# ChatGPT與DeepSeek能通過台灣律師及司法官考試嗎？初探大型語言模型對台灣法律的理解與知識儲備

李 芯*
林其叡**

## 壹、前言

自從ChatGPT於2022年底橫空出世，關於大型語言模型（Large Language Models, LLMs）究竟能力所及為何、是否能取代法律工作者、如何能協助法律工作者的工作效率等等討論不絕於耳[1]。2025年初，DeepSeek的發布再度震驚四界，論者開始比較其與ChatGPT之差異與優劣[2]。

大型語言模型於近年來迅速發展，尤其在自然語言處理與生成（Natural Language Processing, NLP）領域展現驚人的理解與推理能力。然而，儘管大型語言模型如GPT-4o與DeepSeek-R1在各項通用任務上有優異表現，對於「台灣法律」這一高度專業且在地化的知識體系，其實際掌握情況為何，尚缺乏系統性的檢驗[3]。

過往對大型語言模型在法律領域的應用，

＊本文作者係執業律師
＊＊本文作者係常在國際法律事務所律師
（本文投稿日為2025年3月20日，截稿日為2025年6月8日。作者感謝匿名審稿委員之寶貴意見。本文內容為筆者個人立場，不代表事務所立場，惟相關文責由作者自負。）

註1：我國討論例如：黃銘傑（2019），〈人工智慧發展對法律及法律人的影響〉，《月旦法學教室》，200期，第51-54頁；蘇南（2018），〈論人工智慧運用於律師服務的未來展望〉，《全國律師》，22卷6期，第40頁；葉于甄，〈當AI來襲，律師執業的機會與衝擊〉，《在野法潮》，https://dissent.tba.org.tw/special/3216/（最後瀏覽日：2025年6月8日）。

註2：例如文獻上有以程式編寫、思想產出、學習與研究方面比較兩者性能者，參Graham Fraser, *DeepSeek vs ChatGPT-how do they compare?*, BBC (Jan. 28, 2025), https://www.bbc.com/news/articles/cqx9zn27700o。然而，依筆者於2025年6月8日之搜尋結果，**尚無文獻探討ChatGPT及DeepSeek兩者在處理台灣法律問題上之性能或結果產出優劣之文獻**。

註3：關於大型語言模型對於台灣法規、台灣法律問題之表現，依筆者於2025年6月8日於法源法律網及SSRN之搜尋結果，僅有一篇2023年年初之文獻以GPT-4在我國律師考試中的表現，嘗試檢驗GPT-4此一大型語言模型對台灣法律之掌握程度。參Mark Shope, *GPT Performance on the Bar Exam in Taiwan（GPT在台灣專門職業及技術人員高等考試律師考試中的表現）*，SSRN (2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4394826.。

多集中於英文語境下的外國法律系統[4]。然而法律作為在地化之學門與專業，外國之人工智慧法律工具，在台灣之普及性與實用性均不足。因此，對於欲在台灣執業或學習、理解法律之人而言，了解大型語言模型是否能正確理解台灣法規、判斷法條適用，並作為工作輔助工具，成為需要探討的議題。

此外，不同法律專業科目（如民事法、刑事法、公法、商事法）涵蓋範圍廣泛，涉及大量專業術語與法理。我們欲進一步探討大型語言模型是否會在特定法律領域出現明顯短處，從而影響法律人應用AI輔助工具時的選擇與信賴度。

最後，關於時常被討論的「AI是否能取代法官」、「AI是否能取代律師」議題[5]，我們欲先從我國律師及司法官養成的第一關「司法官及律師考試」開始，探討大型語言模型在我國司法官及律師考試之表現，包含是否具備通過考試取得司法官及律師資格之能力。

因此，本研究希望能達成之目的為：

一、嘗試檢驗大型語言模型對於台灣法律的理解與知識儲備。

二、分析不同大型語言模型在各法律專科（民事法、刑事法、公法、商事法等）的表現差異，以找出潛在短處，方便後續增加針對短處之改進。

三、探討模型在準確率、速度與成本的綜合表現，協助法律專業人士選擇最適合輔助實務工作的大型語言模型（即最高CP值模型）。

透過上述分析，期望提供法律專業人士選用大型語言模型的實證參考，並為未來台灣法律AI工具的發展方向提供想法。

## 貳、大型語言模型於法律領域過往之表現

目前我國文獻對於人工智慧於法律之應用，依筆者歸納討論內容概可分為以下幾個主題類型[6]：(1)探討人工智慧於法律領域之可能應用方式與利弊者[7]；(2)探討人工智慧於特定法律程序（例如調解、刑事程序）中之可能應用方式與利弊者[8]；(3)探討如何使用人工智慧於法律領域者，包含其原則與方式[9]；(4)探

註4：例如除了通用之大型語言模型外，美國市場已發展出許多專為法律專業人員服務之人工智慧工具，例如：Harvey AI, CoCounsel, Clio Duo等。參*AI Tools for Lawyers: Improving Efficiency and Productivity in Law Firms, in* AI FOR LAW FIRMS: A COMPREHENSIVE GUIDE, CLIO, https://www.clio.com/resources/ai-for-lawyers/ai-tools-for-lawyers/ (last visited: June 8, 2025).

註5：*See supra* note 1.

註6：依據筆者於2025年6月8日以「人工智慧」、「大型語言模型」、「AI」等關鍵字於法源法律網之搜尋結果。

註7：例如：李函諭、莊弘鈺（2019），〈淺論人工智慧於法律服務業之應用〉，《全國律師》，23卷5期，第74-87頁；蘇南，前揭註1，第31-40頁；黃銘傑，前揭註1。

註8：例如：關於人工智慧於線上調解之應用，參：卓居昌、鍾從定（2024），〈當科技遇上法律：線上調解AI的應用探討〉，《萬國法律》，254期，第69-88頁；關於人工智慧於提升司法通譯能力之應用，參：陳雅齡（2024），〈臺灣司法通譯培訓之改進與AI應用〉，《真理法學論叢》，25期，第55-84頁；關於人工智慧於提升司法通譯能力之應用，參：李維宗，〈論AI犯罪之刑事責任與AI在刑事程序之運用〉，《科技法學論叢》，18期，第213-250頁。

討人工智慧於我國法律領域之表現者[10]。其中，第(4)類型為文獻最稀少者，本研究即旨在針對第(4)類型研究領域貢獻。

探討人工智慧於我國法律領域之表現之文獻幾希，唯一之學術文獻為學者Mark Shope於2023年的撰寫中論文（working paper）。該文中作者以了ChatGPT Plus的GPT-4模型為實驗對象，將111年律師及司法官考試「每題題目分別輸入對話框中」，並「紀錄答案（A、B、C或D）並且忽略ChatGPT4生成的其他內容」[11]。另外，該文「有時會向ChatGPT4解釋輸入內容為選擇題，共有4個選項，並且不需詳答」，且僅於其中一次問答，向ChatGPT4解釋應適用台灣法律[12]。最終，該文研究顯示GPT-4模型無法通過111年律師及司法官考試第一試[13]。換言之，本研究將是國內第一份以實驗方法展現大型語言模型能通過我國律師及司法官考試第一試之研究（如下所述）。

至於探討大型語言模型於各種考試之表現之外國文獻較多[14]，而關於律師考試之表現，最廣為引用之文獻為Katz et al.於2024年發布對於美國律師考試之研究[15]。Katz et al.對於GPT-4的測試如同美國律師考試（Uniform Bar Exam）分為兩部分：選擇題之multistate bar exam（MBE）及開放式、申論式試題之multistate essay exam（MEE）與multistate performance test（MPT）[16]。針對MBE，Katz et al.以標準化的指令（prompt）讓GPT-4進行完整考題作答，用自動化方式檢驗正確與否，並將作答資料儲存於API[17]。針對MEE與MPT，Katz et al.將考題之背景描述與參考資料先格式化為純文字版，之後將題目送至模型回答，並儲存答覆內容，之後由該文其中兩位法律背景的作者（一位教授及一位律師）進行評分[18]。依Katz et al.研究結果，GPT-4已能通過美國律師考試，並表現約相當於人類考生之90%百分位數[19]。學者Martínez後續仿照Katz et al.之方法，成功複製GPT-4之MBE測試分數，足以展現Katz et al.研究之準

註9：例如：王道維、邱筱涵、ChatGPT（2023），〈當ChatGPT來敲法官的門──淺談AI應用於司法審判的原則與方式〉，《當代法律》，18期，第6-28頁；錢世傑、鍾寧（2024），〈人工智慧與法律實踐：從ChatGPT建立審訊模擬系統談起〉，《當代法律》，25期，第88-98頁。

註10：例如：Shope, *supra* note 3.

