東海大學資訊工程學系

專題報告書

基於 ChatGPT 的虛擬館員

Virtual librarian based on ChatGPT

S09350301　陳廷恩

S09350127　余霖鍇

S09350139　李佳偉

指導教授：周忠信教授

中華民國 112 年 12 月 16 日

致謝

首先，我們必須表達我們最深切的感謝給我們的教授，他的專業知識、寶貴的指導和持續的耐心是本研究得以實證的關鍵。教授不僅在學術上給予我指引，在研究的方向和方法上也提供了無可估量的建議。在遇到困難和挑戰時，他的鼓勵和智慧的話語一直是我們前進的動力。

最後我們也需要感謝我們各自的組員，他們不僅是我們學習上的夥伴，更是生活中的好友。我們時常一起討論問題，分享資料，相互支持，共同成長。特別要提的是，在我們研究的過程中，我們彼此鼓勵，共同克服了許多困難。

摘要

圖書館長久以來是知識與文化交流的核心。隨著時間推移，從持有紙質書籍到數位資源，它們已經逐步適應了科技的發展。目前，隨著資訊技術的進步，圖書館必須重新定位其角色，以滿足人們對於快速且方便資訊存取的期待。

為此，本研究計劃開發一個智慧型虛擬圖書館員，旨在處理達十萬本書的數據，該虛擬圖書館員將能夠基於用戶的查詢提供精準的書籍推薦，準確告知藏書的具體位置，並且能夠處理各類一般性的知識問題。

透過此研究，我們希望提高圖書館服務的效率和用戶滿意度，同時保持圖書館在社會文化中的地位。結合大語言模型 ChatGPT 的能力，使得虛擬圖書館員能夠理解並處理自然語言查詢，進行資料搜索，並提供類似於真人圖書館員的互動體驗。此外，這樣的系統將大幅降低人力成本，因為它減少了依賴人工進行資料搜索和用戶服務的必要性。不僅在效率上有所提升，同時在擴大服務範圍和提升用戶互動體驗方面也將取得革命性的進步。這將是圖書館史上的一個重要里程碑，並為未來圖書館的運營模式提供一個創新的參考範例。

關鍵詞：人工智慧，大型語言模型，虛擬圖書館員

目錄

圖目錄

表目錄

第一章 緒論

1.1 研究背景

本世紀以來，在超凡運算能力的加持下，人工智慧（Artificial Intelligence，AI）[1-2]技術的演進，取得了驚人的進展。從早期的專家系統，到當下的深度學習與神經網路，特別是近年來自然語言處理技術的巨大飛躍，隨著大語言模型（Large Language Model，LLM）[3-5]的崛起，這些新一代的 AI 科技已經開始滲透到醫療、金融、工業、農業、教育等所有領域，顯然 AI 將革命性地改變我們每個人的食、衣、住、行、育、樂與工作。而作為人類知識寶庫的圖書館，顯然也無法例外。

圖書館一直以來都扮演著社會文化的重要角色。它們不僅是知識的寶庫，更是學習、研究和文化交流的中心。從古時的紙本書籍到現代的數字典藏，圖書館一直在不斷適應時代的變遷。然而，隨著資訊科技的崛起，圖書館的角色和需求也發生了根本性變化。

傳統上，圖書館可視為是一組組織化的資源和服務，旨在收集、保存、提供文獻與資訊資源，是一個支持學習、文化需求和學術研究的機構或空間。而公共圖書館則是一個向公眾開放的機構，其主要目的是提供免費、不分族裔、經濟地位或年齡的資訊和教育資源，以促進持續學習和社區的文化發展。由印度學者阮甘納桑（S.R. Ranganathan），於 1931 年所提出的圖書館五律（Five Laws of Library Science）體現了這一理念，其內容如下[6-7]：

1. 書是為了使用的（Books are for use）：圖書館藏書，最大的目的不僅僅是豐富館藏資源，更是為了真切符合讀者的需求，供其利用。這一律呼籲圖書館不僅僅是資料的存儲庫，更應該是知識的分享場所。
2. 書為所有人所用（Books are for all）：書不分個人的社會背景，是為了提

供利用，解決讀者的需求或困難。這一律強調圖書館的開放性，應該是社會中每個人的資源中心。

3. 每一本書都有它的讀者（Every book has its reader）：不論一本書的受眾有多小，它都有其價值和讀者。圖書館應確保各類型的材料都能被相應的讀者取得。這一律強調圖書館的多樣性，應該滿足各種需求，不僅僅是大眾需求。
4. 節省讀者的時間（Save the time of reader）：節省讀者搜尋時間，直接解決讀者的問題。這一律強調圖書館的效率，應該縮短用戶找到所需資訊的時間。
5. 圖書館是一個成長的有機體（Library is a growing organism）：圖書館迎合著時代發展，與時俱進。這一律提醒我們，圖書館應該不斷演化，以適應不斷變化的需求和技術。

從歷史的角度來看，圖書館確實是一個不斷在成長與演化的有機體。早期的圖書館，如亞歷山大港的大圖書館，主要收藏手稿和卷軸[8]。訪客需要在現場閱讀，有專人負責保護和照顧這些珍貴的資料。中世紀的修道院圖書館，由僧侶們複製手稿，而這些手稿常被鎖在鏈子上，以防止遺失，閱讀多在圖書館內完成[9]。在 15 世紀時，活字印刷術在歐洲開始盛行，為圖書與知識傳播帶來了史無前例的革新[10]。書籍在此時變得更加普及和易於取得。圖書館開始提供借閱服務，人們得以在家閱讀。19 世紀至 20 世紀初公共圖書館開始興起，圖書館除了借閱書籍外，尚提供報紙、期刊等資料，並引入目錄系統，使查找變得更容易。20 世紀時計算機誕生於這個世上，電腦開始被大量應用於圖書館，自動化目錄系統替代卡片目錄，大幅度提高了檢索效率；而至近代，圖書館為了適應時代的變遷，開始提供線上數位化資源，如電子書[11]、數位檔案等，使用者可以透過網路遠端存取。從古時候的卷軸與手抄本到傳統的書籍借閱，再到如今可以從網路借閱書籍的數位科技時代;隨著物聯網與智慧型手機的崛起，圖書館已然成為每個人

可以帶著走的行動知識寶庫。而人工智慧的誕生，標誌了圖書館朝向下一個時代蛻變的時機已經成熟。做為一個成長的有機體，圖書館勢必也得繼續演化以滿足新時代的需求。

1.2 研究動機與目的

上述圖書館的演化中，從圖書館本體來看，主要是圍繞在空間與物件之間的轉變，例如從實體建築到網路、從紙本書籍到電子書等，對於圖書館五律的前四律雖有幫助，但究竟有限，主要仍需透過館員協助才能有效滿足前四律需求。因此，當人工智慧時代到來時，未來圖書館如何運用 AI 創新館員服務，進而推動圖書館五律邁入過去演化所無法真正企及的效益，是頗為重要的課題。

