圖書資訊學研究20：2 (June 2026)：119-154
DOI:10.30177/JLISR.202606_20(2).0004

大型語言模型在文本自動化標註中的研究設計條件：以任務詮釋需求爲核心的實證分析

Examining Research Designs for LLM-Based Automated Text Annotation: An Empirical Analysis Centered on Task Interpretive Demands

謝吉隆*

Ji-Lung Hsieh

國立臺灣大學新聞研究所副教授
Associate Professor
Graduate Institute of Journalism
National Taiwan University

【摘要 Abstract】

本研究探討大型語言模型（large language model, LLM）用於文本標註時，其回應穩定性與方法論效度如何受到模型控制因素、題項設計與任務詮釋需求層級的共同影響。結果顯示，LLM輸出受提示詞局部語詞、模型版本、模型訓練背景、作答順序與任務性質交互影響。在低詮釋需求的立場判斷任務中，特定文字線索可能觸發模型既有立場先驗；不同模型版本與語言版本的生成慣性，其影響亦可能大於溫度設定。題項設計的關鍵不在量表長度，而在是否符合模型能力特性；理由生成也未必提升判斷品質。在中高詮釋需求任務中，高穩定輸出可能只是穩定地重複同一偏誤。整體而言，LLM評估應區分穩定性與效度，並置於任務詮釋需求與研究設計配置中理解。

*通訊作者：謝吉隆 jirlong@gmail.com
投稿日期：2026年1月31日；接受日期：2026年4月21日

This study examines how the response stability and methodological validity of large language models (LLMs) in text annotation are jointly shaped by model-control factors, item-design factors, and levels of interpretive demand. The results show that LLM outputs do not merely reflect model capability; rather, they are jointly influenced by local prompt wording, language, model training background knowledge, response procedure, and task characteristics. In low-interpretation stance-expression tasks, specific lexical cues in prompts can trigger pre-existing stance priors, suggesting that prompt wording should not be treated as neutral background. With regard to model-control factors, generated stance's tendencies associated with model version and prompt language often exert a stronger influence than temperature settings. In terms of item design, the key determinant of stability is not scale length but the compatibility between task demands and model characteristics. Requiring non-reasoning models to reason before answering, or forcing reasoning models into rigid output formats, substantially increases instability. Reason generation also does not necessarily improve judgment quality; instead, it may reduce both stability and validity. Furthermore, in medium- to high-interpretation tasks, stability and validity do not converge: highly stable outputs may simply reflect the stable repetition of the same bias. Overall, this study argues that methodological evaluation of LLM-based text annotation should not rely solely on stability indicators, but should distinguish between stability and validity while interpreting model performance in relation to task-specific interpretive demands and research design conditions.

【關鍵詞 Keywords】

任務詮釋需求；大型語言模型；提示詞設計；回應穩定性；文本標註

Interpretive Demand Levels; Large Language Models; Prompt Design; Response Stability; Text Annotation

壹、研究背景

隨著大型語言模型（large language model, LLM）的發展，其功能已不再局限於生成文本，而逐漸被應用作為社會科學文本分析工具，用於執行各類內容標註任務，如情緒分析（A. Liu & Sun, 2025; Wen, Clough, Paton, & Middleton, 2026）、立場辨識（Abdurahman, Ziabari, Moore, Bartels, & Dehghani, 2025）、法律文件分類（Hila & Hauser, 2025）、文章主題分析（Pham, Hoyle, Sun, Resnik, & Iyyer, 2024）與框架分析（Cardamone et al., 2025; Chew, Bollenbacher, Wenger, Speer, & Kim, 2023）。然而，當LLM被用作分析工具時，方法論上的關鍵問題已不再僅是模型是否具備良好效能，而在於研究者能否清楚評估，在何種條件下，模型判斷可被視為具有分析上的可用性。

既有研究對LLM方法論效度的討論，多半聚焦於準確率、人機一致性、提示詞設計效果，以及不同模型之間的效能比較（Sallam, 2023）。這些研究奠定了重要基礎，但較少進一步區分不同文本任務在詮釋需求上的差異。事實上，部分任務主要依賴明確的字面線索即可完成判斷；相對地，另一些任務則需整合語境，進一步處理隱含立場、因果歸因或敘事角度等較高層次推論。近年研究亦指出，除任務本身的詮釋需求外，模型生成亦可能受到提示詞語言版本、關鍵語彙選擇、題目表述方式，以及是否提供「不確定」或「無法回答」選項等研究題項設計因素的系統性影響（Dunivin, 2025; Shin, 2025）。

此外，隨著LLM架構與訓練資料持續更新，研究者實際面對的LLM並非靜態工具，而是一組隨時間變動的模型版本。因此，模型在特定時間點所呈現的回應，難以視為具有長期穩定性的特徵。若缺乏可重複的穩定性檢驗機制，研究者將難以判斷既有研究所觀察到的不穩定現象，究竟源自任務與提示設計的結構性問題，或僅為特定模型世代的暫時限制；同時，也難以評估後續新模型是否在這些不穩定面向上獲得實質改善。