註11：Shope, *supra* note 3, at 8.

註12：*Id.*

註13：*Id.*, at 11.

註14：例如Open AI曾於2023年表示GPT-4在各類考試之表現均優於GPT-3.5，包含通過美國律師考試，惟其關於美國律師考試之說明僅係引用Katz et al.之研究結果（參下註16）。*GPT-4*, OpenAI (Mar. 14, 2023), https://openai.com/index/gpt-4-research/#fn-1.

註15：Daniel Martin Katz et al., *GPT-4 Passes the Bar Exam*, 382 PHIL. TRANS. R. SOC. A. 1 (2024), https://doi.org/10.1098/rsta.2023.0254.

註16：*Id.*, at 3.

註17：*Id.*, at 4-5.

註18：*Id.*, at 5, 9.

註19：*Id.*, at 12.

確性；惟Martínez主張應以初次參與律師考試之人類考生之分數為比較基準，在此基準下GPT-4的UBE分數約低於人類考生之69%百分位數[20]。另外，Martínez之研究發現調整模型之溫度（temperature）設定對實驗結果並無顯著影響，惟若使用few-shot prompting之方法（亦即提供人工智慧良好的回答範本以協助人工智慧產生更高品質的答案）[21]，相較於zero-shot prompting將顯著提升人工智慧的答題表現[22]。

## 參、實驗方法

本研究決定與上述Shope、Katz et al.、Martínez之研究相同，使用對大型語言模型輸入指令（prompt）的方式，使大型語言模型回答律師考試選擇題之方式做為研究方法。而與Shope之方法不同，本研究藉由設定max token之方式確保大型語言模型之回答限於A、B、C、D選項，而避免人工篩選「無關」之答覆可能產生之偏誤。此外，借鏡Martínez之研究發現，本研究使用zero-shot測試方法，最小化指令內容對提升大型語言模型表現之影響，以求更精準呈現模型既有的能力。

基此，本研究選擇採用雲端API調用方式，讓大型語言模型回答關於112年、113年司法官及律師考試第一試之試題，並藉由檢視大型語言模型回答這些試題的成果，分析大型語言模型在民事法、刑事法、公法、商事法領域的表現。

### 一、實驗模型選擇

本研究選擇當時市場上最具代表性、性能領先的大型語言模型進行測試[23]，包含OpenAI於2024年發布之GPT-4o與中國公司深度求索（DeepSeek）於2025年發布之DeepSeek-R1，兩者皆為近期國際間重要且具代表性的模型。此二模型為目前關注度最高的大型語言模型之二，且支援完整API調用，具使用之可行性。如此另一方面具有避免本地端部署的高昂成本與技術門檻（如GPU設備需求、推理優化）之特性，亦符合多數法律人可能使用之場景。DeepSeek-R1及GPT-4o之功能比較，可參下表：

**表1：DeepSeek-R1及GPT-4o功能比較表**

| 功能 (Feature) | DeepSeek-R1 | GPT-4o |
|---|---|---|
| 支援的輸入上下文長度 | 128K tokens（約相當於64,000至85,000個中文字） | 128K tokens（約相當於64,000至85,000個中文字） |

註20：Eric Martínez, *Re-evaluating GPT-4's Bar Exam Performance*, ARTIF INTELL LAW (2024), https://doi.org/10.1007/s10506-024-09396-9.

註21：Jonathan H. Choi et al., *Lawyering in the Age of Artificial Intelligence*, 109 MINN. L. REV. 147, 201 (2024).

註22：Martínez, *supra* note 21.

註23：本研究於2025年3月間進行及投稿。

| 功能 (Feature) | DeepSeek-R1 | GPT-4o |
|---|---|---|
| 單次最大輸出長度 | 32K tokens（約相當於16,000至21,250個中文字） | 16.4K tokens（約相當於8200至10,890個中文字） |
| 是否為開源模型 | 是 | 否 |
| 發布日期 | 2025年1月21日 | 2024年8月6日 |
| 知識截止日期 (Knowledge Cut-off Date) | 不明 | 2023年10月 |
| API 提供商 (API Providers) | DeepSeek、HuggingFace | OpenAI、Azure OpenAI Service |
| 1M Tokens價格（Input / Output） | $0.55 / $2.19 | $2.50 / $10.00 |

資料來源：Compare GPT-4o vs DeepSeek-R1[24]

**二、測驗題庫來源與範圍**

本研究測試題庫來源為考選部公告之司法官及律師考試第一試考試試題與答案[25]。本研究測試範圍包含截至截稿為止最近期的二次司法官及律師考試，分別為112年度及113年度司法官及律師考試。

112年度及113年度司法官及律師考試第一試均包含4個試卷，共計15個法律科目：綜合法學（一）（憲法、行政法、國際公法、國際私法）（下稱「公法」試卷）、綜合法學（一）（刑法、刑事訴訟法、法律倫理）（下稱「刑事法」試卷）、綜合法學（二）（民法、民事訴訟法）（下稱「民事法」試卷）、綜合法學（二）（公司法、保險法、票據法、證券交易法、強制執行法、法學英文）（下稱「商事法」試卷）。112年度及113年度司法官及律師考試第一試各有300題，科目分佈如下：

**表2：司法官及律師考試第一試科目與題目數量分佈**

| 科目 | 題目數量 |
|---|---|
| 憲法 | 20 |
| 行政法 | 35 |
| 國際公法 | 10 |
| 國際私法 | 10 |
| 刑法 | 35 |
| 刑事訴訟法 | 25 |
| 法律倫理 | 15 |
| 民法 | 50 |
| 民事訴訟法 | 30 |
| 公司法 | 15 |
| 保險法 | 10 |

註24：*Compare GPT-4o vs DeepSeek-R1*, DocsBot, https://docsbot.ai/models/compare/gpt-4o/deepseek-r1 (last visited: Mar. 20, 2025).

註25：考選部，考畢試題查詢平台，https://wwwq.moex.gov.tw/exam/wFrmExamQandASearch.aspx（最後瀏覽日：2025年3月20日）。

| | |
|---|---|
| 票據法 | 10 |
| 證券交易法 | 10 |
| 強制執行法 | 10 |
| 法學英文 | 15 |
| 總計： | 300 |

資料來源：考選部考畢試題查詢平台

司法官及律師考試第一試試卷內容均為四選一選擇題，每題2分，總分600分。本研究將模仿真實考試情境，對於兩年度司法官及律師考試第一試總計600題題目，要求大型語言模型選擇正確或錯誤之陳述。

**三、作答規範**

本研究採用Zero-shot（0-shot）測試方式，意指不提供任何範例題或提示範例，直接要求模型針對題目進行回答[26]。換言之，模型僅透過內部既有的知識庫與推理能力，來回應真實考題。如此可避免模型受到先前回答內容之影響，更能直接呈現模型既有的能力。

在Prompt（指令）的設計，所有模型均以相同提示進行zero-shot作答，內容如下：

*請回答以下問題：*

*【請務必根據台灣法條回答，不要編造或推測】*

*請僅回應A、B、C或D，不要提供額外解釋。*

此外，模型的生成參數統一設置如下：

- Max tokens[27]=2：確保模型的輸出僅限於選擇A、B、C或D，不會產生多餘的解釋或內容。
- Temperature[28]=0：設定模型的temperature為0，確保獲得較高度之確定性（deterministic），使其更能每次輸出相同的答案，避免隨機性影響結果的穩定性。

**四、分析標準**

- 正確率計算：以考選部公告之司法官及律師考試第一試標準答案比對，計算各模型正確率。
- 錯題分析：統計雙方錯誤交集（both wrong）、單邊錯誤（only GPT-4o wrong、only DeepSeek-R1 wrong），並分析雙方是否傾向選擇相同錯誤選項。

註26：Takeshi Kojima et al., *Large Language Models are Zero-Shot Reasoners*, 36th Conference on Neural Information Processing Systems (NeurIPS 2022), https://arxiv.org/pdf/2205.11916.