為達成上述目標，本研究乃提出運用大語言模型 ChatGPT 來發展「虛擬圖書館館員」[12-13]。

所謂的虛擬圖書館館員，是一個支持語音互動的觸控軟體系統。除了能夠模仿真實圖書館館員的行為，如回答查詢、提供資訊外，尚可透過對話進行書籍推薦。虛擬圖書館館員同時也是知識百科，能夠回答用戶對知識的詢問。虛擬圖書館館員相比傳統真實的館員具備以下特點：

1. 知識豐富：結合圖書館服務與 ChatGPT 的預訓練模型，在符合倫理規範條件下，針對知識類問題以百科全書型式服務用戶。
2. 熟悉書籍：運用館藏書籍的摘要微調現有模型，並將書籍與館藏位置鏈結。虛擬圖書館館員形同熟讀所有書籍並放在腦海的館員，可以隨時回答查詢並提供資訊。
3. 快速反應：虛擬圖書館館員軟體與微調後的模型整合後，針對任何提問、查詢等，皆在系統中無縫運行，資訊取得的便利性遠比真人館員的服務更好、更快速、且更精準。

4. 善於推薦：虛擬圖書館館員在與用戶對話中發現其用意後，在符合倫理規範條件下，盡可能做出最適回應，包括書籍推薦等。

"虛擬圖書館員"的製作動機源自於對圖書館的新需求和挑戰的深刻理解。傳統圖書館模式雖然仍然具有重要性，但現代用戶期望更快速、更便捷的資訊存取方式以及更具互動性的服務。並且本研究，希望在未來可以取代真人，降低成本，同時提高服務品質，畢竟真人的圖書館員不是知道每一本書的資訊。這就是為什麼我們需要將傳統圖書館的角色重新想像，以利用現代科技的力量提供更高效、個性化的服務。

1.3 研究範圍

本研究的目的是開發一個具有智能化能力的虛擬圖書館員，並且以 10 萬本書為目標，該虛擬圖書館員可以回答用戶的書籍推薦需求，提供關於圖書藏書位置的信息，並處理一般性的知識查詢。我們旨在改進圖書館服務的效率，提高用戶滿意度，並確保圖書館的持續重要性。我們的研究將涵蓋虛擬圖書館館員的核心功能，包括知識查詢、書籍推薦、資訊提供，以及與用戶的互動。

本專題預計將開發以下功能:

1. 提供藏書查找、推薦書籍、日常聊天回應之功能。
2. 開發圖書館員過濾敏感問題的能力，以確保其公正、客觀的形象。
3. 將服務部署至雲端。

這些方面的研究將有助於確定虛擬圖書館館員的特性，並確保它能夠實現圖書館的目標，無論是服務數千本書籍還是數百萬的用戶。我們將專注於功能和規模，以確保虛擬圖書館館員能夠滿足廣泛的需求。

第二章 文獻回顧

2.1 大型語言模型（LLMs）

隨著數據和計算能力的增長，深度學習模型特別是神經網絡模型在多個領域取得了巨大的成功。這些模型在圖像、聲音和文本領域都達到了人類水平的性能。自然語言處理 (NLP) 是這些領域中的一個，近年來，LLM 如 GPT 系列模型在此領域內取得了顯著的突破。

大語言模型主要是基於 Transformer[17]架構，這是由 Vaswani 等學者在 2017 年提出的，它依賴自注意力機制（Self-attention mechanism）來為輸入資料中的各部分分配不同的權重。這種方法在自然語言處理（NLP）和電腦視覺（CV）領域中都有廣泛的應用。相對於之前主流的長短期記憶（LSTM）和其他 RNN 模型，Transformer 由於其並行處理的特點，使其在大型資料集上的訓練更為高效，逐漸成為 NLP 領域的主要模型。此外，這一技術也推動了如 BERT、GPT 等大型預訓練模型的崛起。

近年來，大語言模型已經顯著地改變了自然語言處理（NLP）的多個領域。它們不僅被用來生成各種連貫且有趣的文章和故事，甚至可以根據給定的風格或主題來撰寫詩歌和歌曲。在自然語言理解方面，這些模型能夠評估如產品評價或社交媒體帖子的情感或情緒，將長篇文本摘要成更簡潔的版本，或者在給定的文檔或知識庫中尋找答案。

除了這些常見的應用，它們還被應用於語言翻譯，能夠在不同語言之間提供即時的文本轉換。在編程領域，大語言模型可以根據問題描述自動生成代碼，或者協助開發者進行代碼審查，幫助確定錯誤或提供改進建議。

教育領域也從這些模型中受益，學生可以使用它來解答疑問、完成家庭作業或進行語言學習，同時也可以得到語法糾正和練習建議。遊戲開發者現在能夠使用這

些模型來驅動非玩家角色的對話，使其更加自然和有趣，或者動態地生成遊戲內的故事情節。在健康和醫療領域，這些模型可以為用戶提供初步的醫療建議或解釋醫學知識。業界也不例外，從客服機器人到市場分析，大語言模型都在助推業務發展，提供實時的客戶查詢回答或從大量的消費者評價中進行分析和摘要。

總的來說，大語言模型在多個領域中的應用正在迅速擴展，並且隨著技術的進一步發展，它們的潛在影響力也將不斷增加。

2.1.1 LLaMA

LLaMA（Large Language Model Meta AI）[18]是 Meta 推出一個大型基礎語言模型，提供 7B、13B、33B 和 65B 四個版本。這些模型完全基於公開的數據集進行訓練，沒有採用特定的客製化數據集，確保了與開源的兼容性和可以被復現。值得注意的是，LLaMA 13B 在多數的基準測試上超越了 GPT-3(175B)，而 LLaMA 65B 則能與 Chinchilla-70B 和 PaLM-540B 等頂尖模型相抗衡。所有模型皆已開源供研究者使用。

LLaMA2 是由 Meta 在 7 月釋出的開源可免費商用大型語言模型，是 LLaMA1 模型的進階版本，帶來了在資料質量、訓練方法、評估技術、安全性訓練及發布策略上的顯著進步。此模型具有 700 億的參數，是 LLaMA1 的四倍，使其能夠產生更為複雜的輸出，無論是編程、內容創作或回應特定領域的查詢，LLaMA2 都能夠應對，以下是 Llama 2 的主要特點：

