國立臺東大學

資訊管理學系

115(級)畢業專題報告書

以人工評估的方式探討 LLM 經 RAG 訓練前後回復能力比較

——以植保機考證流程爲例

指導教授：廖國良教授

學生：韋帆輝(11112131)

趙語晴(11112138)

中　華　民　國　1　1　4　年　5　月

謝誌

在本論文完成之際，心中充滿無限的感激與敬意。此研究能夠順利完成，承蒙許多師長、同學、親友及家人的協助與支持，在此謹致上誠摯的謝意。

首先，衷心感謝我的指導教授廖國良老師，在整個研究過程中不僅提供了寶貴的學術指導，亦在研究方向、資料分析與論文撰寫上給予耐心的教導與細心的指正。老師嚴謹的學術態度與熱忱的研究精神，將成為我日後學習的典範。

其次，感謝廖國良老師與謝昆霖老師在口試或課程中所給予的寶貴意見，使本研究內容更加完整與周延。同時也感謝系上各位師長於課堂中所傳授的知識，讓我建立了堅實的學術基礎。

感謝專題研究的夥伴，在研究過程中互相協助、討論與鼓勵，使研究過程更加順利且充滿收穫。

最後，特別感謝我的家人，感謝父母長期以來無私的支持與鼓勵，是他們的理解與陪伴，使我能專心投入學業並完成本論文。也感謝所有曾經給予我幫助與啟發的人們，因為有你們的支持，這份研究才得以圓滿完成。

謹以此文，向所有曾經協助與關心我的人致上最誠摯的感謝。

韋帆輝、趙語晴　謹誌於

國立臺東大學資訊管理學系

中華民國一一四年五月

摘要

自 2018 年起，大型語言模型（Large Language Model，LLM）陸續問世並迅速發展，現今已成為現代人工作與學習中不可或缺的工具之一。然而，本研究發現，LLM 在特殊專業領域的知識掌握仍顯不足，加上容易受到「AI 幻覺」等問題影響，可能產生錯誤資訊，進而誤導使用者。

本研究旨在提升 LLM 於特定專業領域中的回答正確率，以「植保機考證流程」為例，透過導入相關知識文件與參考資料至語言模型（DeepSeek），採用補充資料的方式進行訓練，進而提高模型的應答準確性。

在研究過程中，透過比較模型訓練前後的回答內容，並結合人工評估的多項量表進行分析，評估其表現與準確率的變化。結果顯示，雖然訓練後的整體分數與正確率有所提升，但幅度並不顯著，且在表達方式與問題聚焦能力等方面亦未見明顯改善。

本研究結果顯示，模型表現可能受到導入資料內容、實驗環境及提問方式等多重因素影響，進而影響語言模型的應答表現與反應效率。

關鍵字：大型語言模型、植保機考證流程、檢索增強生成

目錄

圖目錄

表目錄

第一章、 導論

1.1 研究背景與動機

在當今社會快速變革的背景下，人工智慧技術的訓練與發展具有深刻的經濟價值和社會意義。世界經濟論壇預測 2025 年 AI 將替代 8500 萬崗位 ，但同時創造 9700 萬新崗位 [4] 。

現如今，有部分的商業巨擘相當重視員工的 AI 技能，例如亞馬遜投入 7 億美元培訓 10 萬員工掌握 AI 協作技能[2] ，可以說明合理地使用 AI 是未來的重要技能之一。

生成式 AI（如 ChatGPT、 Deepseek ）是衆多 AI 應用的其中之一，其催生內容創作產業變革的案例比比皆是，信息爆炸與知識更新加速的社會環境下，基於檢索增強生成（RAG）的大語言模型技術正成為破解「數據迷霧」與「知識過載」的關鍵工具。其價值不僅體現在技術性能提升，更在於重構知識生產與應用模式。

目前傳統大模型依賴靜態訓練數據（如 Deepseek qwen2 數據截止至 2023 年 12 月），而 RAG 可實時接入新聞、論文、市場報告等動態信息源，使得許多企業導入 RAG 來輔助工作，雲端客戶關係管理（CRM）平台 Salesforce 實測顯示 RAG 系統使客服問題解決率提高，RAG 具備高準確度及效率，可令企業省下成本並增加工作效率 [3] 。

綜上所述，RAG 訓練技術在各個方面都擁有較高的實用價值，使得 RAG 技術的訓練頗具重要性。

1.2 研究目的

本研究目的在於使用本地的 7b 蒸餾版大語言模型 Deepseek r1 ,運用檢索增強生成技術訓練，以其回復與未訓練的本地蒸餾版大語言模型做比較並研究兩者的區別，瞭解檢索增強生成訓練對於蒸餾版大語言模型的效用。

1.3 研究流程

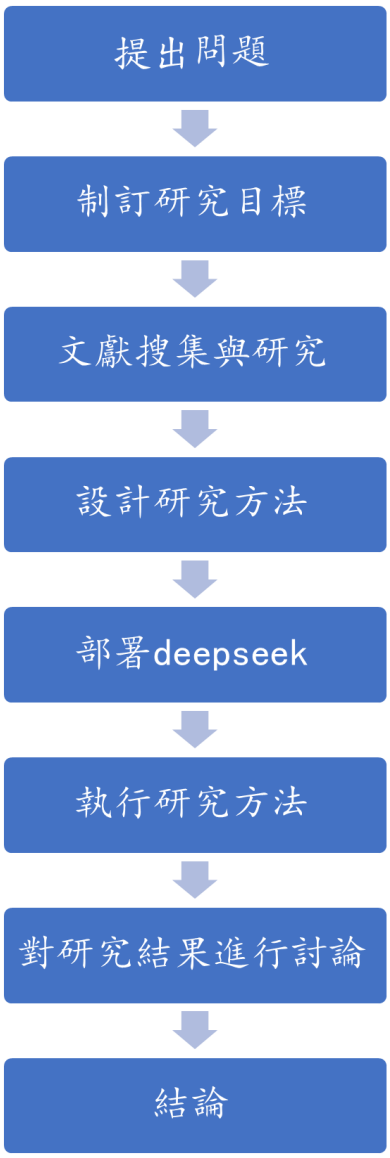

圖 1- 1：研究流程

第二章、文獻探討

2.1 大型語言模型

大型語言模型（Large Language Model, LLM）是一種基於深度學習技術，這類模型經過大規模語料庫的訓練，具備強大的語言理解與生成能力，可應用於機器翻譯、文本摘要、對話系統、程式碼生成等領域。

