表使用者對其體驗的穩定度更高；傳統平台則較不一致。

(4) 傳統平台在語意理解與情境匹配上受限:

其回覆差異較大，反映關鍵字搜尋無法精準捕捉語意需求，造成使用者體感落差較大。

(5) 智慧搜尋系統展現語意檢索的核心價值，RAG 結合 LLM 能有效處理：

- 抽象語意
- 多面向問題
- 情境式查詢
- 長文本 Podcast 的跨段落資訊對齊

整體而言，本系統在搜尋品質與回答正確度上具顯著優勢，受試者一致認為智慧搜尋系統提供的內容更完整、精確、具邏輯性，也更能對應真實資訊。

4.5.2 Podcast 智慧搜尋系統滿意度分析

此部分讓使用者評估本研究之系統相較於傳統 Podcast 平台，在回覆與使用直覺性上的表現。本研究將評分結果進行平均數與標準差的計算比較，分析結果如表 四-7，

表 四-7 系統滿意度之平均數與標準差比較結果

題目	平均數	標準差
Q1	4.69	1.39
Q2	5.02	1.33
Q3	4.98	1.33
Q4	5.05	1.36
Q5	4.45	1.55

(1) 整體滿意度維持中高水準

五項題目的平均分數介於 4.45～5.05，顯示使用者整體對系統持正向評價，對搜尋品質與回覆皆有良好感受。

(2) 資訊可信度為主要改進方向

其中「系統提供之資訊可信度」平均分數略低（μ = 4.45），代表雖仍達正向評價，但使用者期望系統在來源呈現方式、證據透明度或回覆引用上更清楚。

(3) 回覆清晰度與操作直覺性受高度肯定

其餘題目平均分均接近或高於 5 分，反映使用者認為系統提供的答案具可理解性、內容完整且能有效協助解決查詢問題。

(4) 評分一致性良好

標準差偏小，顯示使用者在系統回覆的體驗差異不大，整體感受一致，系統表現穩定可靠。

(5) 智慧搜尋系統展現語意搜尋的實質價值

使用者普遍認為本系統在以下方面優於傳統平台：

- 語意理解能力佳
- 能提供脈絡化內容
- 回覆結構清楚
- 搜尋結果更貼近需求

綜合五項題目結果可知，系統在「操作便利性」、「回覆品質」、「內容相關性」

等面向皆獲得中高程度肯定，實證語意型 Podcast 智慧搜尋系統具備實際應用價值。

第五章、 結論與未來研究

5.1 結論

本研究的主要貢獻如下：

(1) 提出一套創新的系統架構，將大型語言模型與向量資料庫整合，有效提升語意理解能力與 Podcast 內容的搜尋與檢索準確性。
(2) 完成一套可實際運作的 Podcast 智慧搜尋系統，具備良好的應用與擴充潛力，可進一步延伸至跨主題、多語言或更大規模的音訊內容分析情境。
(3) 設計並實作「動態去冗分群機制」，透過分群檢索最精準的資料，實驗結果顯示，此機制有效提升模型輸出的事實性，其 Cohen’s d 效果量達 6.17，並經 t 檢定確認達顯著水準（$p < 0.001$）。此外，動態分群亦成功減少搜尋時間，使平均查詢時間改善 19.5%，中位數改善 12.4%，而標準差則改善 56.2%，顯示此機制不僅能加速搜尋效果，也強化輸出品質，具高度實務應用價值，並可作為其他 RAG 架構之參考依據。
(4) 建立自我校正迴圈，透過「事實性、一致性、切題性」三項評分進行前置檢核，提升 LLM 回覆的可靠度並降低生成內容的幻覺風險。
(5) 透過使用者問卷驗證本系統優於傳統 Podcast 平台，結果顯示在回覆品質、內容完整性、情境理解與操作直覺性等方面皆具明顯優勢，本系統在三項情境題均達顯著水準($p < 0.05$)，在滿意度中總平均為 4.84，證實本系統能有效改善 Podcast 的搜尋體驗。

本研究成果亦成功呼應研究目的：一、動態去冗分群機制使系統能依查詢自動縮小搜尋範圍，明顯縮短查詢時間並提升穩定度，達成「提升效能、降低冗餘」的目標；二、在搜尋結果呈現方面，前端介面的設計讓使用者能清楚理解內容，增加「使用者的體驗感受」。語意理解層面方面，透過 LLM 與向量資料庫，使系統得以「掌握自然語言查詢背後的語意脈絡」，改善傳統關鍵字式搜尋的語意侷限；加上自我校正機制有效提升回覆的事實性與一致性，使最終回答更為可信，符合「強化內容可靠度」的研究目的。

整體而言，本研究提出的 Podcast 語意搜尋系統在技術與使用者體驗上皆展現良好成效，不僅達成研究目的，也為未來多主題、多語言與大規模音訊搜尋提供可行的架構與方向。

5.2 未來研究

本研究雖已完成一套具語意理解能力的 Podcast 搜尋系統，但在更大規模、多語言與更高精準度的應用情境中仍有提升空間，未來研究可從以下方向持續拓展：

(1) 縮短搜尋時間

未來可探索向量壓縮（Vector Quantization）、近似最近鄰（ANN）索引等方法，以在保留語意品質的前提下進一步縮短查詢延遲。同時，也可將搜尋流程改為並行式或分散式架構，以提升大型資料庫下的整體反應速度。

(2) 提升搜尋精準度

若引入更精細的排序與重排序技術，例如使用 reranking 模型、長文本理解模型等，用來量測語意相似度，提升最終檢索結果的準確度，使系統可在複雜查詢情境中提供更貼近語意需求的搜尋結果。