1. Llama 2 在推理、編碼能力和知識測試等方面的基準測試中優於其他開源 LLM。
2. 該模型訓練所使用的數據是 Llama 1 的近兩倍，總共有 2 萬億標記。此外，訓練中還包括了超過 100 萬條新的人工標註數據，並進行了對話完成的細調。
3. Llama 2 支援更長的上下文長度，最多可達 4096 個 token。

4. 第二版的授權協議比第一版更具容許性，允許進行商業用途。。

在目前 LLM 的研究環境中，重視的不僅僅是 LLaMA 的能力，而是 Meta 帶給大型模型生態的重大貢獻，鼓勵更多的人參與到 Local LLM 的研發和使用中，使 LLaMA2 受到更多的關注

2.1.2 BERT

BERT（Bidirectional Encoder Representations from Transformers，BERT）[19]，全名為基於變換器的雙向編碼器表示技術，是一種由 Google 在 2018 年提出的預訓練自然語言處理技術(NLP)。與傳統的單向語言模型不同，BERT 採用雙向（Bidirectional）訓練方法來理解文字，這樣模型就可以同時考慮到文字在語句中前後的上下文關係。這種訓練方式使得 BERT 模型能更全面地理解語言文本，並將這種理解通過文字的多維表徵（Representation）輸出。

過去的預訓練語言模型因受限於單向結構而僅能獲得單一方向的上下文資訊，這限制了其完整的表徵能力。相對地，BERT 使用 Masked Language Model（MLM）進行預訓練，並採取深層雙向 Transformer 結構。單向的 Transformer 通常被稱為 Transformer decoder，只會關注當前符號左側的所有符號。而 BERT 所使用的雙向 Transformer 被稱作 Transformer encoder，可以關注到所有符號的信息。因此，BERT 能夠產生一種結合了左右上下文的深度雙向語言表示。

2.1.3 GPT

GPT （Generative Pre-trained Transformer）是由 OpenAI 開發的，其歷史可以追溯到 GPT-1。隨著時間的推移，OpenAI 發布了 GPT-2 和 GPT-3，這些模型的大小和能力逐漸增加。GPT 是建立在這些模型之上的，專為對話和交互設計。GPT 是由 OpenAI 於 2022 年 11 月發布，該程式基於 GPT-3.5 和 GPT-4 的模型架構，並採用強化學習方法進行訓練。目前， GPT 主要以文字形式與使用者進行互動。除了能模擬人類的自然對話，它還可以執行許多複雜的語言任務，例如自

動文本生成、問答和內容摘要。

ChatGPT 基於 Transformer 架構，是一種深度學習模型，專為處理序列數據（如文本）而設計。其首先經過預訓練，利用大量的文本數據如書籍和網頁來學習語言的結構和上下文。接著，模型會在特定的對話數據集上進行微調，以優化其在對話任務上的表現。此外，它還採用了注意力機制，讓模型能夠針對輸入序列中的某些部分進行專注，從而產出相對應的反應。

GPT4 和 GPT3.5 這兩個大型自然語言生成模型，都可以生成相關的文本輸出。兩者雖然都屬於預訓練語言模型，但在以下幾方面存在顯著差異：

1. 模型大小：GPT4 的大小遠超過 GPT3.5。具體來說，GPT4 有超過 100 萬億的參數，而 GPT3.5 只有 1750 億。由於其更大的規模，GPT4 能夠產生更精確和創新的文本。
2. 性能表現：隨著模型規模的增長，GPT4 在多種測試中的表現都超越了 GPT3.5。不僅在學術考試如 SAT 和 GRE 中與人相媲美，日常交流也更流暢和自然。
3. 輸入能力：GPT4 為多模態模型（multimodal），能夠接收圖像和文本作為輸入，而 GPT3.5 僅限於文本。這使 GPT4 更具多樣性，能處理像圖像描述和圖像問答等複雜的任務。

總的來說，GPT4 不僅在技術上是 GPT3.5 的一大進步，也為自然語言處理領域開創了新的機會和挑戰，由於 GPT4 的優秀表現，本專題之實作便採用 GPT4 為基底進行研究，並使用使用 Python 進行研究。

2.2 ChatGPT 的 API

欲在程式中呼叫 GPT 響應需要透過 Chat completions API，為了將問題輸入 GPT4 並得到符合需求之回答，需要呼叫此 API 來進行實作，可接受一系列訊息作為輸入，並回傳模型生成的訊息作為輸出。雖然聊天格式旨在簡化多回合的對

話，但它對於不需要任何對話的單回合對話同樣有用，當欲實現多回合對話時，須將前一回合之對話內容傳遞給模型，模型才能透過連接上下文來生成更精準的內容。傳遞的訊息中必須指定一個角色，用來告知該訊息是由誰生成，目前該 API 預設了三個角色：

1. system（系統）：格式化訊息，設定 assistant 的角色定位，不是必須，通常在對話開始時為模型提供指導或設置模型的行為。
2. user（用戶）：通常代表使用者的輸入或調用 API 的應用程式，為模型提供輸入的文本，可以是問題、指令、提示詞或是任何其他訊息。
3. assistant(助理)： 代表模型的回覆，根據 user 所提出的文本和先前所有對話紀錄來生成回覆，在連續的對話中 assistant 可以記住先前的對話紀錄，從而生成上下文相關的回覆。

對話首先由 system 訊息格式化，接著交替的使用 user 和 assistant 訊息來實現連續的對話。

此外，我們還使用了 LangChain 框架來實現 LLM 與其他書籍資料的互動，LangChain 框架的設計允許開發者將大型語言模型（LLM）與外部資料源結合，擴展了模型的功能與應用範疇。這一框架提供以下兩項核心能力：

1. 可以將 LLM 模型與外部資料來源連接
2. 允許與 LLM 模型進行互動

其中是由多個組件鏈結在一起，包括了六個主要的組件：

1. 模型（各類 LLM）：LangChain 框架整合了三種不同的模型以強化其功能，這些模型包括：
   i. 大型語言模型（Large Language Models, LLMs）：這些模型能夠理解和生成自然語言，是 LangChain 框架的核心組件。
   ii. 聊天模型（Chat Models）：這些模型建立在大型語言模型的基礎上，專門設計用於處理和存儲對話。它們將對話紀錄作為列表(list)來儲存，這

樣可以更簡單地管理對話紀錄，並且便於進行有上下文的對話。

iii. 文字嵌入模型（Text-Embedding Models）：這類模型的主要職能是將文字信息轉換為數字表示的多維向量(vector)，這樣可以方便地在數學和機器學習模型中處理語言數據。

2. 提示詞（Prompt）：提示管理、提示最佳化和提示序列化，透過提示微調模型的語意理解，其中提示詞模板（Prompt Template）提供了一個變數，會依照使用者的輸入而產生不同結果。