大型語言模型的發展可追溯至自然語言處理（NLP）研究。早期的語言模型主要採用統計方法，透過計算詞與詞之間的出現概率來進行文本生成與預測。然而，這類模型在處理長距依賴方面表現不佳，且生成的語言品質較為有限。直至 2013 年，Mikolov 等人發表了 Word2Vec 模型的相關研究 [17] ，這是一種能夠將詞轉換為向量表示的模型，使得語言模型可以更好地理解詞與詞之間的語義關係。這一方法為後續的語言模型奠定了基礎。

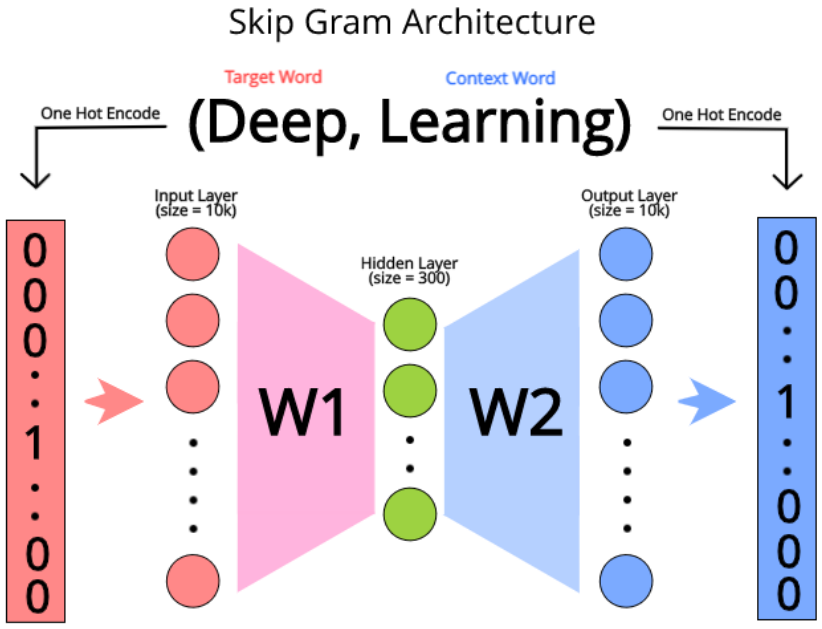

圖 2- 1： Word2Vec 模型

資料來源：知乎·《真 · Word2Vec 从理论到实战（结合 gensim）》.

近年來LLM發展具有代表性的有2017年Google 提出 Transformer 架構 [8]，google 引入了注意力機制（Attention Mechanism），能更高效地處理長距離相依性問題和大規模的計算[6]，繼 google 之後 2018 年 OpenAI 也發布了 GPT-1（Generative Pre-trained Transformer）[21]，首次應用無監督學習於語言模型，且訓練速度因為引入了並行計算而大大提升[7]。

2019 年， Google 發布了 BERT（Bidirectional Encoder Representations from Transformers）[1]，其架構源自於 Transformer 的編碼器(Encoder)[19]，能夠增強模型對文本的理解，實現雙向語境理解。

2020 年：GPT-3 憑藉 1750 億參數的規模，顯示出卓越的文本生成能力 [5] 。**2023 年後** GPT-4、Claude、PaLM 2 等更先進的 LLM 推出，擴展多模態處理能力，如圖像理解與程式碼生成，GPT 能力大幅提升，擁有較高的處理速度及推理能力，也支持長達 25,000 字的文本生成 [13]。

2.2 蒸餾（Distillation）

蒸餾法最初由 Geoffrey Hinton 所提出。蒸餾法通常需要至少兩個模型，分別為老師模型和學生模型。通常會使用一個比較強大的模型當成老師模型來教導一個比較小型的學生模型。其目的是讓學生模型模仿老師模型，不僅讓模型變的更小同時表現得也很好。[12]

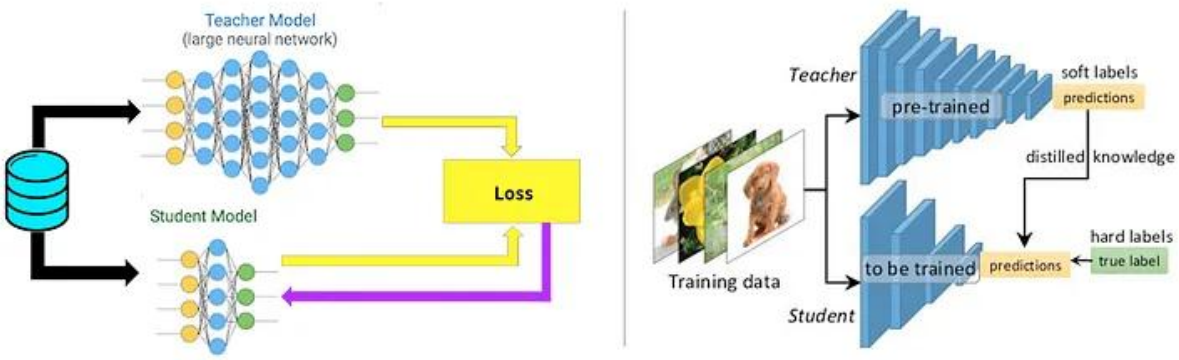

圖 2- 2：蒸餾法示意圖

資料來源：medium，《蒸餾版的 ViT: DeiT (Data-efficient image Transformers)》

2.3 檢索增強生成 (retrieval augmented generation，簡稱 RAG)

RAG(retrieval augmented generation) 首先在 2020 年 Meta(當時為 Facebook)發表的研究論文《適用於知識密集型 NLP 任務的檢索增強生成》 中提出 [8] ，是一種結合了信息檢索和文本生成的 AI 架構。

RAG 能夠在不重新訓練大型語言模型 (LLM) 的情況下，使其利用更廣泛的資料資源，從而提升生成式 AI 的品質。RAG 模型能夠根據特定組織的資料構建知識庫，並持續更新，以幫助生成式 AI 提供即時

