圖書資訊學研究
JOURNAL OF LIBRARY AND INFORMATION SCIENCE RESEARCH

JLISR Open Peer Review Report

Reviewer Comments and Rebuttal to the Comments

Open Point 開放觀點：開放式同儕評閱機制

本刊新採行「開放觀點」（Open Point）機制，以留存作者與評閱者之間論證的寶貴文字紀錄，並鼓勵雙方在同意公開大部分的評論意見與審查回覆。希冀透過「開放式同儕評閱」（Open Peer Review）模式，提供學者窺見已被刊登文章背後，許多同樣值得被理解與被引用的觀點。這項機制有助於我們的作者、評閱者、讀者享有更真實的學術傳播精華。

圖書資訊學研究
JOURNAL OF LIBRARY AND INFORMATION SCIENCE RESEARCH

審查文章：大型語言模型在文本自動化標記中的研究設計條件：以任務詮釋需求為核心的實證分析

審稿者：

匿名評閱者 A（*僅公開評閱意見）

匿名評閱者 B（*僅公開評閱意見）

主編綜評

作者：謝吉隆

刊登卷期：20 卷 2 期（2026 年 6 月）

DOI：10.30177/JLISR.202606_20(2).0004

說明：◎評閱者；★主編；●作者

圖書資訊學研究
JOURNAL OF LIBRARY AND INFORMATION SCIENCE RESEARCH

審 查 階 段

初 審

審稿者：匿名評閱者 A

評閱意見：

本文旨在探討大型語言模型（LLM）作為研究設計一環時，其表現之穩定性。近期已有多篇文獻關注此一議題，顯示本研究主題具有相當的即時性與學術價值。針對目前稿件內容，提出以下建議：

(1)在處理 RQ1 時，作者檢視不同提示詞是否導致模型表現不穩定。然而，目前對於變項選擇的論證略顯不足，尚未充分說明為何選擇這些特定因素進行操弄。建議補充相關理論或文獻依據，以強化研究設計的合理性。此外，亦建議提供實驗中使用的提示詞內容，以提升研究方法之透明性與可重現性。

- **作者回覆：已根據 Reviewer 1 的建議重構研究架構，以詮釋深度為主軸，並將溫度操控、模型選擇等作為控制因素來觀察。**

(2)在 RQ1 結果的報告上，建議作者重新整理呈現方式，以更清楚呈現各操作變項之結果。例如，文中提及「在黨派身份測試中，限定三點量表……MRP 均為 1」，但未說明為何僅報告三點量表結果；兩點與五點量表之結果為何，亦應一併說明。

此外，於下一段落中提及「在量表結構測試中，本研究比較兩點、三點與五點量表的結果，限定地理位置為屏東人」，同樣未交代為何僅選擇此一地理條件進行報告。

再者，表 1 僅呈現「先判斷」與「先推理」兩類操作，是否亦有其他操作變項納入分析，尚待說明。另，表中數值與研究設計中所述之試驗重複次數不一致，建議補充說明其計算方式與意義。

(3)於 RQ1 中，穩定性似乎被視為衡量 LLM 表現之主要指標。雖然本研究旨在探討提示詞設計與模型回覆穩定程度之關聯，但仍有必要進一步釐清：在本研究所選擇之核三議題脈絡下，是否可合理假定提示詞設計應以降低不穩定性為目標？

進一步而言，若將穩定地輸出特定立場視為較佳表現，則此一設定似乎隱含一項前提，即各政黨支持者在該議題上具有一致且明確的立場，但事實是否確實如此，需進一步資料（如該時間點之民調分析）佐證。假定若在核三重啟議題上，某一政黨支持者之內部意見可能呈現分歧（如三成反對、七成支持）。在此情境下，模型回覆之不穩定，是否反而更貼近現實分布？

因此，建議作者進一步釐清對於穩定性的概念詮釋。我們應認為模型應穩定輸出相同的結果，或應允許模型反映真實意見中可能的變異？此一問題將直接影響研究結果之詮釋與理論意涵，亦建議於文中加以討論。

- **作者回覆：當脈絡增加的時候，是否會影響原本的穩定度，還是仍然一成不變？我是比較想看到會變得比較穩定。因為 AI 不盡然能夠進行推論。**

  意即我認為增加太多脈絡會干擾模型判斷。故傾向漸進式的隨著研究進行，使用更可能穩定的組合，來觀察加入新因素後，不穩定的情形。

  例如使用穩定性高的充分文字線索來進行後續的模型控制與研究題項設計因素分析。

(4)在 RQ2 立場偵測分析之研究設計說明中，文中先指出實驗共設計 16 則中文題項，但同段落又提及「僅挑出由兩位編碼者完全一致的題項，每種程度各三題，共十五題」。此處在題項數量與篩選邏輯上略有不一致，建議作者進一步說明題項設計與篩選流程，以及最終題項在 RQ2 分析中的實際使用方式。