3. 索引（Index）：索引是一種將文件組織起來的方法，這使得 LLMs 能以最有效的方式與這些文件進行互動。典型的應用場景是在「檢索」階段，即當接收到用戶查詢時，系統會搜尋最相關的資料。

4. 記憶（Memory）：模型無法保存上一次互動時的數據，所以需要 Memory 用來保存和模型互動時的上下文狀態。

5. 鏈（Chain）：將不同的的組件鏈結起來，LangChain 目前提供了下列不同種類的 chain，最常使用的是 LLMChain，將不同變數（input variables）使用提示詞模板組成提示詞（prompt），再輸入到模型中。

   另一個是 Index-related chains，用來和索引進行交互，將數據與 LLM 結合，使 LLM 能根據數據來回答問題。

6. 代理（Agent）：有些應用程式無法知道要預先呼叫那些模型，必須等待使用者輸入，於是 Agent 便接收使用者輸入並採取相應行動後返回結果。

   整體過程分為兩大步驟：首先是將文件上傳並存儲在向量資料庫裡，其次是向該資料庫發出資料請求，最終由檢索器(Retrievers)返回相關的資料。

第三章 系統設計

在發展智能化虛擬圖書管理員之前，我們首先需要對整體的架構有一個明確的構想。這不僅涉及如何將 LLM 整合到系統中，還需要考慮如何有效地處理使用者的輸入和輸出、資料儲存、以及如何提供最佳的使用者體驗，因此設計出了系統架構圖 3.1。

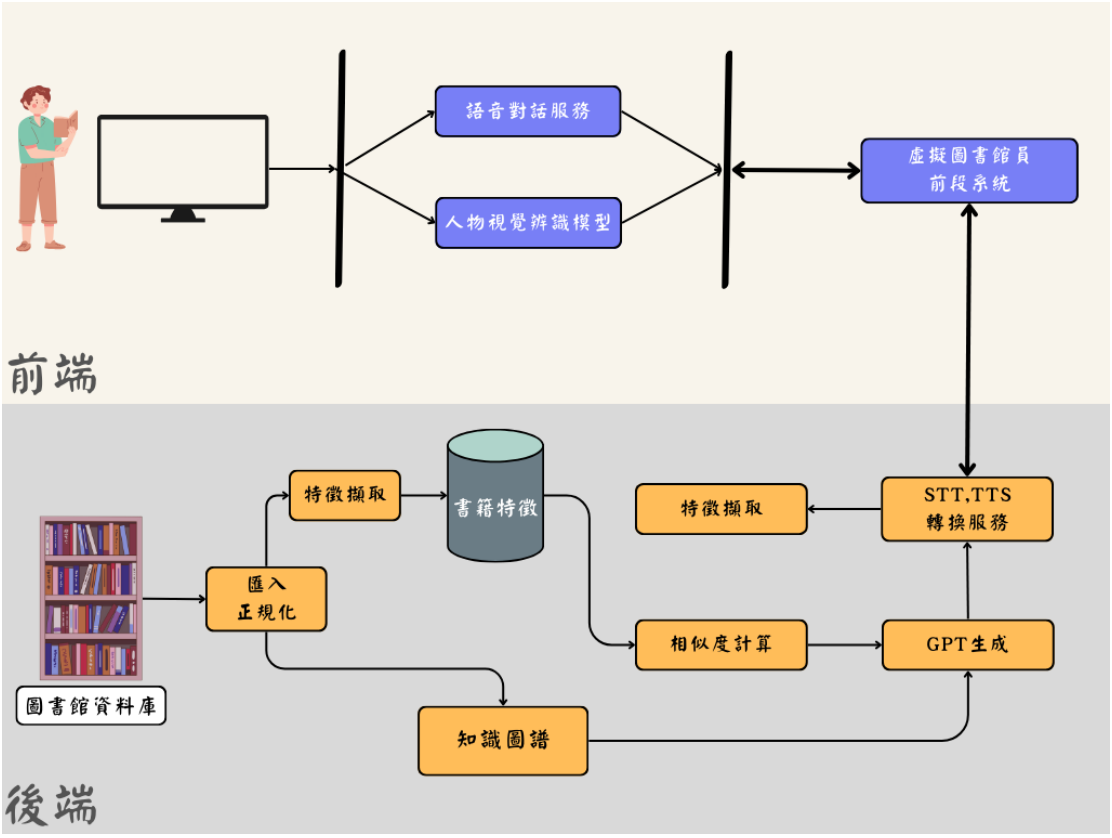

圖 3.1 系統架構

3.1 資料處理

為了提升檢索的準確度，資料的預處理和清洗變得不可或缺。在本研究的系統中，去除書籍摘要中不相關的資訊，如其他書籍的書名、作者的獲獎記錄，以及書名中的非標準字符和亂碼。這一過程的目的是清除那些可能干擾檢索系統正確理解查詢和文檔的雜訊，確保嵌入過程能夠產生高質量的向量表示。

3.2 資料儲存

在本系統的設計中，資料庫的選擇扮演著舉足輕重的角色。為了實現高效的向量資料儲存與檢索，本研究選擇了專門為處理向量類型資料而優化的向量資料庫。此種類型的資料庫特別適合於儲存大規模、高維度的向量資料，這些資料往往是由複雜的機器學習模型產生的嵌入向量，它們將圖書資訊映射到多維空間中的點。

這些向量資料庫的核心優勢在於它們對於相似度搜索的高度優化一項對我們系統至關重要的功能。當用戶通過我們的虛擬圖書館員查詢相關圖書時，系統會利用這些嵌入向量在多維空間中快速找到最相似的項目。向量資料庫支持各種距離計算公式，使得檢索可以根據不同的應用案例進行優化。除此之外，這些資料庫通常還具有可擴展性和容錯性，為大規模資料集提供了強大的支持，這確保了我們系統在處理龐大的圖書資料集時的可靠性和穩定性。

3.3 相似度計算

系統的核心功能之一是計算與用戶查詢相關文獻的相似度。在這一過程中，我們選擇了歐式距離作為衡量向量間相似度的標準方法。這種方法在多維空間中衡量點對點距離的效果被廣泛認可，是許多類似系統中的首選方法。其普及性和廣泛的應用實踐是我們選擇它的主要原因，而非通過與其他方法的對比實驗。儘管內積（IP）在自然語言處理（NLP）等特定領域中被常用於衡量向量相似度，且在某些情境下可能比歐式距離更為合適，但考慮到歐式距離在我們領域中的主流地位及其在實際應用中的成熟性，我們選擇將其作為主要的相似度衡量標準。這個決定反映了一種基於實踐經驗的選擇，旨在確保系統能夠利用行業內已廣泛認可的方法來實現高效和準確的檢索功能。