基於上述背景，本研究主張，若要將LLM視為可信的分析工具，關鍵不應僅止於比較模型是否更為準確。特別是在來源與內容多元、語意複雜的社群文本中，生成表現亦不宜被視為模型本身的固定屬性。因此，在應用 LLM 於文本標註之前，有必要系統性檢視模型生成回應的穩定性，並分析其可能影響因素。在此方法論脈絡下，本研究從模型控制、研究題項設計與任務詮釋需求三個核心面向，系統性檢驗 LLM 回應穩定性的邊界與條件。

近年使用LLM協助質性分析的研究，如Wen等人（2026），除關注模型與人類標註之間的一致性（annotation consistency）外，也開始重視模型在不同執行條件下的變異性（variability）與不穩定性（instability）。本文即在此脈絡下採用「穩定性」（stability）概念，並將其界定為「模型在固定輸入條件下重複執行時，其生成結果是否維持一致的特性」，而非泛指整體工具的可信度。此一界定有助於區別於較廣義的測量可信度（reliability），亦避免與編碼者間一致性（consistency）概念混淆。

貳、文獻探討

一、LLM作為文本自動化標註工具的研究現況與方法論問題

近年LLM已逐漸被應用於文本自動化標註、演繹式編碼與主題分析等任務。相較於傳統需針對特定任務做大量標註與特徵工程的分類模型，LLM能在零樣本或少量樣本的情境下，直接一句自然語言提示來執行情緒分析（A. Liu & Sun, 2025; Wen et al., 2026）、立場判斷（Abdurahman et al., 2025 ）、主題分析與文件分類與框架分析（Cardamone et al., 2025; Hila & Hauser, 2025）等任務，因此被認為具有提升分析規模，並大幅降低編碼人力與時間成本的潛力（Chew et al., 2023; Gilardi, Alizadeh, & Kubli; 2023; Morgan, 2023; Sallam, 2023）。既有研究也顯示，在部分演繹式編碼任務中，LLM 與人類標註結果可達到中度至高度一致，顯示其作為初始編碼者或輔助編碼者的可能性（Chew et al., 2023; Hila & Hauser, 2025; Wen et al., 2026）。

然而，LLM 在文本分析中的方法論效度，不能僅化約為單次準確率或與人類編碼者的一致性。Chew等人指出，LLM 輔助內容分析的關鍵不只是讓模型產生分類結果，而在於研究者能否辨識模型何時理解了編碼任務、何時可能只是近似隨機猜測（Chew et al., 2023）。Wachinger等人也發現，ChatGPT 雖能提出表面上具說服力的主題、引文與理論連結，但其結果仍須經過審慎人工檢視，否則易被其語言流暢性與連貫性掩蓋方法論風險（Wachinger, Bärnighausen, Schäfer, Scott, & McMahon, 2024）。因此，若要將 LLM 視為分析工具，而非僅是生成工具，研究重點便不應只停留在「模型準不準」，而必須進一步探問，模型表現究竟受到哪些研究設計條件所形塑？根據前人研究，可分為如何運用提示詞來設計研究與研究任務本身的特性兩者。具體而言，一方面涉及研究

者如何透過提示詞、模型控制與題項設計形塑模型生成過程，並處理其中的不確定性與推理解釋問題；另一方面則關乎任務本身的詮釋需求差異，亦即不同分析任務在語境整合與推論負荷上的結構性差異。以下文獻探討即依此兩個層面展開。

二、提示詞、題項設計與不確定性管理

（一）提示詞的語言形式影響

既有研究顯示，LLM對提示詞的語言形式高度敏感；提示詞被認為直接參與模型如何理解任務、界定邊界與作出分類判斷的詮釋框架，包含提示詞中的語言用法、指令順序或概念標示上的差異所可能造成的系統性變化（Dunivin, 2025; Ziems et al., 2024）。在Dunivin（2025）針對社會歷史文本的研究中，單純更改文本標註的代碼名稱，即便不改動其定義描述，也能顯著提升模型的分類效能。同時，將關鍵指令前置、改寫為更具字面意義的名稱，也有助於提升模型理解與編碼一致性（Dunivin, 2025）。Ziems等人（2024）的研究也對提示詞進行語義等效的微小變動，證實了這些細微改動會導致顯著的結果變異，從而影響最終的分析穩定性。