且符合情境的回應。

透過 RAG，可以在不改動基礎模型的前提下，針對特定資訊優化 LLM 的輸出，使其獲取比 LLM 本身更即時的內容，並適用於特定產業或組織需求。

這意味著，生成式 AI 系統能夠根據最新的資料提供更貼合情境的回應，使其回答更具時效性與精確性 [10] 。

而 RAG 的運作包含兩個主要步驟，檢索與生成。當使用者輸入查詢後，RAG 會在外部文檔、內部資料庫或知識庫中執行相關性搜尋，以獲取符合查詢內容的資訊。接著，RAG 會提取這些相關資訊片段，並將其整合到內容視窗的提示中。接著，這些提示與額外資訊一同輸入至 LLM，使其能夠結合內建知識和檢索到的內容並產生回覆。

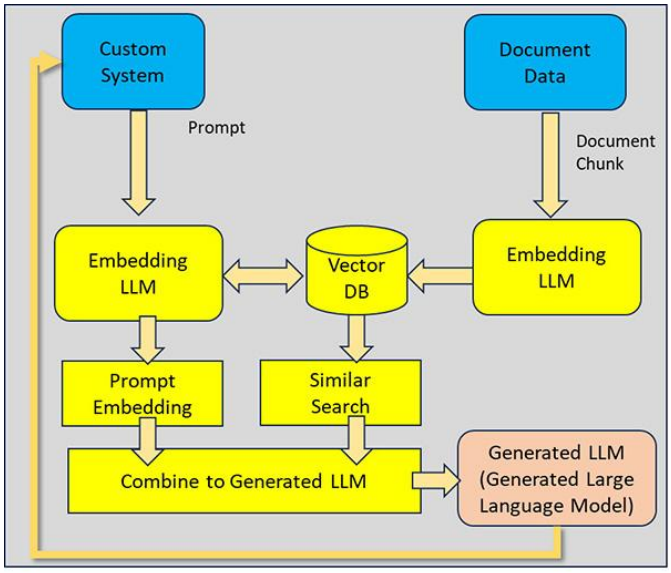

圖 2- 3：RAG 架構

資料來源：王文泰 ·RAG 解決方案規劃與實作 (CIO TAIWAN)

第三章、研究方法

3.1 研究架構

本研究使用 RAG(檢索增強生成)訓練 LLM，以達到增強模型文本品質，並改善其在特殊領域知識及問題方面的準確性及邏輯性，達到能夠準確回答特殊領域問題的目的。實驗分為三個階段，資料蒐集與處理、模型訓練以及實驗和評估。

在第一階段，本研究蒐集了關於考取無人機證照的相關網路資源及資料，並整理成一份檔案。

第二階段，本研究將提供整理後的檔案給模型，透過提問相關領域的問題來測試模型的準確性，並藉由逐步修改檔案內容的方式，以提升模型的準確性。

第三階段的實驗與評估，將透過比較模型經訓練前後的回答來檢測訓練前後的準確率差異，再以人工評估的方式設置指標並衡量模型訓練前後的答案品質差異，找到前後算數平均分數分差最高者，探討 RAG 訓練後改善幅度最大的問題。

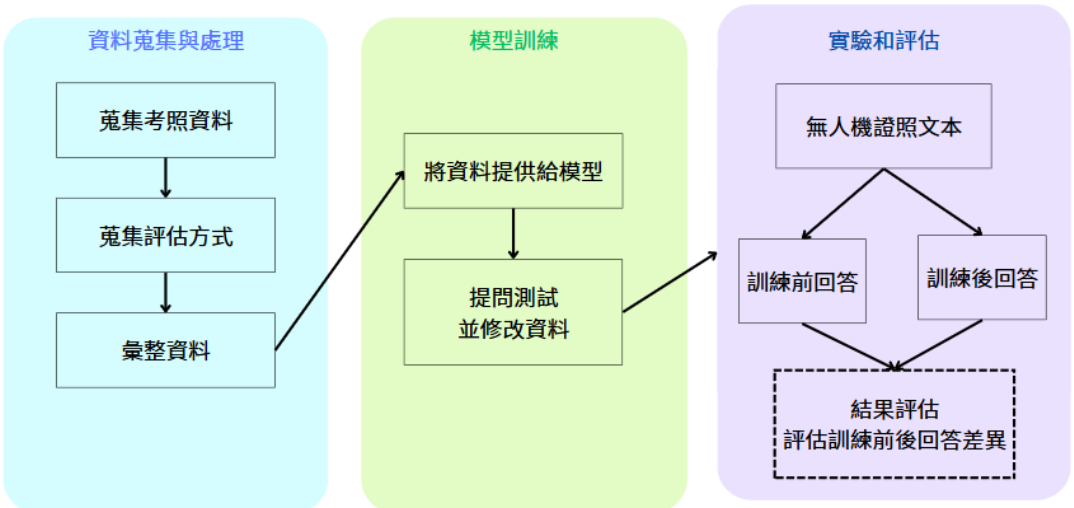

圖 3- 1：研究架構

3.2 資料集介紹

本研究所使用之資料集是以融合自身專業知識與多方資源所彙整而成。透過蒐集相關文獻資料、公開資訊及個人實務經驗，進一步加以篩選、整合與整理成一份資料。如：

無人機考照流程
考完高級學科考試后可以考基本級I2 或基本級I 基本級I2 考完可以考高級Ia2 基本級I考完一個月后可以考基本級II或高級Ia 高級Ia 考完一個月后可以考高級Ib 高級Ib 考完三個月后可以考高級IIc 基本級II考完三個月后可以考高級IIc 高級IIc 考完三個月后可以考基本級III 基本級III考完一個月后可以考高級IIId

表 3- 1

成爲植保機農藥代噴技術人員證照條件
若想成爲植保機農藥代噴技術人員，處理需要通過高級學科測驗以外，還需要術科專業高級證照以及 G2 無人機投放物品證照和藥毒所發佈的共同科目與專業空中施作證照并且具有法人資格