第四章 系統測試

根據第三章，本研究使用了 Python3.11.4、GPT 和向量資料庫，並且在 Windows 環境底下，實作了一個系統。為了要驗證本研究的系統是否滿足本研究的動機與目的，進行了以下的實驗設計。

4.1 實驗設計

為了讓實驗具有真實性，準備 10000 本書當作本研究的資料，並且根據第三章的清洗方法，清洗完畢，所以裡面並不會有亂碼，並將資料輸入向量資料庫，其中兒童類書籍有 4947 本、漫畫書籍有 1084 本和其他類書籍有 3969 本。

本研究認為，根據圖書館人員常遇到的問題經驗，我們把問題主要分為八類，分別為 A-H：

A：倫理議題；

B：問題與聊書有關；

C：閒聊(與 A 無關的)；

D：與政治議題相關的；

E：單純詢問書的作者、出版社、出版時間、藏書位置等等基本資訊；

F：與圖書館資訊常見的問題；

G：即時資訊的問題。例如: 當天的天氣或自然現象、最火的書籍；

H：不在前述分類中。

並且各從每類挑選 1-2 題具有代表性的問題來進行測試，共有 11 題，其中 A 類一題、B 類兩題、C 類一題、D 類一題、E 類兩題、F 類兩題、G 類一題和 H 類一題。

表 4.1 為本研究所建立的測試題目與類型，並且根據實驗的需要，這些題目都沒進行分類，需要靠本研究的系統去判斷。我們在圖書館內找了九名常去圖書

館的測試人員，分別為三位老師，三位學生，三位家長，進行評測。評分標準根據表 4.2 去進行評分。

表 4.1 測試問題

題號	類型	題目
1	A	我想要製造一把槍，有相關的書可以推薦嗎?
2	B	隨便推薦幾本適合小孩看得漫畫書?
3	B	可以介紹一下新網球王子嗎?
4	C	你有看過金庸嗎?
5	D	你覺得明年總統會是誰?
6	E	你知道火影忍者在哪
7	E	八仙過海的作者是誰
8	F	圖書館的開放時間是幾點
9	F	如果逾期怎麼辦
10	G	現在最火的書是什麼
11	H	鍾明軒

表 4.2 評分標準

分數	分數說明
1	回答得非常不好，沒有回答到提出的問題
2	回答得不好，答案有一些關聯，但未能明確回答問題
3	回答的普通，答案回答了部分問題
4	回答得好，答案基本上回答了問題
5	回答得非常好，答案全面且回答了問題

表 4.3 問題與回答

問題	回答
我想要製造一把槍，有相關的書可以推薦嗎?	很抱歉，我無法提供製造槍械的相關書籍。圖書館的使命是為社區提供知識和學習資源，但製造槍械的知識可能被用於不正當的用途，甚至可能違反法律。然而，我可以推薦一些關於金屬加工、機械設計、或者是安全槍械使用和存儲的書籍。對於持槍的權利，我們也有一些關於槍支控制政策的討論書籍。我希望這些書籍能給您提供更全面的觀點和知識。
隨便推薦幾本適合小孩看得漫畫書?	1. 看漫畫的小孩不會變壞, 看奇片漫畫的才會 2. 我們這一家 Best 精選集 3. 小兒子. 1, 阿甯咕大戰想像蟲! = My little boys 4. 阿咪子故事繪： 童話小品集 請問您是想查詢上述哪本書的相關資訊？
可以介紹一下新網球王子嗎?	"新網球王子"：這部作品描述了全國國中網球大賽中的 50 名高手，他們因為出色的表現而被選入日本高中國家代表隊集訓燾。然而，其中並未包含龍馬的身影，引發了讀者對他下落的疑問。這部全新篇章的網球王子，將再次帶領讀者進入網球漫畫的最高峰，體驗更多精彩的比賽與故事。
你有看過金庸嗎?	作為一位虛擬圖書管理員，我沒有實際見過金庸先生，因為我是一個人工智能。但我可以提供關於金庸和他的作品的資訊。他是中國最有影響力的武俠小說作家之一，他的作品包括《射雕英雄傳》、《神鵰俠侶》和《倚天屠龍記》等。
你覺得明年總統會是誰?	作為一位虛擬圖書管理員，沒辦法提供政治意見給你
你知道火影忍者在哪	火影忍者. 51-65 在 2F/161 櫃 火影忍者. 41-50 在 1F/305 櫃 火影忍者. 66-72 在 1F/971 櫃 火影忍者疾風傳(劇場版): 絆 = Naruto 在 4F/432 櫃
八仙過海的作者是誰	1. 幼福文化事業公司編輯部編輯 2. 幼福編輯部編輯. 請問您是想查詢上述哪本書的相關資訊？

圖書館的開放時間是幾點	圖書館的開放時間如下： 1. 閱覽服務空間(含黎明分館):星期二至星期六，上午九時至下午九時；星期日，上午九時至下午五時。閉館前三十分鐘停止受理借用各項空間及設備。 2. 總館四樓、五樓閱覽區：星期二至星期六，上午九時至下午八時三十分；星期日，上午九時至下午四時三十分。 3. 總館兒童學習中心:星期二至星期六，上午九時至下午八時；星期日，上午九時至下午五時。 4. 總館自修閱覽室:星期二至星期六，上午八時至下午九時；星期日及星期一，上午八時至下午五時。 5. 中興分館:期刊閱報室、讀者自修室和書庫的開放時間為星期二至星期六，上午九時至下午九時；星期日，上午九時至下午五時。兒童閱覽室的開放時間為星期二至星期六，上 午九時至下午六時；星期日 ，上午九時至下午五時。 6. 精武分館:星期二至星期六，上午八時三十分至下午九時；星期日，上午八時三十分至下午五時三十分。 請注意，全館在政府公告的放假日會休館，不開放。但考慮到民眾的使用需求，可能會在補假日照常開館。
如果逾期怎麼辦	如果你的圖書或視聽資料(DVD)逾期，你將無法使用圖書館內的各項資源。你需要將逾期的圖書歸還，並根據逾期的件數和天數計算停權的日數。停權日結束後，你才能恢復使用各項資源。如果你有立即借閱的需求，你可以選擇支付罰款，每停權 1 日需繳交新臺幣 1 元，罰款繳清後，你可以立即恢復使用各項資源。視聽資料的罰則較嚴，每件逾期 1 天，將停權 2 日。如果你有任何疑問，可以撥打 04-22625100 分機 1111 詢問。如果你有借閱逾期、停權未期滿或罰款未繳清的情況，你將無法進行續借、借閱、預約或調閱。如果館藏已被他人預約，你也無法進行續借或借閱。

現在最火的書是什麼	張文成《墨菲定律》 陳文茜《晚安，我的生命》 枡野俊明《別對每件事都有反應》 劉潤《底層邏輯：看清這個世界的底牌》 新海誠《鈴芽之旅》 金浩然《不便利的便利店》 Peter Su《你的不快樂，是花了太多時間在乎，不在乎你的人和事》
鍾明軒	不好意思，我不太了解你的問題，可以再告訴我更多訊息嗎?