再從語言與邏輯結構的角度來看，Hou等人（2024）提出，在設計社群訊息標註的編碼提示時，應考慮減少代名詞（如「它」或「他們」）的使用而改以具體名詞取代，能有效降低模型的理解歧義並提高準確度。Shin（2025）在科學教育對話分析中也觀察到，透過在提示詞中加入明確的排除準則（如「排除缺乏理由的簡短意見」），能使模型與人類研究者之間的一致性大幅提升。甚至，提示詞的語言，如英文相對於韓文，亦可能展現出更高的準確性（Kim et al., 2024）。這些研究共同強調，提示詞工程並非僅是簡單的達意文字輸入，而是一項需要精確措辭與邏輯架構設計的提示詞工程，以確保研究結果的可驗證性與可重製性。這種對語言高度敏感的特性，使提示詞本身成為一個關鍵的分析變項，而不只是單純的技術細節，因而需要謹慎使用檢驗，為運用 LLM 做自動化編碼流程時必須面對的核心挑戰。

（二）題項設計與模型的不確定性管理

除了語彙與結構之外，題項如何安排判斷責任，以及質性編碼中原本即存在的判斷邊界與歧異管理機制之差異，也會影響模型生成。例如，LLM在面對高不確定性、價值判斷模糊或規範風險較高的任務時（如政治立場判斷或框架分析等高詮釋需求應用），往往會出現「無法

回答」、「拒絕判斷」或以高度保留語言回應的現象，以避免產生潛在誤導或爭議（Wachinger et al., 2024）。這類「非回應行為」未必源自單純的技術失誤，而可能與大型語言模型在訓練與部署階段所採取的安全與風險控制設計有關；相關研究指出，模型本身的黑箱特性、訓練資料限制與潛在偏誤亦可能影響其輸出結果（Sallam, 2023）。然而，實際上在傳統質性內容分析與問卷設計中，「無法回應」原本即被視為反映概念邊界模糊或證據不足的重要方法論資訊，研究者通常不追求唯一正確歸類，而是透過編碼者協商、操作性定義修訂或明確揭露歧異，以維持詮釋透明度與信度。而Chew等人（2023）在評估LLM進行演繹式編碼時亦指出，當模型選擇拒答或低確定性回應時，往往集中於代碼邊界重疊、操作型定義不足，或文本本身缺乏明確判斷線索的情境，反映的並非模型「不會做」，而是其對任務可判斷性的內部評估結果。進一步地，Dunivin（2025）將此類回應視為一種「演算法層級的不確定性揭露」，而非錯誤回應。因此，近年已有研究主張，應將「無法回答」納入模型評估的重要指標，而非僅以回應是否一致或準確作為唯一績效標準。此一觀點對於本研究中所關注的輿情論述之框架分析尤為關鍵，因為框架判斷這類高度詮釋需求任務，當邊界模糊時，極有可能促使模型產生非回應行為。

（三）要求推理解釋的生成風險

另一項重要設計條件，是研究者是否要求模型對其分類或判斷提供理由。部分研究發現，要求模型生成理由、逐步說明判斷依據（chain-of-thought prompting），可能提升某些質性編碼任務的表現。例如，Dunivin（2025）的研究即指出，在部分詮釋性質性編碼的任務中，當LLM被要求提出判斷依據時，其與人類黃金標準的對齊程度優於未要求理由時的表現。Chew等人（2023）也指出，要求模型給出編碼決策的理由，有助於研究者檢視模型是否理解編碼任務，並協助辨識何時模型可能只是以近似隨機方式作答。

然而，要求LLM解釋其判斷並不必然等於提升方法論透明度。Wachinger等人（2024）提醒，LLM所生成的理論連結與解釋，往往具有高度表面合理性，但不代表模型真的具備可審計的分析過程；其「理由」可能更接近生成後的語言性合理化，可稱為「自我連貫」或「自圓其說」，實際上並非研究者意義上的可回溯推理。在質性分析與主題分析的應用中，也有研究指出，當模型被要求進行較開放式的主題命名、摘要與解釋時，雖能快速產生看似連貫的分析架構，但也可能傾向於給

出過度寬泛、平滑的主題（De Paoli, 2024; Wen et al., 2026）。因此，要求LLM進行推理解釋在研究設計中具有雙重角色，它既可能協助模型聚焦判斷依據，也可能進一步介入生成流程，放大模型自我連貫與事後合理化的傾向。

三、任務詮釋需求的層級差異：從直接分類到高詮釋性判斷

若將前述提示詞設計、不確定性管理與推理解釋要求放在更高層次來看，這些因素之所以重要，乃是因為不同文本任務本身具有不同程度的「詮釋需求」。部分任務較接近明確分類或字面辨識，主要依賴文本表面線索；另一些任務則涉及語境整合、隱含立場辨識、角色推論、因果歸因與敘事框架判斷，在編碼過程具有更高的詮釋負荷。Hou等人（2024）曾將演繹式編碼任務區分為「語境獨立」與「語境依賴」兩類，並指出前者通常較容易達到較高的一致性，而後者則更仰賴對上下文的理解與推理整合能力。