表 3- 2

預設考證案例
假如我要考高級Ib 流程就是： Step 1. 專業基本級操作證術科（入門證，相等於需要這個證書才能往下一步邁進） Step 2. 專業 1A (1A 為 2~15KG 的等級，如果你要操作的無人機是介於 2~15KG 的話，那考「專業基本級操作證+1A 證照」即可）

Step 3. 專業 1B (1B 為 15~25KG 的等級，如果你要操作的無人機為 15~25KG 的話，那需要考「專業基本級操作證+1A 證照+1B 證照」)

表 3- 3

術科考證前置條件
所有術科考試都需要先通過高級學科考試 術科考試基本資格需要年滿 18 歲，並通過體檢（同駕照體檢）

表 3- 4

逐級考解釋
民航局的操作證是採取「逐級考」的制度,逐級考的意思是若你要考的證照有前置證照 ,那就必須將前置證照考到並等待數月才能接著考所需證照

表 3- 5

共同科目訓練及藥代噴技術人員訓練說明
藥毒所共同科目訓練與考試 課程內容：農藥法規、安全施作知識、藥劑調配等理論課程。 時間與費用：8 小時課程（含 1 小時考試），需通過筆試。 通過後：取得「共同科目訓練及格證明」，方可報考下一階段專業空中施作。 藥毒所專業空中施作訓練與考試

課程內容：實際操作無人機噴灑、航線規劃、緊急處理等。 時間與費用：14 小時課程（含 2 小時考試），總計 16 小時。 通過後：取得「農藥代噴技術人員訓練及格證明書」

表 3- 6

本研究參考了部分網路資源，如交通部民用航空局：

交通部民用航空局（簡稱民航局）隸屬交通部，負責規劃、管理與監督我國民用航空相關事務。其主要職責包括航空運輸管理、飛航安全監督、機場設施規劃、航空器適航審查及航空人員執照核發等。

其官網收錄了考照流程、註冊及考試等相關資訊，以及航空安全、政策等飛航相關的各大資訊，其中也有收錄本研究需要的報考及考照流程等資訊。

圖 3- 2：交通部民用航空局 · 新逐級報考流程圖

3.3 語言模型介紹

本次研究使用 ollama（一個可在本地運行 LLM 的服務工具）運行 Deepseek r1 的 7b 蒸餾版進行測試。

DeepSeek R1 語言模型是 AI 新創公司 DeepSeek（深度求索）於 2025 年 1 月推出一款免費人工智慧應用程式，以相較於 GPT-o4 約十分之一的開發價格達到類似的對話水準，并且而 Deepseek r1 7b 蒸餾版則是深度求索公司運用蒸餾技術以 Deepseek 完整版訓練出來的擁有七億參數量級的縮減版本。

3.4 人工評估

本研究采用人工評估的方式分別評估 LLM 經過 RAG 訓練前後的回復，人工評估采用雙人獨立評估，即兩名實驗人員在獨立狀況各自下給 LLM 的回復做一至五分評分。

評分標準	分數
不符合	1
較不符合	2
普通	3
較符合	4
符合	5

表 3- 7：人工評估標準

評分標準主要參考以下方面

（1）任務完成度

對指令的理解是否準確（例如遵循格式、步驟要求）？

（2）語言自然度

回答是否流暢、符合語法？

用詞是否貼近人類的表達習慣？

（3）事實準確性

輸出的資訊是否符合事實？

是否避免虛假或誤導性陳述？

（4）邏輯性與連貫性

回答是否邏輯清晰、結構合理？

上下文是否一致？

（5）相關性

回答是否緊扣問題核心？

是否過度冗長或缺乏重點？

第四章、研究設計與結果

4.1 實驗環境

本研究使用的電腦設備規格如下表所示。

OS	Windows 11
CPU	AMD Ryzen 7 4800H with Radeon Graphics
GPU	AMD Radeon Graphics Processor (0x1636)
記憶體	16 GB
LLM	deepseek-r1:7b

表 4- 1：電腦配置

4.2 實驗結果

4.2-1 題目與回復

題目	要怎麼拿到藥毒所發布的共同科目與專業空中施作這兩份證照
訓練前	若您想要取得中國藥典委員會發布的《共同科目》與《專業空中施工作業證書》，請依照以下步驟操作： 1. **造訪官方網站**：請前往中國藥典委員會的官方網站（https://www.cpc.org.cn/）或相關培訓機構網站。 2. **查詢報名資訊**： 　◦ 在官網或培訓機構網站中，尋找《共同科目》及《專業空中施工作業證書》的報名連結。 　◦ 依照指示填寫個人資訊並完成註冊流程。 3. **關注官方公告**：請持續留意官方網站及其平台所發布的證書與考試通知，確保掌握最新資訊。 4. **聯繫當地培訓機構**：如有疑問或需要進一步協助，建議聯繫正規藥典培訓機構或當地的諮詢服務中心。 按照以上步驟操作，您將能順利取得所需證書。
訓練後	要獲得藥毒所發布的共同科目與專業空中施作證照，通常需要按照以下步驟辦理： 1. **確定期望的資格條件**：確保自己符合相關資格條件。例如，可能需要具備一定的教育背景、專業訓練或相關工作經驗。 2. **了解課程與要求**： 　◦ **共同科目（General Subjects）**：包括農藥法規、安全施作知識、藥劑調配等理論內容。 　◦ **專業空中施作科目（Professional Aerial Application）**：可能涉及更多實用技能，如操作無人機、噴灑技術等。

	3. **準備相關材料**： ○ 學歷證書（如大專及以上） ○ 職業相關的技術證照或背景證明 ○ 通過相應測驗的成績或證明 ○ 資格證書和其他相關證照 4. **報名並參加課程**： ○ 查詢藥毒所或相關官方管道的正式課程安排。 ○ 填寫報名表並繳交相關費用。 ○ 參加課程並完成學業。 5. **通過考試**： ○ 成功通過課程的考核，通常包括筆試和實作考核。 ○ 確保達到所需的通過標準。 6. **申請正式證照**： ○ 提交所有完成的文件和證書。 ○ 繳費並按照指定程序申請正式證照。 7. **後續步驟**： ○ 開始實際工作並確保符合證照要求。 ○ 可能需要定期更新或維護證照，以保持有效性。 其他建議： • 查閱藥毒所的官方網站或相關公告，取得最新的課程與申請資訊。 • 聯繫專業的培訓機構或為空勤工作提供服務的機構，尋求專業指導和報名服務。 • 加入相關專業團體或論壇，獲取更多資訊與支援。 請提供更多背景資訊，以便我能給您更具體的指導。