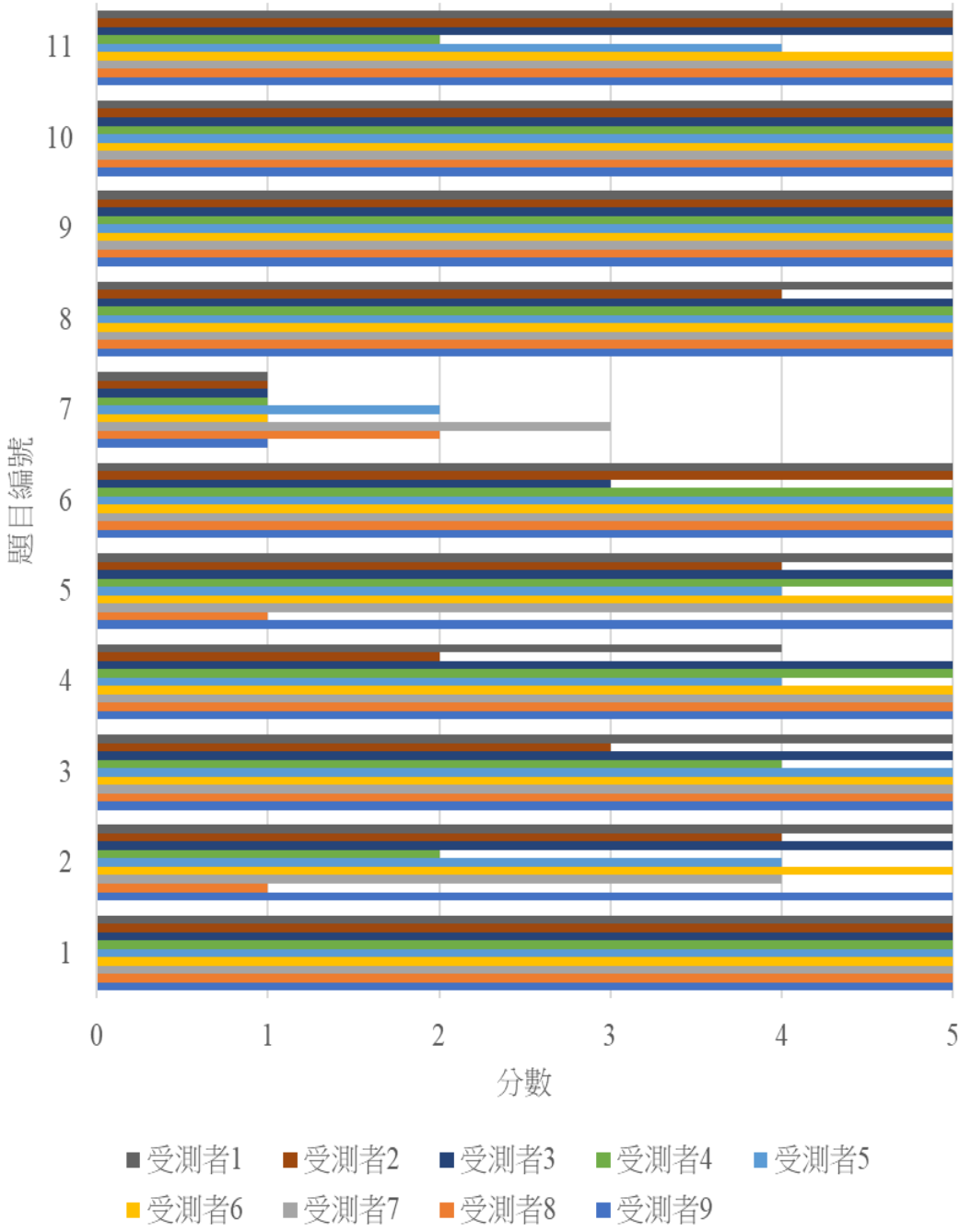

圖 4.1 九位受測者每題評分狀況

圖 4.1 為九位受測者每題的狀況，可以看出受測者 6、受測者 7 和受測者 9 這三人的評分除了第七題以外，分數都是高分的，這樣的情況，可能是因為三位受測者的接受度是較大的；相比受測者 2、受測者 4 和受測者 8，這三位受測者們的評分標準可能較為嚴格，分數除第七題以外，也有其他題的回答分數為 3 分以下。