註27：即完成一次生成（run）所需的最大token數。OpenAI Platform, *Create Thread and Run*, https://platform.openai.com/docs/api-reference/runs/createThreadAndRun (last visited June 21, 2025).

註28：Temperature是大型語言模型的一個參數，用來控制選擇token時的隨機性。Temperature為0時，透過始終選擇機率最高的token，可以產生高度確定性的輸出，從而最大程度地減少運行過程中的差異。反之，較高的temperature會為產生的反應帶來更大的可變性和創造性。Jihong Zhang et al, *Leveraging Interview-Informed LLMs to Model Survey Responses: Comparative Insights from AI-Generated and Human Data*, https://arxiv.org/pdf/2505.21997 (last visited June 21, 2025).

・速度與成本：完整紀錄每題作答所需時間（latency），分析各模型平均作答速度（秒/題）、總花費。

## 肆、實驗結果

司法官及律師考試第一試將篩選出前三分之一之應考生，通過者方能參與該年度司法官及律師考試第二試。依據考選部公告結果，112年、113年律師考試第一試及格門檻分別為374分[29]與354分[30]，112年、113年司法官考試第一試及格門檻分別為384分[31]與362分[32]。

依據本研究結果，以總分而言，GPT-4o在112年、113年律師及司法官第一試分別獲得總分400分、368分；DeepSeek-R1在112年、113年律師及司法官第一試分別獲得總分460分、460分。**兩者在兩年度均超過該年度律師及司法官考試第一試及格門檻**。若以總分而言，GPT-4o在112年律師、司法官考生中將排名前19.41%[33]、前24.00%[34]，在113年律師、司法官考生中將排名前25.01%[35]、前29.57%[36]。DeepSeek-R1在112年律師、司法官考生中將排名前1.97%[37]、前2.71%[38]，在113年律師、司法官考生中將排名前0.50%[39]、前0.65%[40]。

### 一、各試卷測試結果

GPT-4o及DeepSeek-R1回答112年、113年律師及司法官第一試之各科測試結果，如下表所示。GPT-4o在各科均可以答對約60-70%的題目，DeepSeek-R1在各科均可以答對約70-80%的題目，且在各「試卷」間表現並無顯著差異：

註29：考選部，112年專門職業及技術人員高等考試律師考試第一試應考人成績統計表，https://wwwc.moex.gov.tw/main/controls/wHandEditorExtend_File.ashx?Fun=Property&menu_id=335&item_id=6288&file_id=14596（最後瀏覽日：2025年3月20日）。

註30：考選部，113年專門職業及技術人員高等考試律師考試第一試應考人成績統計累計表，https://wwwc.moex.gov.tw/main/controls/wHandEditorExtend_File.ashx?Fun=Property&menu_id=335&item_id=6829&file_id=18027（最後瀏覽日：2025年3月20日）。

註31：考選部，112年公務人員特種考試司法官考試第一試應考人成績統計累計表，https://wwwc.moex.gov.tw/main/controls/wHandEditorExtend_File.ashx?Fun=Property&menu_id=335&item_id=6288&file_id=14595（最後瀏覽日：2025年3月20日）。

註32：考選部，113年公務人員特種考試司法官考試第一試應考人成績統計累計表，https://wwwc.moex.gov.tw/main/controls/wHandEditorExtend_File.ashx?Fun=Property&menu_id=335&item_id=6829&file_id=18026（最後瀏覽日：2025年3月20日）。

註33：見前揭註30。

註34：見前揭註31。

註35：見前揭註32。

註36：見前揭註33。

註37：見前揭註30。

註38：見前揭註31。

註39：見前揭註32。

註40：見前揭註33。

表3：GPT-4o及DeepSeek-R1之112年度及113年度第一試各試卷測試結果

| 112年度 | 綜合法學(一)（公法） | 綜合法學(一)（刑事法） | 綜合法學(二)（民事法） | 綜合法學(二)（商事法） | 總分／每試卷所需時間 | 該年度及格門檻／應試人員比例 |
|---|---|---|---|---|---|---|
| GPT-4o | 正確率：66.67%（50/75題） | 正確率：65.33%（49/75題） | 正確率：65.00%（52/80題） | 正確率：0.00%（49/70題） | 400分<br>平均每試卷程式運行時間：36.3秒 | 律師門檻：374分<br>司法官門檻：384分<br>律師：前19.41%<br>司法官：前24.00% |
| DeepSeek-R1 | 正確率：76.00%（57/75題） | 正確率：77.33%（58/75題） | 正確率：73.75%（59/80題） | 正確率：80.00%（56/70題） | 460分<br>平均每試卷程式運行時間：6228秒（約104分鐘） | 律師門檻：374分<br>司法官門檻：384分<br>律師：前1.97%<br>司法官：前2.71% |

| 113年度 | 綜合法學(一)（公法） | 綜合法學(一)（刑事法） | 綜合法學(二)（民事法） | 綜合法學(二)（商事法） | 總分／每試卷所需時間 | 該年度及格門檻／應試人員比例 |
|---|---|---|---|---|---|---|
| GPT-4o | 正確率：62.67%（47/75題） | 正確率：65.33%（49/75題） | 正確率：57.50%（46/80題） | 正確率：60.00%（42/70題） | 368分<br>平均每試卷程式運行時間：53.5秒 | 律師門檻：354分<br>司法官門檻：362分<br>律師：前25.01%<br>司法官：前29.57% |
| DeepSeek-R1 | 正確率：81.33%（61/75題） | 正確率：78.67%（59/75題） | 正確率：76.25%（61/80題） | 正確率：70.00%（49/70題） | 460分<br>平均每試卷程式運行時間：5525秒（約92分鐘） | 律師門檻：354分<br>司法官門檻：362分<br>律師：前0.50%<br>司法官：前0.65% |

資料來源：作者自製、考選部應考人成績統計表

另從112年、113年兩個年度之試卷「難度」而言，若考量樣本數龐大（全程到考人數：112年律師9864人[41]、113年律師10232人[42]、112年司法官8866人[43]、113年司法官9014人[44]），可合理假設兩年度律師及司法官第一試考試之人類應考生之作答能力並無顯著差異。又由112年律師及司法官第一試及格門檻高於113年及格門檻，可合理假設113年之律師及司法官第一試試卷「難度」高於112年之試卷。如此「難度」差異在GPT-4o之測試結果較為明顯，即GPT-4o在113年試卷之表現相較112年總分下降32分、正確率總體下降近10%。上開結果在DeepSeek-R1並不可見，蓋DeepSeek-R1在兩年度試卷獲得了相同的總分，故本研究中DeepSeek-R1之表現並未受兩年度試卷「難度」差異影響。