表 4- 2：題目一回答

題目	民航局的操作證採用「逐級考」制，「逐級考」是什麼意思
訓練前	「逐級考」是「分級考試」的簡稱，是指民航局在發放操作證（例如航空乘務員、航空維修師、飛行員等）時所採用的**分階段考試制度**。這種制度要求考生依照等級順序逐步通過各項考試，每通過一級，才有資格參加下一級的培訓與考試。 具體內容通常包括： • **理論知識考試**：例如航空法規、操作手冊、緊急應變程序等； • **實作／模擬操作測驗**：包括模擬器訓練或實際操作評估； • **實習與工作經驗審查**：部分等級還要求一定的實際工作時數。 此制度的設計，能有效確保每位航空從業人員的專業能力作安全水準，為航空安全提供保障。
訓練後	「逐級考」是指民航局的操作證書考試採用按照資格層次逐步考核的方式。具體來說，申請人需要按照資格等級從低到高依次考試，通過後才能獲得相應的操作證。例如，從 A 類到 J 類，每個級別都需要通過特定的考試和實踐考核才能升級。這種方式旨在確保申請人具備相應的專業能力和安全操作技能。

表 4- 3：題目二回答

題目	民航局的術科考試有什麼樣的前置條件
訓練前	民航局的術科考試通常有以下前置條件： 1. **教育背景**：通常要求考生具有相應的學歷背景，如高中畢業或以上文憑。不同職位可能有不同的學歷要求。 2. **工作經驗**：對於部分需要實際操作技能的工作（如機械師、飛行管制員等），可能需要一定的相關工作經驗。例如，飛行員的基本訓練通常包括空中交通管制和飛行員訓練。 3. **身體條件**：考生需符合民航局規定的身體標準，包括心臟病史、視力、體型等要求。 4. **語言能力**：考生需能流利使用英語或中葡語溝通，具備基本的聽說讀寫能力。部分崗位可能需要中葡語能力（如香港特別行政區）。 5. **專業培訓**：考生需完成民航局指定的專業培訓課程，並取得相應的證書或結業證明。 6. **心理素質**：考生需具備良好的心理素質，包括適應壓力、判斷力以及應對緊急狀況的能力。 7. **其他要求**：不同崗位可能有不同的額外要求，如語言能力、特殊技能等。 建議考生提前了解所報考崗位的具體要求，並備齊所有必備條件，才能順利參加術科考試。
訓練後	民航局的術科考試通常有以下前置條件： 1. **專業教育背景**：考生需先完成相應專業的專科或本科教育（例如農業、林業等相關專業），並取得相應的畢業證。

	2. **理論知識通過**：考生需通過民航局指定的理論科目考試，內容包括農藥法規、安全施作知識、藥劑調配等基於實際作業需求的內容。 3. **技術考試通過**：考生需通過民航局舉辦的技術科目考試，內容涵蓋操作技能、無人機使用、藥劑調配等內容，並示範實務技能。 4. **工作經驗積累**：在相關領域中積累一定的實際工作經驗，並可能需要通過民航局的考核或認證。 5. **法人資格**：通常需具備一定的法人資格（如公司註冊或商業登記），並能有效執行相關操作。 6. **語言能力**：需通過民航局指定的語言考試，並具有基本的英文交流能力。 具體前置條件可能因不同崗位而有所不同，建議查看民航局官方網站或相應崗位公告。

表 4- 4：題目三回答

題目	如果我想考到高級Ib 操作證至少需要多久時間
訓練前	要獲得**高級IB 操作證**，通常所需的時間約為 **3 至 6 個月**。具體所需時間可能因以下因素而有所不同： 1. **學習內容的複雜性**：操作證涉及的專業知識和技能需要一定的時間來學習和掌握。 2. **學習方式**：如果選擇了遠距學習或者參加短期培訓班，可能所需時間較短；但如果選擇自學，則可能需要更多時間安排複習和實踐。 3. **學習進度的差異**：不同人的學習能力和知識儲備也會影響所需的時間。 4. **考試準備時間**：除了學習內容外，還需要時間來準備考試（如模擬測試、複習等）。 總而言之，如果選擇系統的訓練並安排充足的複習時間，3 至 6 個月是合理的預期。具體需求建議諮詢相關培訓機構或報名後詳細規劃。
訓練後	考取高級IB 操作證的所需時間因具體崗位和學習方法而有所不同。一般來說，如果考生有相關的工作經驗並且實際學習相應的理論知識、操作規範和風險控制技能，則可能需要幾個月到一年之久來充分準備和通過考試。 具體來說，學習時間可能會根據以下因素有所差異： 1. **學習內容的複習與理解**：需要充分掌握無人機操作的基本法規、藥劑調配、安全施作等相關內容，通常需要至少 3-6 個月的時間。 2. **實際操作練習**：理論學習後，還需要在實際操作中逐步提高技能，這可能需要更多的時間和實踐經驗。

	3. **考試準備與通過**：如果考生有充分的準備，可能需要 1-3 個月來復習、參加考試並通過。 總之，具體所需時間因人而異，但通常至少需幾個月到一年之久才能充分準備並通過高級IB 操作證考試。具體建議根據自己的工作經驗、學習能力和發展目標來制定個人化的學習計劃。

表 4- 5：題目四回答

題目	如果我想成為無人機農藥代噴人員，我需要擁有那些證照(無人機重 30kg)
訓練前	成為無人機農藥代噴人員通常需要一定的資格證書來保障安全性和規範性。具體所需的證書可能包括： 1. 共同科目證書 • 所有無人機作業都需要先通過「共同科目」的考試或登錄，這是無人機操作的基本知識和技能的基礎。 2. 專業空中施工作業證書 • 根據無人機的體重（≤30 公斤）和具體任務性質，可能需要不同等級的證書。例如： ○ **普通無人機作業許可證**：適合一般性無人機操作。 ○ **農藥代噴特種許可證**：具體規定用於農藥代噴的特殊規範。 3. 農藥操作證書 • 取決於具體的農藥代噴規程和規定，可能需要取得「農藥操作人員」的相關資格證書或訓練證明。這通常是透過職業資格考試或操作訓練考核獲得的。 4. 其他相關資格（若適用） • 根據當地法規或企業規定，可能還需要其他額外的資格證明。