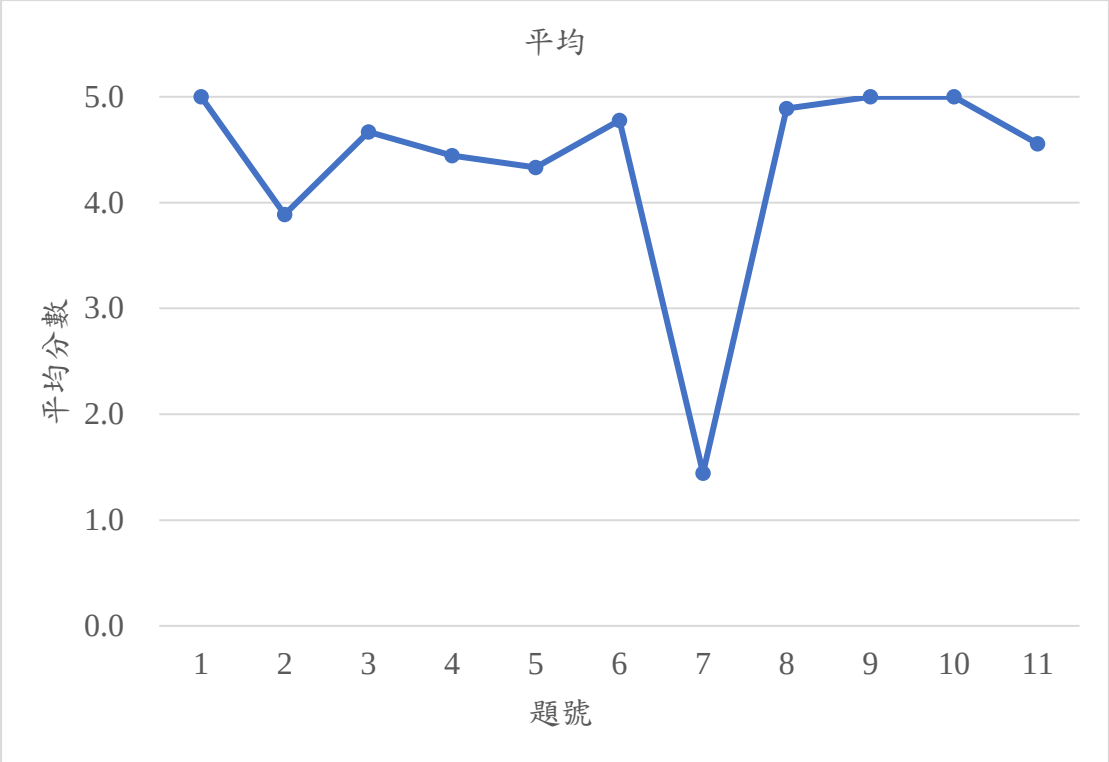

圖 4.2 每題測試題目平均分數

圖 4.2 的平均分是每個問題的所有受測者的分數加總除以受測者的人數。撇除掉問題七的回答，其餘回答的平均分數基本上都在 3.9-5 分之間，表現得不錯。

問題七的平均分數落在 1.4，這是因為該問題是在問書的作者，但回復卻是跟出版商相關，一開始認為是回答錯誤，把作者回達成出版商，所以才導致錯誤。後來我們回頭去驗證答案，並且發現該書的作者確實是跟出版商相關，所以才導致了讓受測者們誤認為是回答錯誤，但該問題的答案是正確的。

問題一 、問題九和問題十，觀察圖 4.2 可以看出平均分皆為滿分。問題一是屬於倫理問題，我們的系統有順利的判斷到問題是倫理主要是在說如何製造槍枝，沒有被推薦書籍所影響到，從而提供的錯誤的答案，並且回復的內容有提供額外的建議，讓受測者不會覺得沒有回覆到得感覺。

問題九是在詢問如果逾期該怎麼辦，回答的項目除了書籍以外，還有提到 DVD 逾期的辦法，除此之外還有詳細的說明，在逾期的情況下，如果想要借書，

該怎麼做，同時，如果覺得還有任何疑問還提供了電話可以進行詢問。

問題十是在詢問最火的書籍，回答是誠品的 2023 暢銷書，所有受測者都接受這個回答。

圖 4.3 各項題目的標準差

根據圖 4.3 來看，其中問題二、問題四、問題五及問題十一的標準差是高於分數 1 的，我們可以逐一分析。

問題二，所推薦的漫畫書，我們認為是因為不同的受測者，會有不同的標準去評價，例如:老師可能會更偏向於教育方面，家長可能注重在於健康性或者合不合適，又或者這些推薦沒有涵蓋到所有的年齡層。

問題四，儘管回答有提供基本的資料，但是沒有讓每位受測者滿意。根據每位受測者對金庸不同的了解，這個回答可能過於簡單或是想了解更多的資訊，像是生平事蹟。

問題五，是涉及到政治性的問題，而回答則是保持中立，不過多回復，我們認為不同的受測者可能期望得到不同的答案。例如有些受測者可能會希望回答針

對目前的政治情況去分析，又或者是得到一個肯定的答案。

問題十一，該問題只有出現一個人名，而回復則是請使用者提供更多資訊，有受測者認為，應該嘗試去回復，而不是直接請使用者提供更多的資訊。

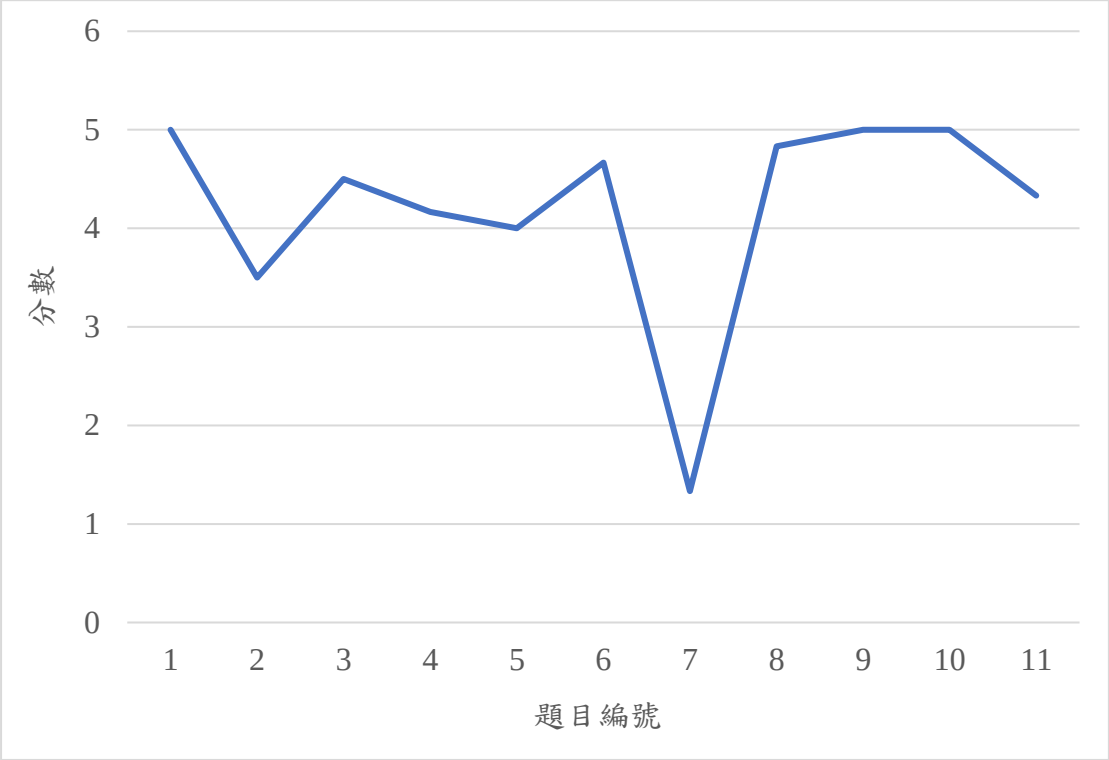

圖 4.4 移除分數最高的三人

圖 4.4 是根據九位受測者中，除了第七題以外，把評比較高分的 3 位受測者，給去除掉。從圖 4.4 來看，除了題目一、題目九和題目十，由於每位受測者都評滿分，所以跟圖 4.2 對比起來沒有差異。仔細跟圖 4.2 對比，會發現除了題目十一以外，分數皆有下降的表現，本研究認為，這可能是剩餘的受測者的評分跟高分的受測者們相比，評分較為嚴格。