### 二、各科目測試結果

若從DeepSeek-R1及GPT-4o兩者在112年、113年各科目測試結果綜合而言，可從本研究

註41：見前揭註30。
註42：見前揭註31。
註43：見前揭註32。
註44：見前揭註33。

結果觀察如下特徵：

（一）DeepSeek-R1及GPT-4o在法學英文的表現最佳，均無任何錯誤。如此結果反應大型語言模型在處理語言類型問題之擅長。

（二）DeepSeek-R1及GPT-4o表現次佳的科目應為法律倫理、國際公法、憲法。DeepSeek-R1及GPT-4o在112年、113年法律倫理試題之正確率均達80%以上。除GPT-4o在113年試題之表現以外，DeepSeek-R1及GPT-4o在國際公法試題之正確率均達80%以上。DeepSeek-R1及GPT-4o在112年、113年憲法試題之正確率均達80%以上。如此結果可能是因為這些科目較無地域性差異，而在各司法管轄區均有共通或類似之原則。

（三）DeepSeek-R1及GPT-4o在刑法相較於其他科目表現均不佳，DeepSeek-R1在112年、113年刑法正確率分別僅63%、80%。

（四）DeepSeek-R1及GPT-4o在票據法相較於其他科目表現均不佳，兩者在112年、113年測驗中正確率均未超過60%。

（五）GPT-4o在保險法表現不佳，112年、113年正確率分別僅有30%及40%；然而DeepSeek-R1並無相同問題，兩年正確率分別仍有70%及80%。

（六）DeepSeek-R1及GPT-4o回答112年、113年兩年度之強制執行法表現差異大，兩者回答112年試題正確率分別有90%及80%，屬於各科目中正確率偏高者；然而兩者回答113年試題正確率分別僅有40%及50%，屬於各科目中正確率偏低者。

（七）DeepSeek-R1在各科表現上普遍優於GPT-4o，除113年強制執行法外，其餘科目DeepSeek-R1正確率均高於GPT-4o。

若單就DeepSeek-R1在各科表現在112年、113年各科目測試結果而言，可從本研究結果觀察如下特徵：

（一）除113年強制執行法外，其餘科目DeepSeek-R1正確率均高於60%，顯示其對於台灣法律的理解與知識儲備佳。

（二）DeepSeek-R1在法學英文正確率為100%，在法律倫理與國際公法兩科目於112年、113年度之表現均超過80%，且在兩年度科目正確率排行中均為前三高之科目，故可以認為DeepSeek-R1在法學英文、法律倫理與國際公法三科均表現較佳。

（三）DeepSeek-R1在112年、113年票據法正確率均僅有60%，分別為該年度科目正確率排行中倒數第一、倒數第二之科目，可知DeepSeek-R1在票據法表現較差。

若單就GPT-4o在各科表現在112年、113年各科目測試結果而言，可從本研究結果觀察如下特徵：

（一）GPT-4o在法學英文正確率為100%，顯示其對語言類型問題之擅長。

（二）GPT-4o在112年、113年保險法正確率分別僅有30%、40%，且均為該年度科目正確率排行中倒數第一之科目，可知GPT-4o在回答保險法問題上還有很大的進步空間。

表4：GPT-4o及DeepSeek-R1之112年度及113年度第一試各科目測試結果

| | 總題數(年) | GPT-4o正確率(113年) | DS-R1正確率(113年) | GPT-4o 正確率(112年) | DS-R1正確率(112年) |
|---|---|---|---|---|---|
| 刑法 | 35 | 51% | 80% | 60% | 63% |
| 刑事訴訟法 | 25 | 76% | 76% | 60% | 84% |
| 法律倫理 | 15 | 80% | 80% | 87% | 100% |
| 民法 | 50 | 64% | 72% | 66% | 74% |
| 民事訴訟法 | 30 | 47% | 83% | 63% | 73% |
| 憲法 | 20 | 85% | 90% | 70% | 70% |
| 行政法 | 35 | 54% | 80% | 60% | 80% |
| 國際公法 | 10 | 40% | 80% | 90% | 90% |
| 國際私法 | 10 | 70% | 70% | 60% | 60% |
| 公司法 | 15 | 53% | 67% | 73% | 73% |
| 保險法 | 10 | 40% | 80% | 30% | 70% |
| 票據法 | 10 | 60% | 60% | 50% | 60% |
| 證券交易法 | 10 | 40% | 60% | 70% | 80% |
| 強制執行法 | 10 | 50% | 40% | 80% | 90% |
| 法學英文 | 15 | 100% | 100% | 100% | 100% |

資料來源：作者自製

另有鑒於各年度、各科題目的難度不一，上開結果之解讀亦有其限制。意即，由於目前不存在各年度、各科題目間難度差異之客觀標準，吾人難以確實得知大型語言模型在這些科目的表現優劣，是因為大型語言模型在這些科目比較擅長，還是這些科目「題目」難度本來就比較高。此外，由於這僅是針對兩年度試題之解讀，對於瞭解大型語言模型在這些科目之掌握程度亦有限制。

### 三、錯題數分析

比較DeepSeek-R1及GPT-4o答錯之題目內容，可以發現DeepSeek-R1及GPT-4o同時答錯之題目，佔DeepSeek-R1答錯題目之比例高。換言之，僅有DeepSeek-R1答錯，而GPT-4o答對之題目數量不多。而由於GPT-4o答錯之題目總數較多，僅有GPT-4o答錯之題目數量亦較多。兩者的錯題數統計如下圖：

此外，觀察在DeepSeek-R1及GPT-4o均答錯時雙方選擇同一錯誤答案之比例，於各試卷間約54-83%不等，比例相較於單純亂數選擇之機率較高。因此，可以推論兩個大型語言模型均使用類似的判斷方式進行答案之選擇，故當該判斷方式錯誤時，兩者傾向於選擇同一個錯誤答案。此外，雙方皆答錯的題目，仍有高比例選擇相同錯誤答案，亦顯示兩大大型語言模型在某些台灣法律細節上存在共通性誤區，這些部分應成為後續針對大型語言模型訓練資料之重點補強對象。