(3) 多主題搜尋

目前系統以戲劇類 Podcast 為主，未來可擴展至多主題語料。可研究跨主題的語意邊界、主題混合時的群集行為，以及如何透過層級分群有效提升跨主題搜尋的精準度，使系統能在更複雜的語意空間中保持穩定表現。

(4) 多語言系統

未來可將系統擴展至英文、日文、韓文等語言，建立跨語言語意搜尋架構，結合多語言嵌入模型，讓使用者能以單一語言查詢不同語言的 Podcast 內容，提升系統的跨文化應用價值。

(5) 降低 LLM 幻覺現象

雖本研究已導入自我校正機制，但仍可進一步使用外部驗證器、知識庫比對等方法，降低模型在未知領域的推測性回答。若能結合強化學習（RLHF/RLAIF），亦可提升系統在事實性上的控制能力。

(6) 強化非監督式學習分群

未來可採用更進階的非監督式分群方法，如密度式分群或深度嵌入式分群，以突破 K-Means 在群形狀單一與需預設 k 值的限制。運用更彈性的分群策略，能使系統更精準反映語意空間的自然結構，使動態分群在不同查詢情境下具更佳的適應度與穩定性。

參考文獻

ARADHYA, K. and T. DIVYA (2024). "VECTOR DATABASES." OPEN ACCESS RESEARCH JOURNAL OF ENGINEERING AND TECHNOLOGY Учредители: Open Access Research Journals Publication **7**(1): 096-104.

Besser, J., M. Larson and K. Hofmann (2010). "Podcast search: User goals and retrieval technologies." Online information review **34**(3): 395-419.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry and A. Askell (2020). "Language models are few-shot learners." Advances in neural information processing systems **33**: 1877-1901.

Carterette, B., R. Jones, G. F. Jones, M. Eskevich, S. Reddy, A. Clifton, Y. Yu, J. Karlgren and I. Soboroff (2021) "Podcast Metadata and Content: Episode Relevance andAttractiveness in Ad Hoc Search." arXiv:2108.11460 DOI: 10.48550/arXiv.2108.11460.

Clifton, A., S. Reddy, Y. Yu, A. Pappu, R. Rezapour, H. Bonab, M. Eskevich, G. Jones, J. Karlgren, B. Carterette and R. Jones (2020). 100,000 Podcasts: A Spoken English Document Corpus, Barcelona, Spain (Online), International Committee on Computational Linguistics.

Devlin, J., M.-W. Chang, K. Lee and K. Toutanova (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers).

Eskevich, M. (2014). Towards effective retrieval of spontaneous conversational spoken content, Dublin City University.

Finstad, K. (2010). "Response interpolation and scale sensitivity: Evidence against 5-point scales." Journal of usability studies **5**(3): 104-110.

Fuller, M., M. Tsagkias, E. Newman, J. Besser, M. Larson, G. J. Jones and M. de Rijke (2008). "Using term clouds to represent segment-level semantic content of podcasts."

Goto, M. and J. Ogata (2011). PodCastle: Recent Advances of a Spoken Document Retrieval Service Improved by Anonymous User Contributions. INTERSPEECH.

Ji, Z., T. Yu, Y. Xu, N. Lee, E. Ishii and P. Fung (2023). "Towards mitigating hallucination in large language models via self-reflection." arXiv preprint arXiv:2310.06271.

Jing, E., K. Schneck, D. Egan and S. A. Waterman (2021). "Identifying introductions in podcast episodes from automatically generated transcripts." arXiv preprint arXiv:2110.07096.

Jing, Z., Y. Su and Y. Han (2025). When large language models meet vector databases: A survey. 2025 Conference on Artificial Intelligence x Multimedia (AIxMM), IEEE.

Kang, H., Y. Zhu, Y. Zhong and K. Wang (2024). "SAKR: Enhancing Retrieval-Augmented Generation via Streaming Algorithm and K-Means Clustering." arXiv preprint arXiv:2407.21300.

Khattab, O. and M. Zaharia (2020). Colbert: Efficient and effective passage search via contextualized late interaction over bert. Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval.

Lee, L.-s., J. Glass, H.-y. Lee and C.-a. Chan (2015). "Spoken content retrieval—beyond cascading speech recognition with text retrieval." IEEE/ACM Transactions on Audio, Speech, and Language Processing **23**(9): 1389-1420.

Lugosch, L., M. Ravanelli, P. Ignoto, V. S. Tomar and Y. Bengio (2019). "Speech model pre-training for end-to-end spoken language understanding." arXiv preprint arXiv:1904.03670.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer." Journal of machine learning research **21**(140): 1-67.

Reimers, N. and I. Gurevych (2019). "Sentence-bert: Sentence embeddings using siamese bert-networks." arXiv preprint arXiv:1908.10084.

Senay, G., G. Linares and B. Lecouteux (2011). A segment-level confidence measure for spoken document retrieval. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE.

Spina, D., J. R. Trippas, L. Cavedon and M. Sanderson (2017). "Extracting audio summaries to support effective spoken document search." Journal of the Association for Information Science and Technology **68**(9): 2101-2115.

Thakur, N., K. Kumar, V. K. Thakur, S. Soni, A. Kumar and S. S. Samant (2022). "Antibacterial and photocatalytic activity of undoped and (Ag, Fe) co-doped CuO nanoparticles via microwave-assisted method." Nanofabrication **7**: 1-27.

Tsagkias, M., M. Larson and M. de Rijke (2008). Term clouds as surrogates for user generated speech. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.

Wu, L., Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu and Q. Liu (2024). "A survey on large language models for recommendation." World Wide Web **27**(5): 60.