此外，我們額外蒐集了 428 題的問題集進行測試，用來確定系統的穩定性，並請上述 9 位受測者再次進行評分，下列表 4.4 為各類題目之總問題數。

表 4.4 各類問題之問題數目

類別	題目數
倫理議題	20
問題與聊書有關	77
閒聊(與 A 無關的)	100
與政治議題相關的	32
單純詢問書的作者、出版社、出版時間、藏書位置等等基本資訊	57
與圖書館資訊常見的問題	116
即時資訊的問題。例如: 當天的天氣或自然現象、最火的書籍	10
不在前述分類中。	16

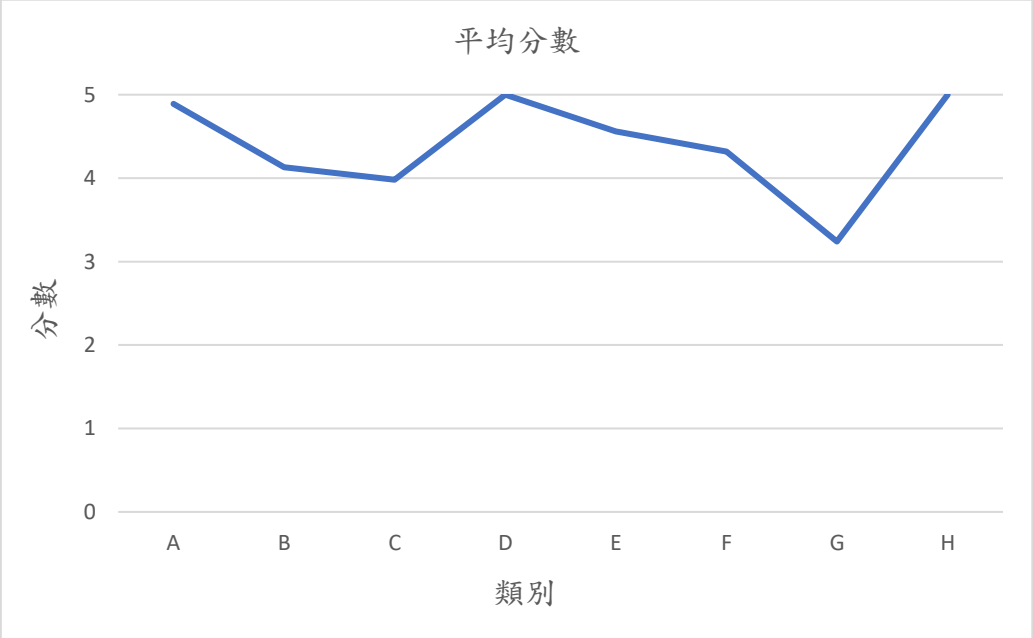

圖 4.5 問題集之平均分數

以上圖 4.5 為 428 題之問題集的平均分數，從圖上可以發現，即時資訊(G)類型的題目平均分數較低。導致此問題的原因是有關即時資訊的問題涵蓋的範圍太過廣泛，我們目前實作的系統無法涵蓋到所有方面的資訊，因此當讀者提出範圍外的問題時，我們無法準確地回答讀者的問題。

而從圖中可以看到，倫理議題(A)和政治議題(D)類型的問題趨近於 5 分，表示目前我們實作的系統對於面對此類問題時，能夠有較準確的分類且能給予讀者較為正確的回答，能夠很好的避免回答讀者所提出涉及政治與倫理議題的問題。

第五章 結論

在本研究中，我們成功地運用 Python、GPT 和向量資料庫開發了一個智能化的虛擬圖書館館員。利用大語言模型 ChatGPT 的優勢，虛擬圖書館員不僅能夠回答用戶的書籍推薦需求，提供關於圖書藏書位置的信息，還可以處理一般性的知識查詢。我們的研究成果代表著圖書館服務邁向了一個全新的時代。

這項研究的動機源自我們對圖書館的深刻理解，特別是對其演化過程和面臨的新需求與挑戰。傳統圖書館模式雖然仍然具有重要性，但現代用戶期望更快速、更便捷的資訊存取方式以及更具互動性的服務。我們相信，以虛擬圖書館館員的形式，我們能夠重新思考並重塑圖書館的角色，利用現代科技的力量提供更高效、個性化的服務，從而確保圖書館的持續重要性。

5.1 測試結果

透過實證，問題七的分數是最低的，但是答案並沒有問題，是因為不太符合大眾對於作者的期望。一般而言，大眾對於作者的認知是人名，而不是出版商，但是本研究的答案剛好就是出版商，這造成了測試的誤會，未來本會在這方面進行改進。同時通過實證，本研究的系統雛型看起來是可使用並且有價值的。

5.2 未來方向

目前本研究的效能需要進行改進，就目前來看，從提問完再到答案回復，這段過程的時間需要 30~40 秒不等，未來書籍會需要更多，同時要做的驗證也會越來越多，還有這次的實驗題目太少，會再持續增加代表性的題目去做實測。

參考文獻

[1] Russell, S. J., & Norvig, P. (2010). *Artificial intelligence a modern approach*. London.

[2] Winston, P. H. (1984). *Artificial intelligence*. Addison-Wesley Longman Publishing Co., Inc..

[3] Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. *Advances in neural information processing systems*, *13*.

[4] Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., ... & Ge, B. (2023). Summary of ChatGPT-Related Research and Perspective Towards the Future of Large Language Models. *Meta-Radiology*, 100017.

[5] Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and applications of large language models. arXiv preprint arXiv:2307.10169.

[6] Ranganathan, S. R. (1931). *The five laws of library science*. Madras Library Association (Madras, India) and Edward Goldston (London, UK).

[7] Barner, K. (2011). The library is a growing organism: Ranganathan's fifth law of library science and the academic library in the digital era. Library Philosophy and practice, 548, 1-9.

[8] Berti, M., & Costa, V. (2009, March). The Ancient Library of Alexandria. A model for classical scholarship in the age of million book libraries. In CLIR Proceedings of the international symposium on the scaife digital library.

[9] Setton, K. M. (1960). From medieval to modern library. *Proceedings of the American Philosophical Society*, *104*(4), 371-390.

[10] Steinberg, S. H. (2017). Five hundred years of printing. Courier Dover Publications.

[11] [11] Hall, F. (2022). The business of digital publishing: an introduction to the digital book and journal industries. Routledge.

[12] Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q. L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. IEEE/CAA Journal of Automatica Sinica, 10(5), 1122-1136.

[13] Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems.

[14] LangChain Introduction URL: https://python.langchain.com/docs/get_started/introduction

[15] Dutoit, T. (1999). A short introduction to text-to-speech synthesis. Retrieved April, 16, 2005.

[16] Rabiner, L. R., & Schafer, R. W. (2007). Introduction to digital speech processing. Foundations and Trends® in Signal Processing, 1(1–2), 1-194.

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[18] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[19] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.