

以人工評估的方式探討LLM經 RAG訓練後回復能力之比較 —— 以植保機考證流程為例

11112131 韋帆輝

11113138 趙語晴

目錄

研究內容與目的

為何選擇Deepseek r1

什麼是RAG / 為何選擇人工評估

研究流程與資料集

評估方式與指標

題目與答案說明

成果數據比較(結論)

摘要

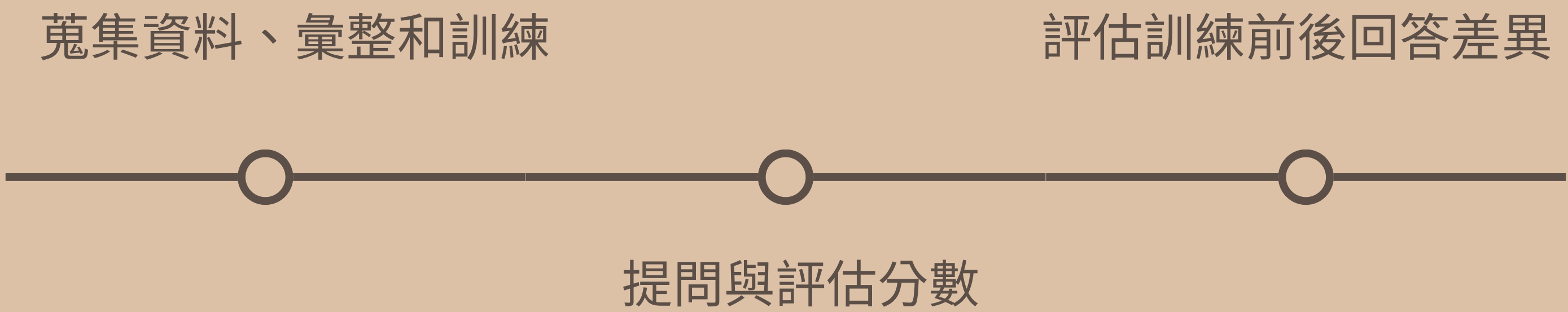
本研究使用Deepseek r1本地蒸餾（7b）進行預先的RAG訓練，以人工評估的方式檢測訓練前後結果，評估人員有兩位，在互不干涉的情況獨自完成檢測。問題是基於無人機考試中的植保機專項，題目有5題9向度總計45項分數。

研究目的

以“尋常人”的使用角度來發想，以人工評估的方式測試RAG訓練後在經驗性邏輯問題上的變化。

我們對“尋常人”的定義是不清楚答案的情況下去詢問LLM以獲得咨訊，問問題近口語話且不會過分追求問題表達準確程度，并且以實用性為主，所以沒有評估LLM的回答是否不符合道德倫理。

研究流程



爲什麼選擇Deepseek r1本地蒸餾（ 7b ）

DeepSeek-R1是於2025年1月20日發布的人工智慧大型語言模型，專門適用於數學、編碼和邏輯等任務，性能對標OpenAI o1

DeepSeek稱該模型用了強化學習訓練，並展現了 o1 沒有公開的完整思考過程

SuperCLUE的2015/1/8榜單deepseek位居第四，前三均是gpt且需付費

什麼是RAG

RAG（檢索增強生成）。檢索指的是檢索外部知識庫，增強生成指的是將檢索到的知識送給大語言模型以此來優化大模型的生成結果，使得大模型在生成更精確、更貼合上下文答案的同時，也能有效減少產生誤導性資訊的可能

方式	作用區域	劣勢	優勢
提示詞	提示詞	長度限制，消耗大，速度慢	方便靈活
RAG	提示詞後，推理前	檢索拖慢響應時間，考驗檢索能力	知識庫易更新，提高數據真確性
微調	大模型本身	訓練成本高，時效性差	通用能力強，消耗少，速度快

爲什麼選擇人工評估

常見的評估方式有4個方向，分別是：

1. 特定任務指標
2. 研究機構發佈基準
3. LLM自我評估
4. 人工評估

高級和專業級在無人機證照裏是同義

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意外，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

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」

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）

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。

考完高級學科考試后可以考基本級 I 2或基本級 I

基本級 I 2考完可以考高級 I a2

基本級 I 考完一個月后可以考基本級 II 或高級 I a

高級 I a考完一個月后可以考高級 I b

高級 I b考完三個月后可以考高級 II c

基本級 II 考完三個月后可以考高級 II c

高級 II c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 III d

假如我要考高級 I b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」）

民航局術科需自備對應重量無人機，建議選擇支援GPS定位模式的機型

法人需通過民航局「能力審查」並投保責任險

學科測驗的成績只保留一年喔

評估方式與指標

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

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

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。

題目與答案節錄(一)