舉例來說，相較於情緒分類或簡單類別判斷等低詮釋需求任務，模型在需要判斷他人立場、意圖或隱含意義的高詮釋需求情境中，回應的不確定性也比較高（Chew et al., 2023; Wen et al., 2026）。要求模型進行「直接表態」或「判斷他人立場」並非等質任務，前者更接近角色模擬，模型僅需生成符合提示期待的立場；後者則涉及對文本線索的整合與詮釋，具有更高的語用不確定性（Wachinger et al., 2024）。

又如，在社群文本探勘任務中，主題分析或框架分析常對應兩種不同的方法論取向，前者多採取由下而上的歸納式邏輯，強調讓主題自資料中浮現；後者則多為由上而下的演繹式分析，研究者事先界定分析框架，並據此判斷文本如何建構因果關係、責任歸屬與敘事角度。既有研究顯示，演繹式分析方法因有事先的清楚定義並以編碼本引導，LLM的回應較為穩定且可控（Tai et al., 2024; Xiao, Yuan, Liao, Abdelghani, & Oudeyer, 2023）。相對而言，歸納式編碼要求模型在缺乏既定類別的情況下生成與整合主題，其表現更高度依賴研究者在分析過程中的詮釋與判斷（De Paoli, 2024; Gao et al., 2024），需透過反覆檢視文本、整合語意線索來回應研究問題，完成任務的詮釋需求更高（Wen et al., 2026）。因此，標註任務的難度與模型表現會隨任務詮釋需求層級而改變，而前節研究設計因素的效果，也可能在不同任務類型中呈現不對稱影響，說明LLM的表現並非固定屬性，而是由任務詮釋需求與研究設計條件共同形塑。

四、研究缺口：從模型效能評估轉向方法論條件檢驗

綜合上述文獻可見，LLM 在文本自動化標註中的應用，已從單純測試模型是否能做出分類，逐漸轉向討論提示詞設計、理由生成、上下文依賴與任務差異等方法論問題。然而，既有研究仍有兩項不足。第一，多數研究雖指出提示詞、語彙框架、不確定性管理與推理解釋要求會影響模型生成，但這些因素往往分散於不同研究脈絡中，較少被整合到同一套研究設計中進行系統比較。第二，現有研究雖已觸及不同任務的語境依賴與詮釋需求差異，但較少進一步檢驗，相同的設計因素是否會在不同詮釋需求層級的任務中產生不同效果，亦即這些方法論條件是否具有任務依賴性。

基於此，本文的研究焦點在檢驗研究設計條件與任務詮釋需求如何共同形塑模型表現。具體來說，本文關注的不是LLM在單一任務中是否「足夠準確」，而是在不同任務詮釋需求下，企圖觀察題項語言、語彙框架、量表設計與推理解釋要求等因素，如何影響模型生成的穩定性與判斷表現。在此脈絡下，本文進一步提出以下研究問題，以系統性分析LLM作為文本分析工具的邊界與條件。

（一）RQ1（模型控制與研究題項設計對回應穩定性的影響）

既有研究指出，LLM的生成不僅受到模型本身條件影響，也會隨提示結構、題項形式與回應程序的差異而產生變化。然而，這些因素何者對回應穩定性影響較大，仍缺乏系統性的比較。據此，本研究整理前人研究將可能影響因素系統性地區分為：1. 模型控制因素（包含模型版本、控制隨機性的溫度與所用語言）；2. 研究題項設計因素（包含題項選項設計、是否要求推理、推理與作答的先後順序、文字表述方式，以及提示詞中的文字線索）是否會影響LLM的回應穩定性？在這些因素之中，何者對回應穩定性的影響較為顯著？

（二）RQ2（任務詮釋需求層級與回應穩定性的關係）

當LLM所面對的任務由立場表述、立場判斷進一步提升至框架分析時，模型所需處理的不只是分類本身，而是不同程度的語用理解、語境整合與概念判準套用。相關研究雖已指出不同分析任務之間可能存在難度差異，但尚少從「詮釋需求」的角度，系統性檢驗其與回應穩定性之間的關係。低詮釋需求任務主要依賴文本表面線索與較明確的態度表達；高詮釋需求任務則涉及對隱含立場、語境脈絡或敘事框架的判讀。準此，本研究欲探問在不同模型控制與研究題項設計條件下，相較於低

詮釋需求的立場表述任務，中度詮釋需求的立場判斷與高度詮釋需求的框架分析，是否更容易出現不穩定或結構性收斂的回應模式？

參、研究方法

本研究旨在系統性檢驗LLM在文本自動化標註任務中的回應穩定性，並釐清其變化是來自模型控制、研究題項設計，或是來自任務本身的詮釋需求差異。為此，本文以「核三廠是否應重啟」作為共同議題場域，在相同公共政策爭議脈絡下，建構不同詮釋需求層級的文本標註任務，並系統性操弄可能影響模型生成的各項條件。此一設計對應前述兩個研究問題：一、檢驗模型控制因素與研究題項設計因素是否影響LLM的回應穩定性，以及何者影響較為顯著；二、檢視在不同控制因素與研究題項設計因素條件變動下，不同詮釋需求層級的任務是否呈現系統性的穩定性差異。