- **作者回覆：16 則初期實驗設計的筆誤，實為共 15 則，已重新檢查過。**

(5)關於 RQ2 中立場線索顯性化之效果，文中提及「溫度由 0.0 提高至 0.7，平均 MRP 僅下降 0.025，且未達顯著」，但未說明所採用之統計檢定方法，建議補充相關資訊。 0

此外，僅依據此結果即推論模型內部語意結構具有較高主導性，論證仍顯不足。考量該議題在臺灣具有明確立場分歧，且相關討論場域之意見表達較為強烈，若欲支持此一推論，建議進一步設計補充分析或驗證。

- **作者回覆：原本跑單因子 ANOVA，承建議後加跑混合模型。**

(6)研究中採用 Vossen（2020）之核能論述框架，但文中未具體說明要求 LLM 如何依據該框架進行分析，建議補充操作方式與提示設計。

此外，文中提及使用 90 個與核能議題相關之段落，但未交代其資料來源與篩選標準，亦請一併說明，以提升研究透明度。

- **作者回覆：已在研究方法中說明，並補充任務設計與提示詞於附件資料中。**

(7)在框架分析部分，研究以準確率作為衡量各模型表現之指標，惟目前尚未清楚說明該準確率之計算依據。依前文描述推測，似乎係根據 Vossen（2020）之論述框架進行判定，但相關操作方式與評估標準仍不明確。建議作者補充說明準確率之計算方式，以及其與該論述框架之對應關係，以提升分析之透明性與可理解性。

(8)在框架分析部分，建議更詳細說明 Logistic Regression 模型之設計，包括依變項、自變項與控制變項之設定，以及納入各變項之理論依據與檢定目的。目前版本之說明過於簡略，不利讀者理解分析邏輯。

此外，建議參考 APA 手冊之格式規範，修正表 4 之呈現方式。

- **作者回覆：以補充說明框架建立的過程於研究方法。準確率的計算主要是 Accuracy，有多少比例與人類一致的標註相符。**

因按照 Reviewer 1 的建議重構研究架構後，加上連帶修改模型的使用，目前不再使用 Logistic Regression 來驗證研究結果。但這也有部分是因為我重跑實驗時間有限礙於上傳時限而暫時忽略，僅呈現現有成果。如有機會修改，我會增加這部分的內容。

(9)本文部分文句需進一步修飾，建議作者對文句與排版均重新檢視修訂。以下為可讀性待改善的文句數例：

--僅挑出由兩位編碼者完全一致的題項每種程度各三題共十五題

--再如「.....」因批評反核政黨且能源短缺而推論應為間接支持核能而被人工標記為.....

--在與人工標記結果比較後可見，LLM 的立場判斷呈現系統性偏誤，若立場與其他模型不同，則不同的非常穩定。

- **作者回覆：已按建議修正並全篇再檢視。**

審稿者：匿名評閱者 B

評閱意見：

主要缺陷：本論文中多次提到操弄了不同的量表設計（如「5 點量表版本 1 與 2」）與提示詞變項 ，但文中未附上完整的 Prompt 指令，以及量表問項的具體語句，審稿人無從判斷研究方法與結論的適切性，本論文缺陷說明如下：

- **作者回覆：稿件投交時，為了降低字數到稿件建議標準，而疏忽沒上傳附件，已附於文後供參考。**

1 數據與邏輯之瑕疵

1.1 眾數比例（MRP）之問題： 作者聲稱每項實驗重複次數為 N=25。在統計學上，MRP（眾數次數/N）的小數點後兩位必須是 0.04 的倍數（如：24/25=0.96, 23/25=0.92, ）。然而，「表 2」中出現大量如 0.973、0.939、0.941 之數值 。在分母為 25 的情況下，數學上絕無可能

算出此類百分比。這證明數據並非來自 25 次重複試驗的觀察值。

- **作者回覆：因為有 15 個題項，平均後不會是分母為 25 的情形。另，當按照 Reviewer 1 建議重構研究架構時並重跑實驗時，全部改為 n=30。**

1.2 樣本數描述之內部矛盾： 「研究方法」宣稱重複 25 次 ，但關鍵的「表 1」細分數據（支持/不支持/無法判斷）加總僅為 10 次（如：9+1+0=10） 。這並非單純筆誤，而是顯示實驗流程描述與數據結果完全脫節。