	考試與頒證流程 1. **共同科目考試**：通常是必備，內容涵蓋無人機操作基本知識、安全規定等。 2. **專業空中施工作業許可證**：根據無人機體重和任務性質，通過相關考試或登錄審核。 3. **農藥操作人員訓練**：透過職業資格考試或操作訓練認證。 綜合建議 1. 仔細閱讀當地或雇主規定的具體資格要求。 2. 可報名參加相關的正式考試或培訓課程，例如： 　○ 品質監管單位舉辦的專業空中施工作業考試。 　○ 農藥操作人員職業資格考試。 3. 培訓中心通常會為需要這些證書的人提供專業的培訓與考試服務。 透過以上步驟，您將能更清晰地了解成為無人機農藥代噴人員所需的證書與條件。
訓練後	成為無人機農藥代噴人員，一般需要具備相應的無人機操作證照，具體來說： 1. **高級ＩＢ操作證**：這是基本要求，用於操作無人機進行農藥代噴作業。 2. **藥毒學科測試（共同科目）**：需要通過無人機與藥毒相關的共同科目測試，內容包括農藥法規、安全施作知識、藥劑調配等。 3. **專業無人機操作證照**：具備專業操作 30 公斤無人機所需的相關證照。

	4. **相關專業知識和經驗**：需要具備農藥噴灑相關的專業知識，以及實際操作經驗。如果是專門從事農藥代噴工作，可能還需要一定的農業或植物保護背景。 具體所需時間因人而異，但一般而言，通過相應的測試和積累經驗後，可以順利獲得相關證照。

表 4- 6：題目五回答

4.2-2 各題目評分

題目：要怎麼拿到藥毒所發布的共同科目與專業空中施作這兩份證照

指標序號	評分指標	訓練前平均得分	訓練後平均得分
1	對指令的理解是否準確？	2.5	3.5
2	回答是否流暢、符合語法？	3.5	3.5
3	用詞是否貼近人類的表達習慣？	2.5	3
4	輸出的資訊是否符合事實？	2.5	2.5
5	是否避免虛假或誤導性陳述？	1.5	2
6	回答是否邏輯清晰、結構合理？	3.5	3.5
7	上下文是否一致？	3.5	3.5
8	回答是否緊扣問題核心？	3	3
9	是否過度冗長或缺乏重點？	2.5	3.5

表 4- 7：題目一評分

題目：民航局的操作證採用「逐級考」制，「逐級考」是什麼意思

指標序號	評分指標	訓練前平均得分	訓練後平均得分
1	對指令的理解是否準確？	4	4
2	回答是否流暢、符合語法？	4.5	4.5
3	用詞是否貼近人類的表達習慣？	3.5	3.5
4	輸出的資訊是否符合事實？	3.5	4
5	是否避免虛假或誤導性陳述？	3.5	3.5
6	回答是否邏輯清晰、結構合理？	3.5	3.5
7	上下文是否一致？	4	4
8	回答是否緊扣問題核心？	4	3
9	是否過度冗長或缺乏重點？	2.5	3.5

表 4- 8：題目二評分

題目：民航局的術科考試有什麼樣的前置條件

指標序號	評分指標	訓練前平均得分	訓練後平均得分
1	對指令的理解是否準確？	3.5	4
2	回答是否流暢、符合語法？	3	3.5
3	用詞是否貼近人類的表達習慣？	3	3
4	輸出的資訊是否符合事實？	1.5	2.5
5	是否避免虛假或誤導性陳述？	1.5	2.5
6	回答是否邏輯清晰、結構合理？	3.5	3.5
7	上下文是否一致？	4	4
8	回答是否緊扣問題核心？	3	3.5
9	是否過度冗長或缺乏重點？	4	3

表 4-9：題目三評分

題目：如果我想考到高級Ib 操作證至少需要多久時間

指標序號	評分指標	訓練前平均得分	訓練後平均得分
1	對指令的理解是否準確？	2.5	2.5
2	回答是否流暢、符合語法？	3.5	3.5
3	用詞是否貼近人類的表達習慣？	3.5	3.5
4	輸出的資訊是否符合事實？	1.5	2.5
5	是否避免虛假或誤導性陳述？	2	2
6	回答是否邏輯清晰、結構合理？	3	3
7	上下文是否一致？	3.5	3.5
8	回答是否緊扣問題核心？	3	3
9	是否過度冗長或缺乏重點？	4	4

表 4- 10：題目四評分

題目：如果我想成為無人機農藥代噴人員，我需要擁有那些證照(無人機重 30kg)

指標序號	評分指標	訓練前平均得分	訓練後平均得分
1	對指令的理解是否準確？	2.5	3.5
2	回答是否流暢、符合語法？	3	3.5
3	用詞是否貼近人類的表達習慣？	2	2.5
4	輸出的資訊是否符合事實？	1.5	3
5	是否避免虛假或誤導性陳述？	2	3.5
6	回答是否邏輯清晰、結構合理？	2.5	3.5
7	上下文是否一致？	2.5	4
8	回答是否緊扣問題核心？	2	3.5
9	是否過度冗長或缺乏重點？	2	3.5

表 4- 11：題目五評分

4.2-3　數據比較

指標序號	前	後	差異
1	3	3.5	0.5
2	3.5	3.7	0.2
3	2.9	3.1	0.2
4	2.1	2.9	0.8
5	2.1	2.7	0.6
6	3.2	3.4	0.2
7	3.5	3.8	0.3
8	3	3.2	0.2
9	3	3.5	0.5
	訓練前	訓練後	
	2.922	3.311	0.389