一、研究材料與不同詮釋需求任務的設計

本文以「核三廠是否應重啟」作為測試議題。此議題兼具明確的政策爭議性、社會討論度與多重論述框架，適合用以比較不同文本標註任務所需的詮釋負荷。相較於低爭議或單一意義明確的議題，核電重啟議題在公共討論中常同時涉及能源供應、空氣污染、核安風險、地緣政治、政黨競爭與世代正義等不同框架，故能提供從表層態度表達到較高層次論述判讀的分析情境。此外，該議題在新聞與公共輿情論述中資料充足，可合理假設LLM在訓練過程中已接觸過大量相關語料以回應本研究中的提示詞，且在後續立場與框架的標註任務結果，和兩位人類編碼者的結果對照下，可達到近80%的準確率。

在此議題下，本文建構三種文本標註任務。第一類為立場表述任務，要求模型直接表達其對「核三廠是否應重啟」的支持、反對或中立立場。此類任務主要要求模型針對一明確政策問題做出自身立場回應，賴以判斷的依據多為題目本身所提供的表面語意與明確選項結構。第二類為立場判斷任務，要求模型根據既有文本內容判斷文本作者對重啟核三的立場。此類任務要求模型不僅需處理顯性態度字詞，亦可能需結合反諷、批評對象、語境暗示與價值指涉進行推論。第三類為框架分析任務，要求模型辨識文本主要採用何種論述框架來理解或表述核三議題。此類任務要求模型將文本歸入較抽象但已經定義好的論述框架類別，需進一步整合文本中的敘事焦點、問題定義、歸因方式與價值訴求。三種任務在表面形式上皆屬分類或標註工作，但在實際處理過程中，分別對

應不同程度的語用理解、語境整合與概念判準套用，因此可視為詮釋需求不同的任務層級。本文即以此操作化方式，檢驗詮釋需求層級是否與回應穩定性呈現系統性關聯。

二、研究變項的操作化

（一）回應穩定性的操作化

本文主要的依變項為回應穩定性與回應準確率。回應穩定性指在相同輸入條件下，模型多次執行後是否能產生一致或高度相似的生成結果。所用指標為「眾數比例」（modal response proportion, MRP），即最常出現的答案占所有回應的比例。模擬問答的程序完全以程式自動化，所有試驗均重複 30 次以計算MRP。若模型在重複執行相同任務與相同提示條件時，多次給出相同分類結果，則可視為穩定性較高；反之則視為穩定性較低，並透過條件的比較探究穩定度較低的原因。回應準確率則是計算 LLM 所生成的選擇，相符於人類標註一致的選擇的比例。

本文將回應穩定性理解為一種方法論上的可重現性指標，而非正確性本身。換言之，穩定並不必然等於正確，不穩定也不必然等於錯誤；但若模型在相同條件下無法維持基本一致，則其作為文本自動化標註工具的研究可用性將受到限制。這樣的區分與前人探究 LLM 自動化文本標註的研究觀點一致，即研究者必須同時關注模型生成的可靠性與其是否真正對應研究者欲操作的構念，而不能僅以單次結果作為判斷基礎（Chew et al., 2023; Wachinger et al., 2024）。

（二）模型控制與研究題項設計因素的操作化

在以詮釋需求低的立場表述任務來進行模型控制與研究題項設計因素的分析之前，將先進行立場表述提示詞文字線索的檢測分析，以瞭解提示詞中的概念要素會如何影響生成穩定度。實際上是將提示詞「如果你是屏東縣民，退役的核三廠就在屏東縣，你會支持重啟核三嗎？」拆解為「身分」、「退役狀態」與「區域」三個事實，並以2^3 完全因子設計來分析這些文字線索的影響。

在模型控制因素方面，本文納入模型版本、溫度設定與提示語言三種因素。模型版本用以比較不同LLM在相同任務下是否呈現不同的穩定性模式；溫度設定則用以控制生成過程中的隨機性，本文設定不同的溫度取值為0.0、0.3、0.5與0.7，前人研究為了控制模型的決定性與可重製性最常用的0.0為基線（A. Liu & Sun, 2025），並以ChatGPT一般對話的介面預設溫度0.7為中高水準，以檢驗當生成自由度增加時，模型生成是

否更容易產生波動而不穩定。提示語言則比較中、英文提示詞在相同任務下的差異。如此設計可對應既有研究中對模型能力差異、提示詞語言效果與生成參數設定的討論（Hou et al., 2024; Wen et al., 2026）。