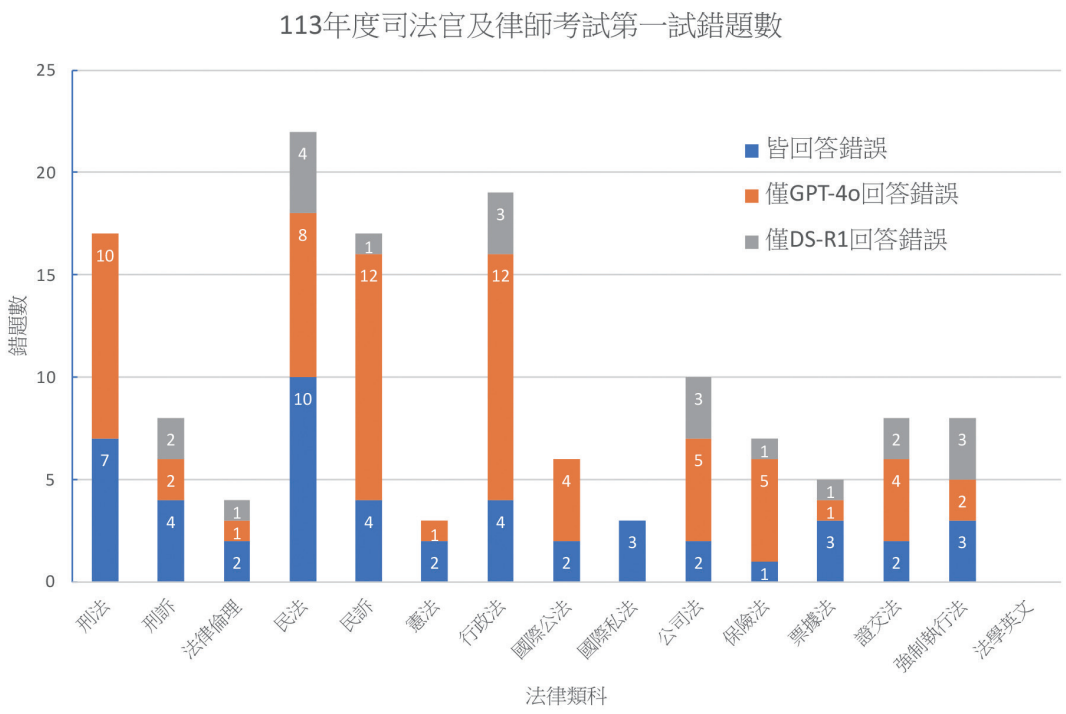


**圖1：113年度司法官及律師考試第一試錯題數**

資料來源：作者自製

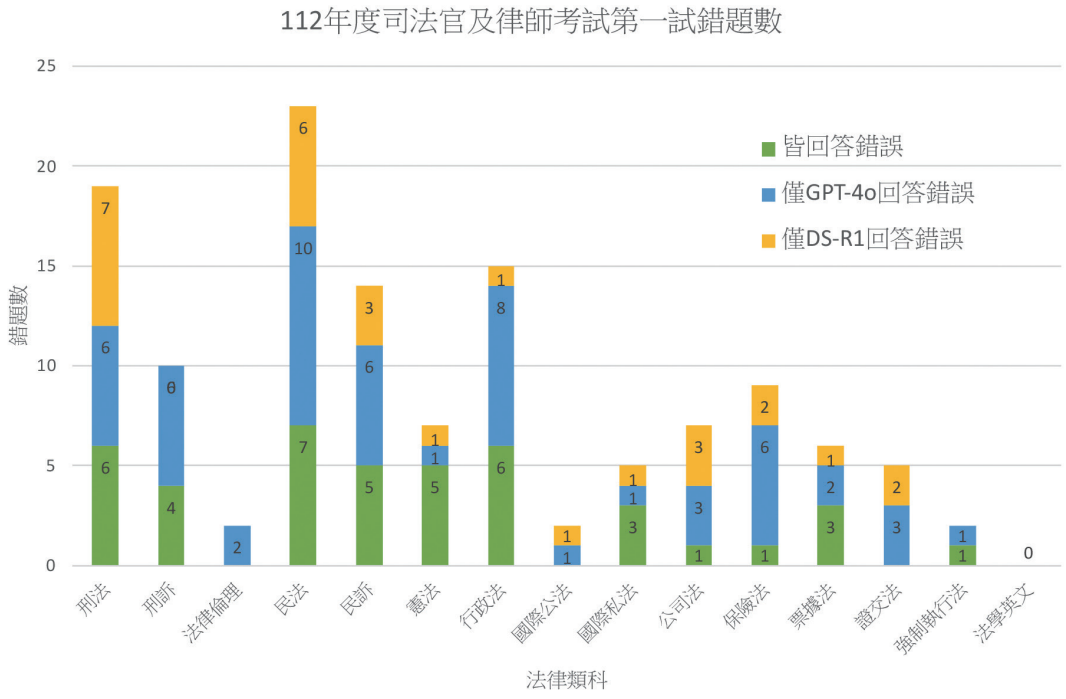


**圖2：112年度司法官及律師考試第一試錯題數**

資料來源：作者自製

表5：GPT-4o及DeepSeek-R1之112年度及113年度第一試錯題數及雙方都錯時選同一錯誤答案比例

| 試卷（113 年度） | 雙方都答錯 | 僅有GPT-4o答錯 | 僅有DeepSeek-R1答錯 | 雙方都錯時選同一錯誤答案比例 |
|---|---|---|---|---|
| 綜合法學(一)（公法） | 11題（2憲法、4行政法、2國際公法、3國際私法） | 17題（1憲法、12行政法、4國際公法） | 3題（3行政法） | 54.55% |
| 綜合法學(一)（刑事法） | 13題（7刑法、4刑事訴訟法、2律師倫理） | 13題（10刑法、2刑事訴訟法、1律師倫理） | 3題（2刑事訴訟法、1律師倫理） | 76.92% |
| 綜合法學(二)（民事法） | 14題（10民法、4民事訴訟法） | 20題（8民法、12民事訴訟法） | 5題（4民法、1民事訴訟法） | 57.14% |
| 綜合法學(二)（商事法） | 11題（2公司法、1保險法、3票據法、2證券交易法、3強制執行法） | 17題（5公司法、5保險法、1票據法、4證券交易法、2強制執行法） | 10題（3公司法、1保險法、1票據法、2證券交易法、3強制執行法） | 63.64% |

| 試卷（112年度） | 雙方都答錯 | 僅有GPT-4o答錯 | 僅有DeepSeek-R1答錯 | 雙方都錯時選同一錯誤答案比例 |
|---|---|---|---|---|
| 綜合法學(一)（公法） | 14題（5憲法、6行政法、3國際私法） | 11題（1憲法、8行政法、1國際公法、1國際私法） | 4題（1憲法、1行政法、1國際公法、1國際私法） | 78.57% |
| 綜合法學(一)（刑事法） | 10題（6刑法、4刑事訴訟法） | 18題（6刑法、8刑事訴訟法、2律師倫理） | 7題（7刑法） | 60% |
| 綜合法學(二)（民事法） | 12題（7民法、5民事訴訟法） | 16題（10民法、6民事訴訟法） | 9題（6民法、3民事訴訟法） | 75% |
| 綜合法學(二)（商事法） | 6題（1公司法、1保險法、3票據法、1強制執行法） | 15題（3公司法、6保險法、2票據法、3證券交易法、1強制執行法） | 8題（3公司法、2保險法、1票據法、2證券交易法） | 83.33% |

資料來源：作者自製

**四、題目長度與正確率**

本研究所使用之兩款大型語言模型GPT-4o或DeepSeek-R1對於其所輸入上下文長度（context length）皆支援高達128,000個tokens，約等同於64,000個至85,000個中文字。相較之下，112年度及113年度司法官及律師考試第一試之所有選擇題，其單題題目平均長度為74字，選項平均長度為137字，合計（題目加上選項）之平均總長度為211字，最長的單題題目與選項合計長度為729字。此外，作答時所額外加入之提示詞（prompt）長度為50字。所有考題之題目、選項及提示詞總計，皆遠低於GPT-4o與DeepSeek-R1支援之最大上下文長度。因此，本研究過程中並無因題目長度過長而影響模型處理能力之情形，兩模型皆能完整接收並處理所有題目內容。

至於題目長度是否會影響語言模型回答問題之正確率，本研究將題目總長度分為短（150字以下）、中（150-250字）、長（250字以上）三組討論，無論GPT-4o或DeepSeek-R1，其在不同長度題目（短、中、長）上的正確率皆並無顯著差異。因此，本研究並未發現題目長度顯著影響模型表現的證據。

**表6：GPT-4o及DeepSeek-R1在不同題目長度分組之正確率**

| 題目加上選項總長度分組 | 題數 | DeepSeek-R1正確率 | GPT-4o正確率 |
|---|---|---|---|
| 短 (150字以下) | 127 | 78.74% | 65.91% |
| 中 (150-250字) | 314 | 76.75% | 63.23% |
| 長 (250字以上) | 159 | 75.47% | 68.66% |

資料來源：作者自製

一般而言，題目長度越長，對考生來說所需的閱讀時間越長，進而會使作答時間拉長。然而，對於大型語言模型而言，是否呈現相同現象，仍值得進一步探討。為檢視此問題，本研究分析題目總字數（包含題幹與選項）與模型作答時間之關聯性。

**表7：GPT-4o及DeepSeek-R1在不同題目長度分組之作答時間**

| 題目加上選項總長度分組 | 題數 | DeepSeek-R1平均每題作答時間 | GPT-4o平均每題作答時間 |
|---|---|---|---|
| 短 (150字以下) | 127 | 71.85秒 | 0.57秒 |
| 中 (150-250字) | 314 | 92.73秒 | 0.59秒 |
| 長 (250字以上) | 159 | 94.96秒 | 0.61秒 |