表 4- 12：訓練前後評分差異表

數據顯示，訓練前分數算數平均數為 2.922 分，而訓練後數算數平均數為 3.311 分，兩者分數相差 0.389 分。

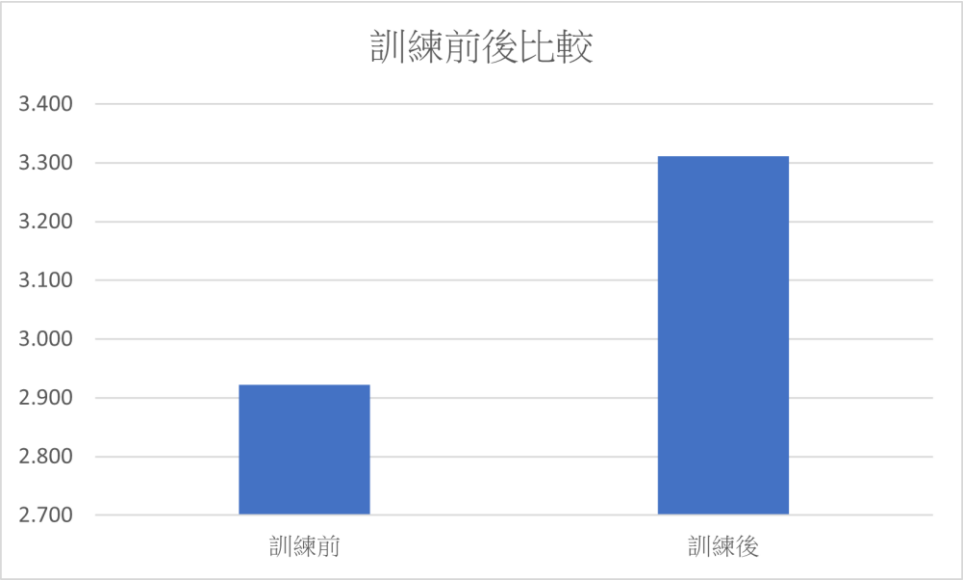

圖 4- 1：訓練前後比較——直方圖

若以各個評分指標的五題平均數來說，訓練前指標平均分最高的是 2 號與 7 號，兩者皆是 3.5，訓練後平均分最高是 7 號。 訓練前指標平均分最低為 4 號和 5 號，僅有 2.1 分，訓練後則是 5 號，為 2.7 分。其中訓練前後分數差異最大的是 4 號指標，相差 0.8 分，分差最小是 2、3、6、8 號，分差為 0.2 分。

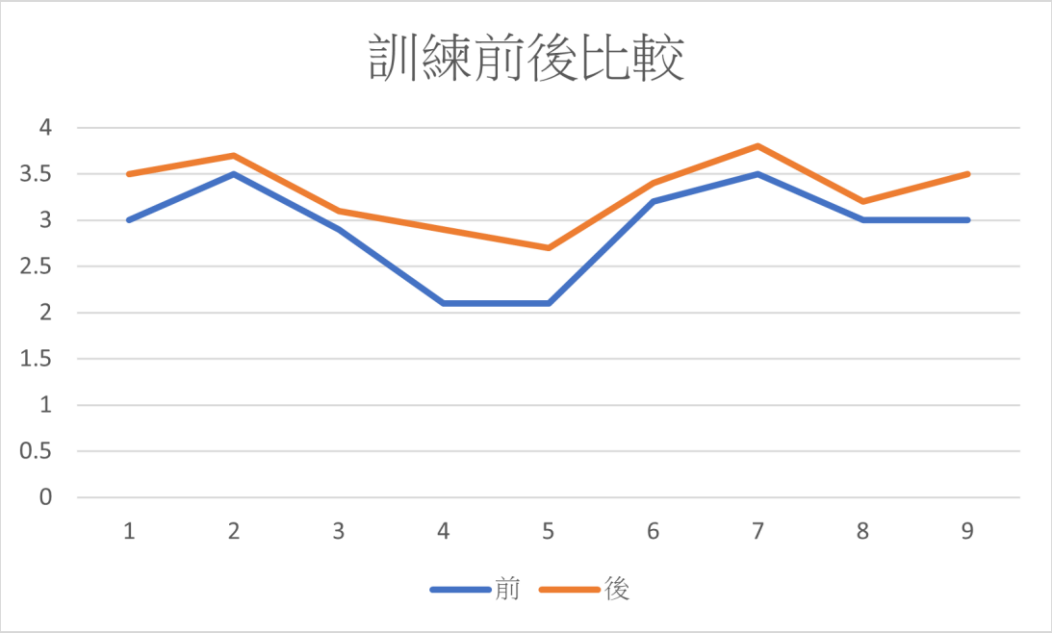

圖 4- 2：訓練前後比較——折線圖

第五章、結論與建議

5.1 結論

依實驗數據顯示，人工評估研究認爲經 RAG 訓練後的 deepseek 的文本生成質量有部分提升，以總平均分來看，訓練後分數較訓練前提升約莫百分之 13，提升幅度較不明顯，有較多的評分指標訓練前後相差較小，如指標 2、3、6、8 號，其前後相差僅 0.2 分，可以說明 RAG 訓練在當前主題小文本訓練下，對於 LLM 回復的語言順暢、表達習慣及問題集中能力并沒有顯著提升，在考量上下文是否有邏輯違背方面的 7 號指標訓練前後分差也較小，僅有 0.3 分。

在 RAG 訓練後，平均分數最低的是 5 號指標，與訓練前最低分指標相同，雖然前後分差為 0.6 分，但仍是蒸餾版 deepseek 在此主題中最不擅長的方向，所以我們認爲在數據有限且不能聯網的狀態下，LLM 較容易誤導使用者，比較無法知道使用者問題的用意。

而同樣是訓練前評分最低的 4 號指標則在本次實驗中分數成長最大，足足有 0.8 分，近乎增加了一個評分階段，可以說 RAG 訓練在改善回復是否真實方面有不小的效果，縱使回復可能不切合問題，但回復的真實性有較大提升。

綜上所述，RAG 訓練在改善內容正確性上較有幫助，包括指標 1、4、5、9 號改善分數差都在 0.5 及以上，然在較爲底層的說話邏輯與說話方式上 RAG 訓練效果相當有限, 并且額外的文本訓練并不能很好的幫助其理解問題。