在研究題項設計因素方面，本文將題項選項設計、要求推理與作答順序，和文字表述語彙的差異納入觀察。選項設計比較了Likert二點、三點與五點量表，同時比較是否提供「無法回答」或「無明確立場」選項，以檢驗分類邊界的設定是否改變模型回應的集中程度或分布型態。要求推理與作答順序的操作則用以比較模型在1. 僅生成答案、2. 先推理後作答、3. 先作答後推理等條件下，是否呈現穩定性差異。文字表述語彙因素的設計則指在維持題意近似的前提下，調整措辭、回應格式與關鍵詞，如「支持／贊成」、「重啟／恢復運轉」、使用數字作為序號代替文字作答，來觀察模型對表述差異的敏感程度。

（三）立場判斷

同樣以「重啟核三」為情境，本文設計LLM文字生成任務以判斷留言中對重啟核三的立場。為降低留言中反諷、語意模糊、離題或立場不明等情形對研究的干擾，本研究擬依立場強度從「明顯不支持」到「明顯支持」抽出五個層級的代表性留言。具體而言，兩位同時從事核電公眾討論研究的研究者（包含作者）對隨機抽取的留言進行編碼，僅保留兩位編碼者具共識的留言，直到五個立場層級各取得三則留言為止，最終樣本共15則。所有留言皆取自臺灣PTT Gossiping討論板於2025年間與核三是否應重啟發電相關之留言（如附錄一）。延續立場表述任務對模型控制與量表設計因素的影響觀察，在立場判斷任務中亦檢驗模型控制因素（模型與溫度）與題項設計的差異。針對15個題項均重複詢問30次以求依變項MRP之平均值來觀察差異。

（四）框架分析設計

本實驗延用Xiao等人（2023）的前人研究結果採用演繹式的框架分析任務，先定義好框架（含框架名稱與描述）。理由在於，若允許模型自行生成框架類型，將使分析結果同時混合「框架建構」與「框架判斷」兩種不同層次的認知活動，不利於釐清模型在標註任務中的不穩定究竟源自提示詞設計、推理解釋需求，或是分類邊界本身的模糊性。

為確保研究焦點集中於LLM的標註表現，而非框架體系的理論爭議，本研究直接參考 Vossen（2020）對荷蘭平面媒體中核能論述所設計的框架，共包含失控科技、進步、永續、生態現代主義（ecomodernism）、權衡取捨、成本效益、公共問責、社會正義與能源獨立等九個框架。各

框架之操作性定義、核心問題設定、因果歸因方式與價值判斷邏輯，皆依據Vossen（2020）之原始描述加以整理如附錄二，並要求模型僅根據此框架進行判斷。呼應其的做法，本研究要求模型針對單一輿情段落判斷「最主要」的論述框架，而非同時指派多重框架，以確保不同文本之間的比較基準一致。透過此種演繹式框架標註設計，本研究得以免於框架驗證的前提下，進一步檢驗LLM在高詮釋需求任務中的標註穩定性與方法論風險。

三、研究執行

所有實驗皆透過Python 撰寫，自動化呼叫OpenAI API執行實驗。實驗使用Pydantic所定義的結構化回應模型來限制LLM的生成輸出格式，使模型只能在預設的選項集合中作答，以避免自由回答產生額外的解析問題。不同題目條件對應不同的回應結構，例如二點量表、三點量表與五點量表，以及是否包含理由欄位，皆以型別約束方式實作，以保證所有回應在收集階段即為有效資料。立場表態、立場判斷與框架分析任務的提示詞如附錄三。

為了降低跨模型差異所帶來的干擾，並利於與其他研究對話，主要實驗模型選用最多研究使用的OpenAI模型，並以gpt-4.1-mini作為基線模型，其原因在於該模型支援溫度參數調整，且計算成本相對可控，適合一般研究進行大量重複查詢。另外選用OpenAI當下最先進的gpt-5.4-mini與gpt-5.4來進行模型比較。選用這兩個模型是因為其模型訓練資料截斷時間為2025年8月31日，包含臺灣2025年的重啟核電公投。除此之外，gpt-5.4為具推理能力的模型，gpt-5.4-mini作為gpt-5.4的速效模型，則可與gpt-4.1-mini做溫度等模型參數調控的比較。模型top-p固定為1.0，不限制最大token數。為避免API速率限制與暫時性錯誤對結果造成影響，每一次呼叫之間皆加入0.5秒延遲，並設定自動重試機制。

肆、研究結果

一、提示詞線索效應

首先進行立場表述的提示詞線索檢驗，以瞭解提示詞中的概念要素會如何影響生成穩定度。2^3 完全因子設計的結果顯示，模型對「重啟核三」的立場不一致，且明顯受模型影響。於無任何語言線索的基線條件下，gpt-4.1-mini與gpt-5.4-mini的主要反應皆為「不支持」（MRP均為0.900）；