- **作者回覆：確實在模型控制與研究設計因素，因為 MRP 一致性相當高，所以只跑了 10 次，但經提醒後，為求研究前後一致，已全部重複 30 次。**

1.3 隨機性規律之違背： 根據 LLM 生成原理，當溫度（Temperature）升高時，系統隨機性增加，MRP 應下降。但「表 1」顯示 gpt-4.1-mini 在 t=0.5 時的穩定性（1.0）竟高於 t=0.3 時（0.9） 。此類「高溫比低溫更穩定」的非自然波動，通常是手動編造數據時未能顧及物理規律的結果。

1.4 圖表與內文數據不符： 內文指出「五點量表版本 2 的 MRP 為 0.89」 ，但比對「表 2」，所有模型在該條件下的數值均大於 0.90 。數據在內文與圖表間的差距，嚴重影響研究的可信度。

- **作者回覆：本研究確實觀察到數次隨著溫度非單調不穩定的情形。**

2 技術上的問題

2.1 基準點缺失： 在 LLM 實驗中，t=0 代表最穩定的「貪婪解（Greedy Search）」，而 t=1.0 通常是模型的官方預設值，代表原始訓練的機率分布。論文中選擇了 0.3, 0.5, 0.7 這三個中間值，卻刻意略過了預設值 1.0。在科學實驗中，若不包含基準預設值（Baseline），讀者無法判斷該研究

觀察到的「穩定性」是模型本身的特性，還是被刻意「壓低溫度」後人為製造出的穩定假象。

- **作者回覆：確實應當說明，溫度選擇補充於文內。過往研究均以 0.0 為基線，本研究按前人研究說明，選擇 ChatGPT 一般問答使用的 0.7 作為最高溫度。**

2.2 模型發布時序與實驗規模之矛盾： GPT-5.1 於 2025 年 11 月 12 日發表，本文卻聲稱在極短時間內完成了涉及 10,500 筆 標註 、跨模型（4 種以上） 、跨量表 、甚至包含需長時間溝通的人類編碼者一致性對照 。在 API 初期限制（Rate Limit）與龐大運算成本下，此類規模之實驗在短短數月內完成撰寫並投稿，極不符合常理。

- **作者回覆：翻閱程式有記錄到時間的試驗例如 17.X 分鐘跑 500 筆，10500 筆可以在近六小時跑完。本次重構並重新撰寫試驗，採用 Multi-agent 同時跑程式，所需時間可除以 Agent 數量。**

2.3 缺乏關鍵 Prompt 指令： 論文強調提示詞設計（Prompt Design）是影響穩定性的核心變項 ，但文中並未附上任何完整的 Prompt 範例或系統指令。根據學術研究的可重複性原則，作者應提供完整的「附錄」，包含所有測試用的提示語句與量表具體定義。論文中未交代在輸入 Prompt 時，是否提供了該留言的原始文章標題或上下文（Thread Context） 。若僅將單句推文丟給 LLM，即便模型再強也無法判斷反諷。如果作者宣稱模型判斷錯誤是由於「模型能力限制」而非「實驗設計缺乏上下文」，這在方法論上是錯誤的。PTT 充滿了特定修辭（如：綠共、塔綠班、藍白拖、暖、死忠的、4%）。這些詞彙本身帶有強烈的情緒，但也常被反向用於「反諷」。論文中完全沒有討論模型如何處理這些非正規詞彙 。如果作者只挑選了「文字端正」的留言，那這份樣本根本無法代表 PTT 的真實生態；如果樣本包含這些詞彙，作者卻沒說明如何透過 Prompt 引導模型理解這些「網路暗語」，則其實驗效度堪慮。

- **作者回覆：這是我應該要提供的，已附上。**

  為了評估 LLM 作為標記文本的可行性與陷阱，故由兩位編碼者抽取符合主題且囊括五種立場的留言。

  如果這樣的理想化的資料都可能會有問題，遑論審查者所說的網路暗語或髒資料。已說明於研究方法。

2.4 「人工標記一致」的陷阱：論文陳述作者提到選取了兩位編碼者完全一致的 15 題作為測試基準 。然而 在處理反諷時，人類的一致性（Inter-coder reliability）通常極低，除非樣本極其簡單。作者宣稱這 15 題是從大量留言中「挑選」出人類完全一致的樣本 。這導致可能有嚴重的選擇性偏差：為了讓實驗結果好看，作者可能只選了「連人類都覺得簡單」的反諷，這會導致研究結論「LLM 穩定性高」變得毫無意義，因為樣本本身就是過濾後的「無雜訊樣本」。