詢問意義：

- 考驗邏輯推理能力(時長計算)
- 內容合理性

結論：

輸出資訊正確度有提升，有掌握複習內容(指標4)

且能夠推理大致時長，但基準仍然不明確(見指標5)

題目與答案節錄(二)

詢問意義：

- 確認模型確實接收到資料集內容
- 確認內容虛假部分與誤導性陳述的出現率

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。

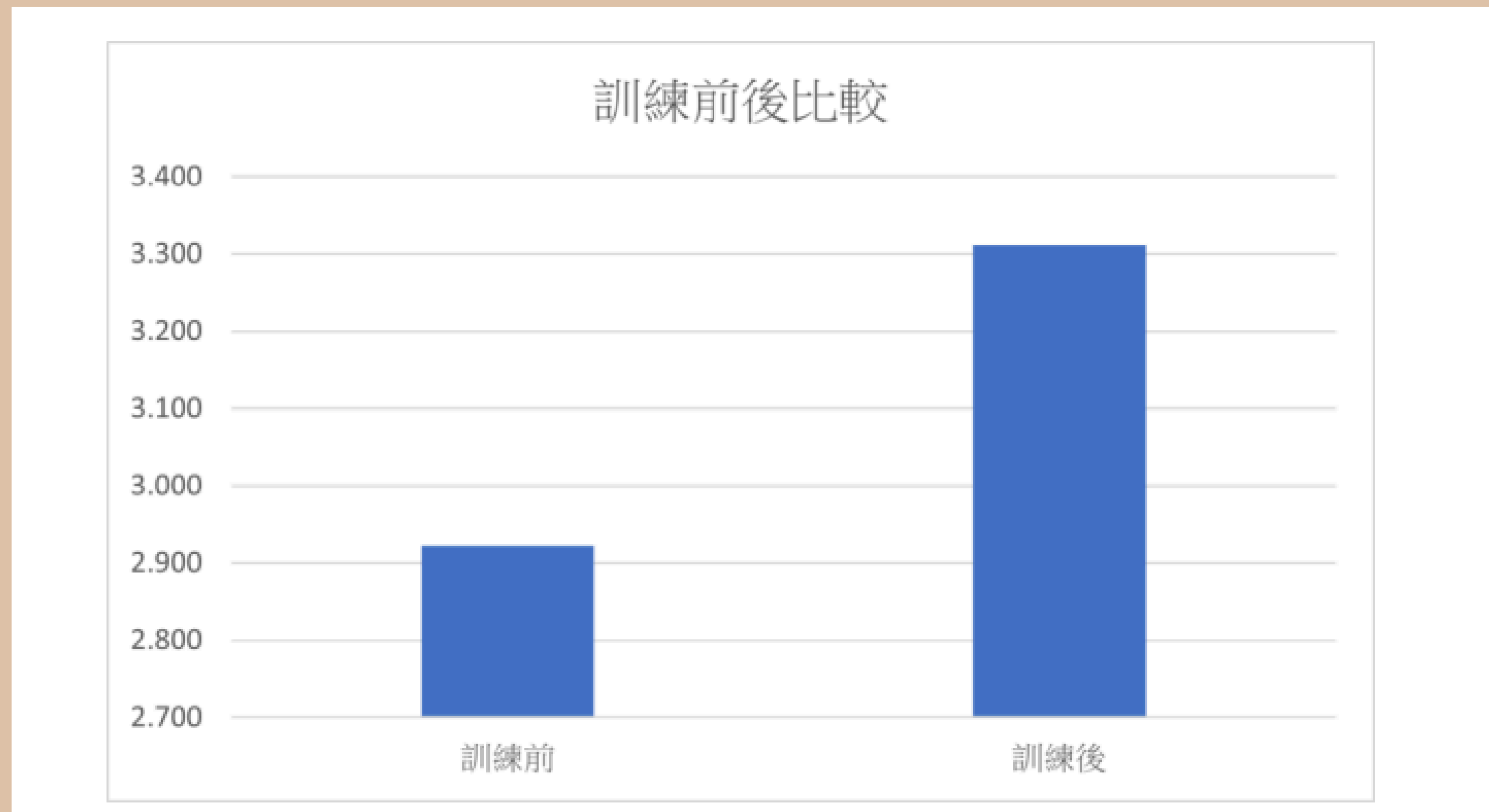
題目與答案節錄(二)