參考文獻

1. 王孟萱，基於知識圖譜與 RAG 方法建構 LLM 解題系統：以論語為例，2024。

2. 愛范兒，擔心機器人搶飯碗？亞馬遜教 10 萬員工這樣保住工作，科技新報，Jul. 2019https://technews.tw/2019/07/16/amazon-pledges-to-upskill-100000-u-s-employees-for-in-demand-jobs-by-2025/。

3. 阿里·本德斯基，什麼是檢索增強生成（RAG）？，Salesforce，2024，

   https://www.salesforce.com/tw/agentforce/what-is-rag/。

4. 張心瑜，AI 已進入日常 世界經濟論壇：機器 2025 年將取代 8500 萬個工作崗位，Yahoo 新聞，2023，

   https://tw.news.yahoo.com/ai%E5%B7%B2%E9%80%B2%E5%85%A5%E6%97%A5%E5%B8%B8-%E4%B8%96%E7%95%8C%E7%B6%93%E6%BF%9F%E8%AB%96%E5%A3%87%E6%A9%9F%E5%99%A82025%E5%B9%B4%E5%B0%87%E5%8F%96%E4%BB%A38500%E8%90%AC%E5%80%8B%E5%B7%A5%E4%BD%9C%E5%B4%97%E4%BD%8D-040246012.html ， Mar. 。

5. 張奕煌，GPT-3 與 GPT-4 的比較，知乎，May. 2023，

   https://zhuanlan.zhihu.com/p/633481799。

6. 陳家駿、許正乾，Google Transformer 模型專利 –ChatGPT 自注意力機制之重要推手，STPI 科技政策研究與資訊中心，2023，

https://iknow.stpi.narl.org.tw/post/Read.aspx?PostID=19886。

7. 蝈蝈俊，從 GPT-1 看 Transformer 的崛起，部落格園，2024，

https://www.cnblogs.com/ghj1976/p/18282541/conggpt1kantransformer-de-jue-qi。

8. TAIDE推動台灣可信任生成式AI發展計畫 如何通過RAG提升生成式AI，2023，

https://taide.tw/index/newsList/newsDetail/4b1141818d642b71018d81b5992d3ace。

9. AI 講講話，大型語言模型（LLM）的發展，Medium，Aug. 2023，

https://medium.com/aigeneral/%E5%A4%A7%E5%9E%8B%E8%AA%9E%E8%A8%80%E6%A8%A1%E5%9E%8B-llm-%E7%9A%84%E7%99%BC%E5%B1%95-3da24f833020。

9. Alan Zeichick，什麼是 Retrieval-Augmented Generation (RAG)？，Oracle，2023 ， https://www.oracle.com/tw/artificial-intelligence/generative-ai/retrieval-augmented-generation-rag/

10. amosgee，真 ·Word2Vec 從理論到實戰（結合 gensim），知乎，Jul. 2020，

https://zhuanlan.zhihu.com/p/157144548

11. SHOPLINE，【2025 最新】ChatGPT 升級 GPT-4、GPT-4o 完整解析功能特點、免費及付費方案，SHOPLINE Blog，Feb. 2025，

https://blog.shopline.hk/gpt-4/

12. Haren Lin，[Notes] BERT / BERT 架構理解，Medium，2021，https://haren.medium.com/paper-notes-bert-bert-%E6%9E%B6%E6%A7%8B%E7%90%86%E8%A7%A3-31c014d7dd63。

13. Simon Liu，淺談 GPT 生成式語言模型(1) — 過去，2023，InfuseAI Blog，https://blog.infuseai.io/gpt-model-past-introduction-1e2558462e41。

14. Authors of "Evaluation of a retrieval-augmented generation system using a Japanese Institutional Nuclear Medicine Manual and large language model-automated scoring"，"Evaluation of a retrieval-augmented generation system …"，期刊 / 公開稿件，2025。 PubMed

15. Authors of "Evaluation of a context-aware chatbot using retrieval-augmented generation for answering clinical questions on medication-related osteonecrosis of the jaw"，"Evaluation of a context-aware chatbot …"，Journal of Cranio-Maxillofacial Surgery，2025。

16. Anum Afzal, Alexander Kowsik, Rajna Fani, Florian Matthes，Towards Optimizing and Evaluating a Retrieval Augmented QA Chatbot using LLMs with Human-in-the-Loop，arXiv，2024。 arXiv+1

17. Geoffrey Hinton, Oriol Vinyals, Jeff Dean，Distilling the Knowledge in a Neural Network，arXiv，Mar. 2025，https://arxiv.org/abs/1503.02531

18. Ivan Iaroshev，Ramalingam Pillai, Leandro Vaglietti, Thomas Hanne，Evaluating Retrieval-Augmented Generation Models for Financial Report Question and Answering，Applied Sciences，2024。 MDPI+1

19. Moris，BERT: Pre-training of Deep Bidirectional Transformers for Language

Understanding，Medium，2019，https://medium.com/ml-note/bert-pre-training-of-deep-bidirectional-transformers-for-language-understanding-d76985771507

20. Nandan Thakur, Ronak Pradeep, Shivani Upadhyay, Daniel Campos, Nick Craswell, Jimmy Lin，Support Evaluation for the TREC 2024 RAG Track: Comparing Human versus LLM Judges，arXiv，2025。 arXiv

21. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean，Efficient Estimation of Word Representations in Vector Space，arXiv，Sep. 2023，https://arxiv.org/abs/1301.3781

22. Xanh Ho, Jiahao Huang, Florian Boudin, Akiko Aizawa，LLM-as-a-Judge: Reassessing the Performance of LLMs in Extractive QA，arXiv，2025。 arXiv

23. YuHe Ke，Liyuan Jin，Kabilan Elangovan, Hairil Rizal Abdullah，Nan Liu，Alex Tiong Heng Sia， Chai Rick Soh，Joshua Yi Min Tung，Jasmine Chiat Ling Ong， Daniel Shu Wei Ting，Development and Testing of Retrieval Augmented Generation in Large Language Models -- A Case Study Report，arXiv，2024，arXiv

24. Yen-Shan Chen, Jing Jin, Peng-Ting Kuo, Chao-Wei Huang, Yun-Nung Chen，LLMs are Biased Evaluators But Not Biased for Retrieval Augmented Generation，arXiv，2024。 arXiv+1