相對地，gpt-5.4則穩定呈現「支持」（MRP = 1.000），顯示不同模型之間存在方向相反的先驗立場，並因模型推論能力而有差異。進一步比較三個語言線索的影響後可發現，身分、核電場退役與所在地等三個線索的任何一個都會將兩個原本傾向「不支持」的模型之MRP由0.900推升至1.000。在原本即穩定「支持」的gpt-5.4中，加入線索後仍未改變其立場方向。換言之，這些線索的作用不會翻轉模型立場，而是強化既有模型先驗生成的結果。此外，三個線索的邊際效應幾乎等效，在gpt-4.1-mini與gpt-5.4-mini中皆呈現ΔMRP = +0.025，且未觀察到明顯交互作用，表示三個語詞線索在本研究條件下屬於可互換的同強度觸發線索。

溫度檢驗則進一步顯示，生成隨機性只會影響無線索題項的穩定度，卻無法稀釋關鍵字效應。以gpt-4.1-mini為例，當溫度由0.0提高至0.7時，作為基線的無線索題項的MRP由0.900下降至0.667，表示在沒有語詞引導時，模型回應確實會因隨機性而鬆動；然而，只要題項中包含任一語言線索，多數試驗仍維持在MRP = 1.000。以gpt-4.1-mini為例，若不包含語言線索，從0.0 ~ 0.7四種溫度的MRP分別為0.900、0.900、0.767、0.667；但若包含任一語言線索（共七種）與四種溫度共28組測試，僅有一組之MRP為0.933（溫度：0.7），其餘MRP均為1.000。這個結果說明，任務中的局部語詞本身就會系統性塑造回答方向的處理因素，未來在設計實驗設計前，恐需建立「關鍵字效應表」，才能避免將語詞觸發誤判為模型對公共議題的真實判斷。

二、模型控制因素的影響

接下來觀察溫度與提示詞語言兩個模型控制因素的影響。OpenAI的溫度控制僅限制於非推理模型，因而比較兩個gpt-4.1-mini與gpt-5.4-mini兩個模型在有完整文字線索下的差異。結果顯示，gpt-4.1-mini的不同溫度與語言組合（共八種）均不影響其生成結果（MRP = 1.000, 不支持）。gpt-5.4-mini則呈現非單調的語言差異，英文提示詞在溫度為0.3、0.5、0.7時，MRP為0.933、0.867、0.967。中文提示詞則不受溫度影響（MRP = 1.000）。

推理模型gpt-5.4的立場方向完全受語言決定。gpt-5.4在提示詞線索中的中文提示詞全為「支持」（MRP = 1.000），英文提示詞則完全相反為「不支持」（MRP = 1.000）。如果延伸再測試文字線索對英文提示詞生成結果的影響，會發現只要有文字線索，無論多寡，生成結果都是「不支持」（所有文字線索的平均MRP = 0.975）。同家族的gpt-5.4-mini則無此立場方向差異。

gpt-5.4翻轉生成結果的可能原因是訓練語料差異。由於該模型收錄資料涵蓋公投前的完整時期，有可能中文提示詞觸發gpt-5.4的中文訓練語料叢集（可能偏向2025年臺灣核能延役公投雖未過門檻但贊成較多的輿情影響），英文提示詞則觸發英文語料的反核／環境論述叢集。據此，可推測中、英提示詞結果可能受訓練語料而反向，而不是簡單的翻譯誤差。

三、研究題項設計因素影響

從提示詞線索探索的結果得知，線索缺乏時會造成模型無從判斷而不穩定；再從模型控制因素的探索中得知，英文提示詞可能會受模型語料本身或溫度影響而呈現非單調的反應，故將使用線索充足且穩定的中文提示詞做研究題項設計的影響觀察。

接續分析研究題項設計對穩定度的影響，以包含量表結構（兩點／三點／五點量表）、語彙表達（使用「支持／不支持」、「贊成／不贊成」、「重啟／不重啟」、「數字代號」等不同方式，以及推理要求及其形式（另要求LLM「先說理後判斷」、「先判斷後說理」與「兩階段判斷」）是否影響LLM的回答穩定性與極端選項使用率？一共組成九組操作，搭配九組不同的前節模型控制因素，共81種研究題項設計與模型控制的搭配如表1。

表1

研究題項設計與模型控制因素

研究題項設計九組	模型控制因素九組
三點, 支持／不支持, 不要求推理（比較基準） 兩點, 支持／不支持, 不要求推理 五點, 支持／不支持, 不要求推理 三點, 贊成／不贊成, 不要求推理 三點, 重啟／不重啟, 不要求推理 三點, 1. 支持／0. 不支持, 不要求推理 三點, 支持／不支持, 先選擇後理由 三點, 支持／不支持, 先理由後選擇 三點, 支持／不支持, 先理由後選擇（兩階段）	gpt-4.1-mini 四組 (temp = 0.0, 0.3, 0.5, 0.7) gpt-5.4-mini 四組 (temp = 0.0, 0.3, 0.5, 0.7) gpt-5.4 (推理模型)

在81種探索組合中，有70種組合呈現MRP = 1.000的完全穩定結果。不穩定的組合則主要受不同模型的推理能力（模型控制因素）和推理要求（研究題項設計）的搭配條件影響。以gpt-5.4而言，多數題項條件下