結論：

- 輸出資訊正確率有小部分提升
- 在正確的內容中混入了誤導性陳述(如資格條件與其他建議的部分)，可見仍需加強

	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。
--	--

成果數據比較(結論)

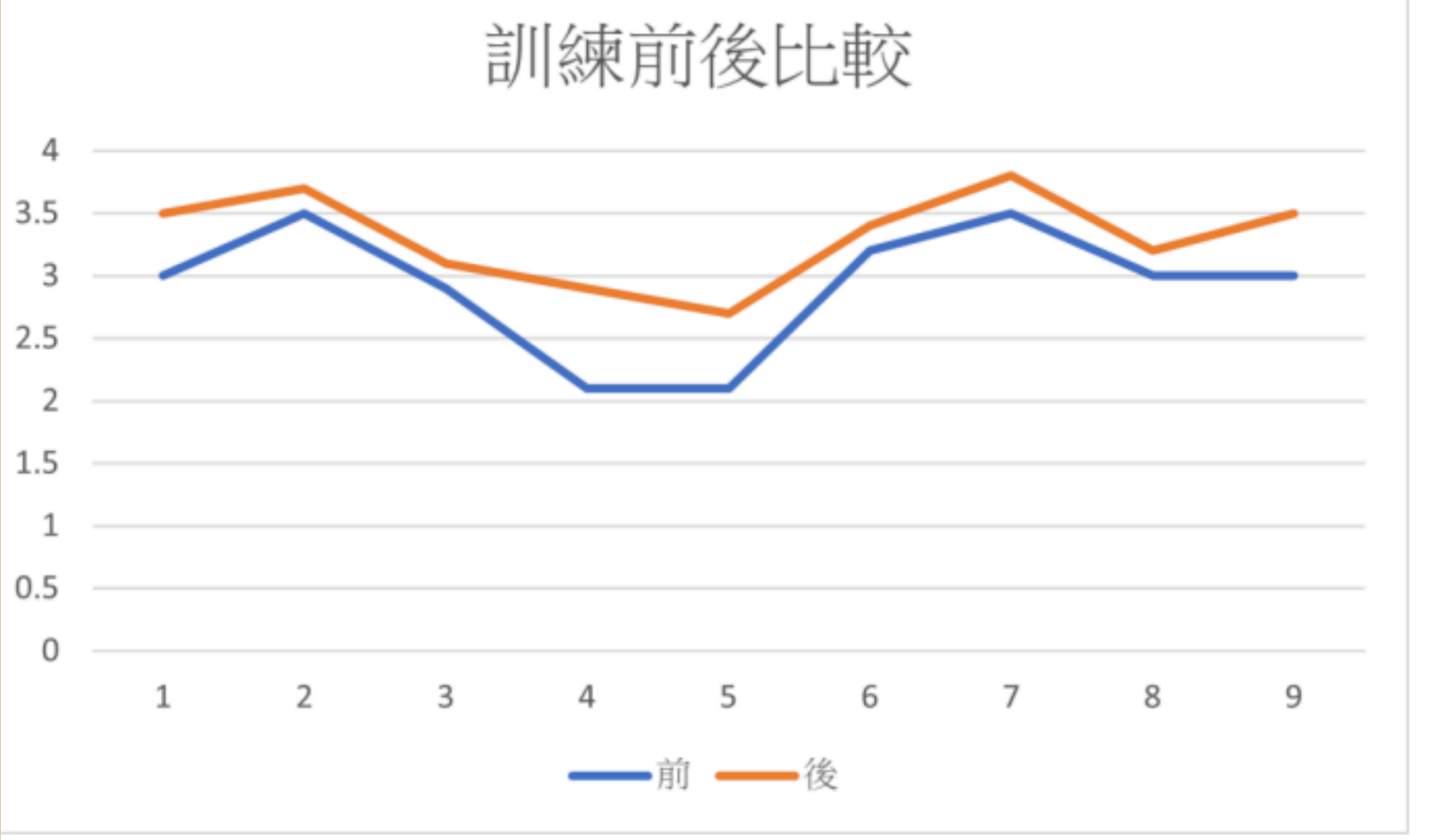


- 訓練後分數較訓練前提提升了約百分之 13，幅度較不明顯

成果數據比較(結論)

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



- RAG模型訓練在改善內容正確性方面有不小的效果
- 模型訓練對於虛假與誤導性陳述這部分並不強
- LLM語言表達能力無太大差異

感謝聆聽