資料來源：作者自製

從單純作答時間長短而言，DeepSeek-R1及GPT-4o之平均每題作答時間，確實均因題目長度變長而作答時間更長。然而，從相關係數而言，**題目長度與大型語言模型作答時間之間呈現低度相關**。以DeepSeek-R1為例，題目總長度與作答時間之相關係數為0.058310，顯示幾乎不具相關性；GPT-4o之相關係數則為0.040426，同樣顯示相關性極低。換言之，雖然人類考生在面對長題時通常需要花費較多時間閱讀與理解，然而對於大型語言模型而言，題目字數並非影響其作答速度的主要因素。可能原因在於，大型語言模型在處理輸入文本時，並不「閱讀」文本，而是直接以併行（parallel）的方式進行內部嵌入（embedding）與推理運算，因此即便字數略長，對其整體處理時間影響有限。因此大型語言模型的回應速度可能更取決於API基礎延遲等雜訊，不見得單純反映題目本身難易或長短。

**表8：GPT-4o及DeepSeek-R1題目總字數與作答時間之相關係數**

| | 題目總字數 | DeepSeek-R1作答時間 | GPT-4o作答時間 |
|---|---|---|---|
| 題目總字數 | 1.00 | 0.003206 | 0.064 |
| DeepSeek-R1作答時間 | 0.003206 | 1.00 | -0.019 |
| GPT-4o作答時間 | 0.064 | -0.019 | 1.00 |

資料來源：作者自製

## 五、作答時間與正確率

### （一）每題作答時間

除了整份試卷之作答時間外，本實驗亦有紀錄每題的作答時間。司法官及律師考試第一試各試卷考試時間長度在80-100分鐘不等，綜合法學（一）試卷（公法）共75題之作答時間為90分鐘、綜合法學（一）試卷（刑事法）共75題之作答時間為90分鐘、綜

合法學（二）試卷（民事法）共80題之作答時間為100分鐘、綜合法學（二）（商事法）試卷共70題之作答時間為80分鐘，平均提供每題的作答時間為68.57-75.00秒。GPT-4o每題花費約1秒鐘作答，而DeepSeek-R1卻要花費約89鐘。

**（二）分科作答時間**

比較DeepSeek-R1及GPT-4o之答題表現，DeepSeek-R1在司法官及律師考試第一試考題上的準確率領先GPT-4o約10%以上，在各個法領域均展現穩定優勢。然而，GPT-4o在速度上展現壓倒性優勢，兩者同樣作答一卷司法官及律師考試第一試考題試卷，GPT-4o在每個類科（70-80題不等）所花費之作答時間平均僅需不到1分鐘，DeepSeek-R1卻平均需要花費大於90分鐘。因此，若考量作答時間，由於司法官及律師考試第一試各試卷考試時間長度在80-100分鐘不等（即提供考生平均每題68.57-75.00秒時間），DeepSeek-R1不一定能在時限內完成全部題目，故不一定能「通過」司法官及律師考試第一試。綜上所述，若僅考慮正確率，DeepSeek-R1無疑更能提供使用者正確率更高的答案；但若從考慮實務應用（例如律師快速提供法律意見、審查契約等）之角度而言，GPT-4o更具即時輔助潛力。

**（三）作答時間與正確率之相關性**

至於作答之時間是否會影響正確率，從相關係數來看，GPT-4o作答時間與正確率的相關性為-0.0308，顯示兩者關聯性不強。此外，從分群平均來看，GPT-4o在答對題目時平均花費0.583806秒，而答錯題目時平均花費0.613278秒。顯示在本資料集中，GPT-4o並不會花較長時間思考正確答案。

從GPT-4o作答時間分布圖（Histogram）可知GPT-4o在答對與答錯題目時的作答時間分布比例，多數題目的作答時間集中於0.5至1秒區間，其中超過80%的題目均於1秒內完成，顯示GPT-4o具備極高的反應速度。無論是答對或答錯，兩者的分布型態幾乎重疊，顯示GPT-4o在「回答錯誤」時，並未花費顯著較長的時間進行思考或推理。例如80.6%的答錯題目於0.5-1秒內完成，與79.9%的答對題目相差無幾，佐證兩者無明顯差異。

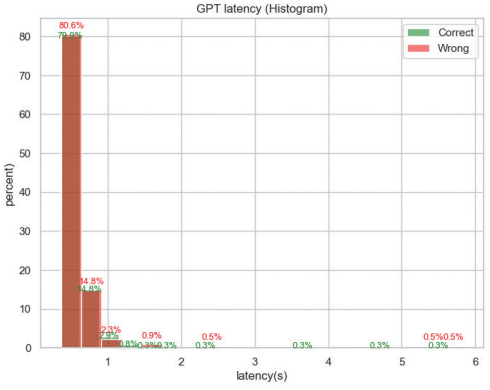


圖3：GPT-4o作答時間分布圖（Histogram）
資料來源：作者自製

以下GPT-4o作答時間之正確與否箱型圖（Boxplot）分為「答對」與「答錯」兩群。兩組資料的中位數、上下四分位數幾乎重疊，顯示作答時間之分布趨勢一致。答錯題目雖出現少數異常值，但整體分布差異極小。

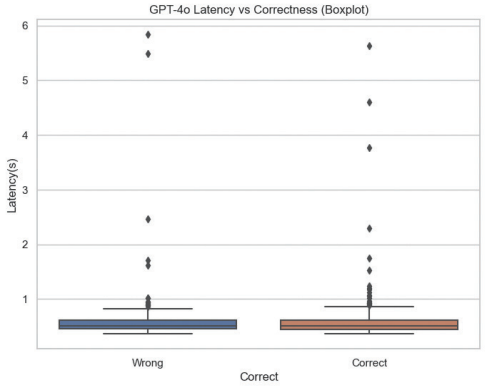


圖4：GPT-4o作答時間之正確與否箱型圖（Boxplot）
資料來源：作者自製

而就DeepSeek-R1而言，從相關係數來看，DeepSeek-R1作答時間與正確率的相關性為-0.2241。此外，從分群平均來看，DeepSeek-R1在答對題目時平均花費74.52秒，而答錯題目時平均花費136.48秒。顯示在本資料集中，DeepSeek-R1不但不會花較長時間思考正確答案，更有可能是思考時間越長，錯誤的可能性越大。從DeepSeek-R1作答時間分布圖（Histogram）可知，DeepSeek-R1僅約10%的正確題目沒有在160秒內完成，但有超過1/4的錯誤題目在160秒內完成。此外，不論答案正確或錯誤，DeepSeek-R1有許多時候使用超出律師及司法官考試第一試期望的68.57-75.00秒時間作答。

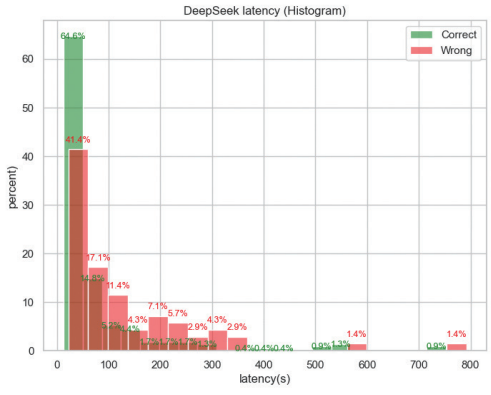


圖5：DeepSeek-R1作答時間分布圖（Histogram）
資料來源：作者自製

從DeepSeek-R1作答時間之正確與否箱型圖（Boxplot）可知，DeepSeek-R1在作答錯誤時，整體作答時間明顯較長。具體而言，錯誤答案的作答時間中位數（約70秒）高於正確答案之中位數（約40秒），且錯誤作答的四分位距範圍較大。換言之，作答錯誤時可能伴隨著作答時間長、波動大的特性。