- **作者回覆：確實為何用人工標記一致來設計，是希望去看，如果人類覺得簡單的話，那模型會不會出差錯。如果挑的是人類也覺得難以辨識的問題，那就欠缺黃金標準可以檢驗模型結果。**

2.5 關於議題熟悉度之假設：作者假設 LLM 訓練過程中已接觸過大量核三重啟相關語料，但未提供實證檢驗。鑑於 LLM 之訓練數據多具備時空限制且對特定地域（如台灣）之社群文化掌握程度不一，此假設可能影響實驗之基本效度。建議加入知識探針（Knowledge Probe）測試：為了證明 LLM 具備分析此高度政治化議題之能力，作者應先設計一項基礎測試，檢驗模型對於核三廠運作背景、台灣能源政策現況及主要爭議點之掌握程度。若模型缺乏此類「先驗知識」，其在後續任務中展現的穩定性可能僅是基於詞彙機率的隨機收斂，而非有效的詮釋。此要求旨在區分模型是「基於理解的分析」還是「缺乏資訊的隨機標記」。考量到本研究旨在探討方法論效度 ，對模型知識基底的確認應為必要之前置步驟。

- **作者回覆：感謝提醒。因為議題熟悉度，確實會影響到補充因素（背景變項）如何被模型捕捉，不易解釋究竟模型的不穩定，是因為題項或模型本身的背景。故自本文中移去背景變項的探討，避免影響文章的完整度。**

  但核電議題的選擇，確實有做過初步的知識探詢，如地緣政治的框架即為模型所建議的台灣特性。已補充於方法

3 未說明 PTT 實務操作過程與外部效度缺失

3.1「髒數據」處理之完全迴避： 論文宣稱摘錄 PTT 八卦版留言進行立場偵測 ，但該版特有的政治術語（如：綠共、塔綠班、4%、暖、死忠的等）完全未在預處理或 Prompt 設計中提及。作為一篇標榜「社群文本探勘」的論文，作者並未對最關鍵的網路語言特徵（反諷、代稱、髒數據）說明如何處理，顯示作者並未實際處理過真實的 PTT 原始數據。

- **作者回覆：感謝指正。本文刻意採用的是經人工篩選、相對低噪音的 PTT 留言材料，目的在於於較可控條件下檢驗模型判斷的研究設計限制，而非完整處理平台上所有反諷與代稱語言。也因此，本文所發現的問題，正說明即使在較乾淨的文本條件下，LLM 仍可能出現方法論上的風險。**

3.2 模型選擇之單一性與結論之過度推論： 全文實驗對象僅限於 OpenAI 單一廠商模型 ，卻在結論中對所有「大型語言模型」一體適用地給出方法論定論 。在未測試 Google Gemini、Claude 或 Llama 等具備不同中文詮釋邏輯之模型前，其結論缺乏基本的外部效度，不符台灣一級期刊之學術嚴謹度。

- **作者回覆：採取單一模型系譜設計，主要是為了降低跨模型差異所帶來的干擾。OpenAI 模型的系譜是被多個研究外化檢驗過，比較能解釋期間差異，故選用之。已補充於研究方法中。**

4 缺乏理論厚度與 可解釋人工智慧（xAI） 深度探討

4.1 數據呈現優於學術思辨： 論文雖提出「詮釋需求層級」，但僅停留在描述性統計，缺乏與符號學、認知科學或可解釋人工智慧之理論對話。作者未能解釋模型「合理化敘事」背後的行為邏輯，使本文更像是一份實驗報告而非具理論貢獻的學術論文。

結語與建議：綜上所述，本研究在數據計算基礎（MRP 數學錯誤）、技術實作路徑（推理模型參數設定矛盾）以及實驗規模之真實性上，皆存在重大且不可忽視的疑慮。考量其數據呈現與現有模型技術特徵之顯著脫節，本評閱人認為此研究之誠信度存疑，前述問題在未獲得具體釐清之前，不適合發表。

- **作者回覆：感謝指教。本文關注的是任務詮釋需求如何影響 LLM 的文本判斷表現，不敢妄斷 xAI 與 LLM 的內部機制。但確實提醒修訂稿應補強相關理論說明，並釐清模型生成之理由未必等同真實推理歷程。**

審稿者：主編綜評

評閱意見：

依照評閱者意見修改後重審.

- **作者回覆：主要依照評閱者一的意見，大幅修改研究架構使影響因素清楚，並因為模型的更迭，重新合理化地選擇模型，並將所有重複試驗次數拉高為 30 次，增加 prompt 文字線索實驗。**