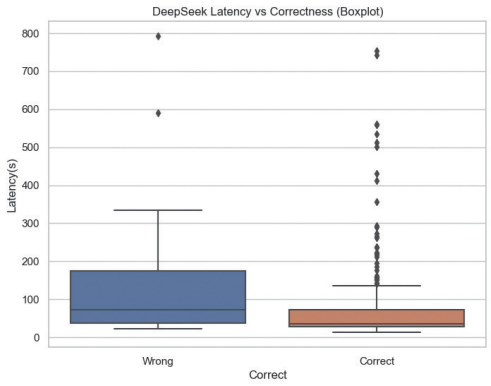


圖6：DeepSeek-R1作答時間之正確與否箱型圖（Boxplot）
資料來源：作者自製

## 六、Token數與花費金額

為進一步評估大型語言模型在法律專業應用上的成本效益（Cost-Effectiveness），本文記錄並分析兩款模型完成所有考題所產生的token數量與實際花費金額。特別值得注意的是，所有請求均設定max_tokens=2限制輸出長度，確保所有回答僅包含選項（A、B、C、D），避免不必要之文字生成。

**表9：GPT-4o及DeepSeek-R1花費之token數與花費金額**

| 模型 | 總Token數 (Input + Output) | 總花費金額 | 幣別 |
|---|---|---|---|
| DeepSeek-R1 | 1,212,305 | 10.51 | CNY |
| GPT-4o | 220,647 | 0.40 | USD |

資料來源：作者自製

從表中可以看出，DeepSeek-R1在處理完整試題過程中，共消耗1,212,305 tokens，總費用為人民幣10.51元；相較之下，GPT-4o僅消耗220,647 tokens，總費用為0.4美元。

### （一）Token數量差異

在本研究中，DeepSeek-R1所消耗的總token

數高達1,212,305 tokens，而GPT-4o則為220,647 tokens。與GPT-4o相比，DeepSeek-R1約為其5.5倍。針對此一現象，近期有使用者於社群指出，DeepSeek-R1的token計算包含大量「推理過程（reasoning tokens）」，這些推理token為模型在生成最終答案前進行內部運算所耗用[45]。DeepSeek-R1的推理token數可達實際回答（completion tokens）的2至5倍，這意味著，儘管最終輸出只有數個字元，背後卻隱含大量模型內部推理所消耗的token。因此，僅管DeepSeek-R1每一百萬個token的價格約為GPT-4o的五分之一，此一隱藏成本亦應成為考量之重點。

**（二）費用比較**

以總金額而言，DeepSeek-R1人民幣10.51元，GPT-4o美金0.40元。若折算匯率（假設1美元大約等於7.2元人民幣），則約為DeepSeek-R1新台幣46元，GPT-4o約新台幣13元。GPT-4o在此次測驗中，無論token數或總費用均明顯低於DeepSeek-R1，展現出更具經濟效益的一面。

## 伍、研究限制

本研究以讓大型語言模型回答112年、113年律師及司法官考試第一試試題進行，顯示大型語言模型回答內容均通過該年度律師及司法官及格門檻。然而，本研究至少有以下限制，仍有待後續研究發展：

**一、選擇題並未涵蓋律師及司法官考試第二試及第三試範圍，難以確認大型語言模型是否有通過該等考試的能力**

除第一試選擇題以外，律師考試之考生尚須通過第二試之申論式試題，在第二試考生中獲得前三分之一成績且總分超過400分，方能通過該年度律師考試。而司法官考試額外還有第三試口試須通過。本研究僅針對第一試選擇題作答，故無法瞭解大型語言模型在律師及司法官考試第二試及第三試之表現。

從實驗設計觀點而言，大型語言模型固然能回答第二試申論式試題，然而如何評量作答成果，將有疑義。即使考選部於試後會公告「司法官、律師考試第二試法律專業科目評分要點」，然而是否實際上每個試卷都客觀地依照該評分要點評分，無從得知。縱使是人類考生，亦常有面臨拿到分數後不明所以、不知評分標準何在之疑惑。而司法官第三試是口試，更難以有具體評分標準，甚至評分委員亦不一定是完全針對「回答內容」本身判斷。因此，如何測試大型語言模型是否能完整通過台灣的律師及司法官考試（包含第二試、第三試），將十分具有挑戰性。

**二、選擇題與法律實務的落差**

律師及司法官第一試選擇題僅能測試考生

註45：Zhu Liang, *"DeepSeek R1 is better and cheaper." -Wrong*., X (Jan. 25, 2025), https://x.com/paradite_/status/1883049686246031463?fbclid=IwY2xjawJCiz9leHRuA2FlbQIxMAABHT25X2_jdvV3f_f1ZbBEOMgRxwhtJpTAqcJJDZKQ1e1v1QJgFeehAZ-bpw_aem_wDoOjzTvOc2wqLsmJRVtqg.

對法律條文與理論知識的基礎理解，但法律實務涉及案件分析、證據評估、談判技巧、倫理判斷等能力，這些能力無法僅透過選擇題測試出來。選擇題亦無涉創新性論述，例如對現有法律進行思辯、產生創新的解釋或提出有說服力的法律論點。此外，法律實務工作許多時候是處理「人」的問題，包含如何察言觀色、如何與他人互動等等能力，亦無法透過選擇題測試。

### 三、本研究無法排除DeepSeek-R1及GPT-4o在訓練過程中已作答過112年、113年律師及司法官考試第一試試題之可能性

由於DeepSeek-R1及GPT-4o兩模型均未完整公開其團隊於開發過程中，或是曾取得並使用哪些資料進行訓練，故筆者無法知悉兩模型之訓練素材為何，而無法完全排除DeepSeek-R1及GPT-4o在訓練過程中已作答過112年、113年律師及司法官考試第一試試題，進而影響其表現之可能性。然而，除非與開發兩模型之團隊合作，否則外部人無法知悉兩模型是否曾有、如何地被以這些考題訓練，故此研究限制適用於所有外部研究團隊。

## 陸、對法律專業的啟示

### 一、大型語言模型對台灣法律的理解與知識儲備進步

依據學者於2023年進行之研究，GPT-4在111年度司法官及律師考試第一試獲得342分，並未通過該年度及格門檻[46]。本研究使用更先進之大型語言模型即GPT-4o與DeepSeek-R1實驗，發現兩者均能通過112年度及113年度司法官及律師考試第一試及格門檻。如此顯示，大型語言模型對台灣法律的理解與知識儲備已有進步，且根據本實驗結果，已經超越通過台灣司法官及律師考試第一試之標準，甚至DeepSeek-R1在回答問題之正確率上已超越約97-99%之人類考生，可謂非常卓越。

同時，不僅針對「台灣」法律，大型語言模型對於美國法律之理解與知識儲備亦有進步，例如依據學者於2023年進行之研究，GPT-3.5無法通過美國律師考試，且表現約相當於人類考生之10%百分位數[47]；但GPT-4已能通過美國律師考試並表現約相當於人類考生之90%百分位數[48]。由此可見，大型語言模型對於各國法律之理解與知識儲備可能均在進步中。

---

註46：Mark Shope, *GPT Performance on the Bar Exam in Taiwan（GPT在台灣專門職業及技術人員高等考試律師考試中的表現）*，SSRN (2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4394826.

註47：Pablo Arredondo, Q&A with Sharon Driscoll & Monica Schreiber, *GPT-4 Passes the Bar Exam: What That Means for Artificial Intelligence Tools in the Legal Profession*, STANFORD LAW SCHOOL BLOG (Apr. 19, 2023), https://law.stanford.edu/2023/04/19/gpt-4-passes-the-bar-exam-what-that-means-for-artificial-intelligence-tools-in-the-legal-industry/ .

註48：Katz et al., *supra* note 16, at 12.

## 二、哪個大型語言模型更適合輔助台灣的法律專業人士？

比較GPT-4o與DeepSeek-R1兩者回答112年、113年司法官及律師考試第一試試題結果，兩者均對於台灣法律展現充足的理解與知識儲備，均通過該年度及格門檻。而就正確率而言，DeepSeek-R1展現更高度之正確率，故若講求正確率，應以DeepSeek-R1作為輔助台灣法律專業人士之工具較佳。DeepSeek-R1表現較佳之原因可能是出於其在中國之開發背景，不僅得以獲得更多中文資料並對中文問題掌握較佳，更可能出於中國與台灣法律背景之相似性[49]。

然而，從回答問題所需之成本而言，GPT-4o在所需時間、所需Token數、所需費用上均顯著低於DeepSeek-R1，故具有更高之可近用性。GPT-4o作答時間約僅為DeepSeek-R1之1%，且DeepSeek-R1回答時間有時候甚至大於司法官及律師考試設定給人類考生之作答時間。此外，本研究中回答相同試卷，GPT-4o所花費總Token數約僅有DeepSeek-R1之18%，且僅需約28%之費用。故若考量使用之成本，GPT-4o更具經濟效益。

## 三、對於司法官及律師考試之啟示

不論對於人類考生或人工智慧而言，時常有論者主張考試之分數高低、通過考試與否，並不是衡量受試者是否有處理法律專業工作之能力、或其能力之高低之指標[50]。此外，選擇題無法衡量處理法律專業人士日常工作內容之能力，且由於考試結果之重點僅在於通過與否，人類受試者普遍並無誘因大幅提升考試分數[51]。甚至以筆者經驗而言，周圍考生時常放棄準備司法官及律師考試第一試之「冷門」或「第二試不會考」的科目（例如國際公法、票據法等），並抱持總分通過第一試門檻即可之心態。由此可見，司法官及律師考試第一試之存在，對於考生能力之衡量效果不一定顯著，且不一定有助於提升考生學習特定法律科目之誘因。

而當大語言模型能在司法官及律師考試第一試獲得及格分數，甚至可能表現優於大部分考生，則司法官及律師考試第一試是否仍有存在的意義？論者或謂選擇題本來就僅是針對法條內容、實務穩定見解等內容設計，無法測試法律專業人士所需要的其他能力，且當律師等法律專業人士需要這些知識時，可以從大型語言模型取得即可，故司法官及律師考試第一試並無存在意義。然而，筆者認為，司法官及律師考試第一試仍有存在意義。第一，是它能測試考生對於「法律是什麼」這一能力之掌握程度，而該能力正是法

註49：例如關於中國民法與台灣民法之相似性，可參：莊錦秀（2020），〈中國大陸民法典對臺灣民法總則編的啟示〉，《財產法暨經濟法》，62期，第77-117頁；姚志明（2017），〈中國大陸民法總則新發展〉，《月旦民商法雜誌》，57期，第47-55頁。

註50：*See e.g.*, Joe Patrice, *New GPT-4 Passes All Sections Of The Uniform Bar Exam. Maybe This Will Finally Kill The Bar Exam*, ABOVE THE LAW (Mar. 14, 2023), https://abovethelaw.com/2023/03/new-gpt-4-passes-all-sections-of-the-uniform-bar-exam-maybe-this-will-finally-kill-the-bar-exam/.

註51：Martínez, *supra* note 21.

律專業人士之基本，甚至反面而言，當大型語言模型均能通過這個考試，則不能通過之人或許並不是那麼適合從事法律專業工作；第二，從務實的考量而言，有鑒於每年報考律師及司法官考試之人數眾多，若全部直接採第二試申論式測驗或第三試口試，則恐怕並無足夠評分者應對，留下第一試選擇題，有助於淘汰許多考生，而有效保留評分者之量能。

### 四、大型語言模型將大幅提升法律專業的可近用性

隨著大型語言模型對於法律之理解與知識儲備增加，將提升人民對於法律專業之可近用性，這至少可分兩點說明：第一，大型語言模型與搜尋引擎不同，採自然對話方式提供回覆，故對於不黯法律、不熟悉法學術語及專有名詞之一般民眾而言，大型語言模型是比搜尋引擎更容易、更簡便獲得答案的方式。第二，大型語言模型的出現亦能縮小民眾與律師間、律師彼此間的知識差距，若律師均能善用大型語言模型或以大型語言模型為基礎之法律科技，大型、費用高之律師事務所展現之優勢將縮小，如此亦能使民眾用更親民之費用獲得足夠品質之法律服務[52]。

### 五、大型語言模型還不能取代法律專業人士

如前所述，本研究僅針對大型語言模型在律師及司法官考試第一試選擇題之表現測試，尚無法評估大型語言模型在案件分析、證據評估、創新性論述等其他法律專業人士需要擁有之能力方面之表現。然而，不論大型語言模型等法律科技如何發展，終究難以完全「取代」法律專業人士，而僅能「輔助」法律專業人士更有效率進行工作，因為終究是人類方受到法律倫理之拘束，並需要依照法律倫理進行決策判斷，完全由人工智慧提供法律服務將有倫理疑慮[53]。

### 柒、總結

本研究探討大型語言模型對於台灣法律的理解與知識儲備，並評估其是否能通過台灣的司法官及律師考試，以分析大型語言模型是否具備取代或輔助法律專業人士的潛力。本研究主要使用兩種大型語言模型進行測試：2024年OpenAI發布的GPT-4o與2025年DeepSeek發布的DeepSeek-R1，並讓其作答112年、113年台灣司法官及律師考試第一試的選擇題。

本研究結果顯示，大型語言模型已具備通過律師及司法官考試第一試的能力，GPT-4o在112年、113年第一試的總分分別為400分、368分，DeepSeek-R1則為460分、460分，皆超過當年律師及司法官考試第一試及格門檻。若以人類考生的排名標準計算，GPT-4o

註52：Arredondo, *supra* note 48.

註53：Lara Kimmel, *ChatGPT Passed the Uniform Bar Examination: Is Artificial Intelligence Smart Enough to be a Lawyer?*, INTL & COMP. L. REV. (2023), ChatGPT Passed the Uniform Bar Examination: Is Artificial Intelligence Smart Enough to be a Lawyer? | International and Comparative Law Review.

約位於約70-80%百分位數，DeepSeek-R1則已超越97-99%的人類考生。比較兩者之表現，DeepSeek-R1正確率較高，但GPT-4o速度較快、且更省成本。DeepSeek-R1在所有科目的表現皆優於GPT-4o，整體正確率高出約10-15%。然而，DeepSeek-R1的作答時間遠超GPT-4o，平均每題需約89秒，而GPT-4o僅需不到1秒。從成本而言，DeepSeek-R1的Token消耗量約為GPT-4o的5.5倍，顯示其在推理過程中可能使用較多內部計算資源。此結果顯示DeepSeek-R1雖能提供較高的正確率，但若考量法律實務應用的即時性與成本，GPT-4o或許更具實用價值。

檢視GPT-4o與DeepSeek-R1的錯題內容，約54-83%的錯誤題目中，GPT-4o與DeepSeek-R1選擇了相同的錯誤答案，顯示兩者可能依賴類似的法律推理模式，而這些錯誤點可能反映大型語言模型對台灣法律某些部分仍存在共通性誤區。兩者在票據法、保險法、強制執行法領域表現較差，但在法學英文、法律倫理、國際公法、憲法等科目表現較佳。此外，大型語言模型的法律表現與題目長度、作答時間關聯性低。

綜上所述，大型語言模型的法律理解與知識儲備正在進步，甚至DeepSeek-R1的成績超越約97-99%人類考生，顯示其對台灣法律的理解能力十分充足。但是雖然大型語言模型能通過律師考試第一試，由於法律工作涉及案件與證據分析、法理思辨、創新性論述等複雜能力，目前大型語言模型仍無法完全取代法律專業人士，應作為法律從業人員的輔助工具，提升法律從業人員的工作效率。如此將可降低法律資訊的取得門檻，使非法律背景的民眾更容易獲得法律建議。此外，大型語言模型可縮小大型律師事務所與小型律師事務所的競爭差距，使法律服務更為普及。隨著科技日新月異，期待大型語言模型的持續進步，能協助台灣民眾獲得更親民、更簡便、更正確的司法環境。

（投稿日期：2025年3月20日）