皆呈現穩定的「支持」立場（MRP = 1.000），但仍有四種組合出現不穩定現象。例如，在兩點量表中，模型雖整體仍傾向支持，但MRP降為0.900，顯示缺少「無法回應」的二元選項的作答形式可能弱化了 gpt-5.4 原本的穩定性；當改以數字代號作答時，模型立場更進一步翻轉回「不支持」（MRP = 0.767）。若要求模型先選擇再回答，其結果仍維持以支持為主（MRP = 0.967）；但若改為二階段作答，即先推理再選擇，則結果翻轉為不支持（MRP = 0.933）。此結果顯示，對gpt-5.4而言，作答形式與推理程序皆可能影響其最終立場表現。

相較之下，gpt-4.1-mini模型本身為一非推理模型，當要求其進行推理時，採取先選擇後推理的條件下，仍可維持完全穩定（MRP = 1.000）；然而，一旦要求模型先推理後選擇，不論是一次性結構化輸出，或是分成兩階段作答（先說明看法，再標註選項），其穩定性均明顯下降。在這些條件下，MRP並隨溫度上升呈現非單調下降，共有七種組合落入此類，平均MRP為0.833，顯示 gpt-4.1-mini 在被賦予推理任務時較易產生不穩定結果。另一方面，gpt-5.4-mini在所有研究題項設計探索條件下皆一致回覆「不支持」，且所有組合均為MRP = 1.000，為三個模型中表現最穩定者。

綜合以上對研究變項設計的探究發現如下：（一）研究題項中的推理要求是否與模型本身的推理特性相匹配，可能是影響回應穩定性的關鍵因素；溫度的作用則較可能是放大這種不匹配，而非獨立造成不穩定的主要來源。（二）在提示詞線索明確的前提下，研究題項設計因素的影響比預期小，除了中性選項可能受明確的提示詞線索影響外，量表與語彙均不影響模型回答的穩定性。（三）另發現，無論是哪種模型、無論是選擇支持或不支持，在此立場表述任務中幾乎不會選擇極端值，呈現中間傾向偏誤。在不同溫度、不同模型的五點量表之270次回答中，僅有一次（gpt-4.1-mini, temp = 0.7）選擇了「非常不支持」。

四、立場判斷任務中的穩定性與判斷偏誤

（一）立場判斷任務的穩定性

RQ2 欲探問在不同模型控制與研究題項設計條件下，不同詮釋需求層級的任務是否會呈現穩定性差異。在觀察立場表述任務後，接續觀察中等詮釋需求的立場判斷任務，LLM被要求判斷15則人類標註者標註過的留言是否支持「重啟核三」，每組條件各跑30次以觀測MRP值。

首先觀察模型控制因素的描述性差異。gpt-5.4、gpt-5.4-mini、

gpt-4.1-mini三個模型在立場判斷任務中展現高穩定性，MRP分別等於0.990、0.968、0.930。隨著模型溫度上升，模型gpt-5.4-mini單調下降約0.02 ~ 0.05，gpt-4.1-mini單調下降約0.01 ~ 0.05。在這類較高詮釋需求的任務中，雖然仍舊保持高穩定度的表現，但模型相較於前節立場表述任務略不穩定，且會隨著溫度上升而略為下降。

再觀察研究題項設計對立場判斷任務穩定性的影響，可發現題項語彙表述比量表長度更能解釋回應穩定性的差異。整體而言，第二種文字敘述（明顯支持 ~ 明顯反對）的平均MRP為0.940，低於基準文字敘述（支持／不支持）的0.968，差距為0.028（paired $t(269) = 3.84$, $p < .001$; Wilcoxon $p < .001$）。相較之下，三點量表與五點量表的平均MRP分別為0.961與0.947，差距為0.014，未達顯著（$t(269) = 1.53$, $p = .127$）。Two-way ANOVA 進一步顯示，語彙表述具有顯著主效應（$F = 8.64$, $p = .003$），而量表長度主效應不顯著（$p = .142$），兩者交互作用則為邊緣顯著（$p = .099$）。此一差異主要集中於屬於非推理的前代模型gpt-4.1-mini，尤其在五點量表結合第二種語彙敘述下，其MRP降至0.866。具推理能力的前沿模型gpt-5.4則在各條件間維持高度穩定。

為排除留言內容差異對MRP的干擾，本研究進一步建立以留言內容為隨機截距的混合效果模型。結果顯示，留言內容的群聚效應為組內相關係數（intraclass correlation coefficient, ICC）= .101，支持將留言內容隨機效果納入模型。在固定效果方面，五點量表版本二相較其他三種量表條件呈現顯著較低的MRP（$\beta = -.038$, $SE = .010$, $z = -3.72$, $p < .001$）。另一方面，溫度的負向效果亦達統計顯著（$\beta = -.055$, $SE = .016$, $z = -3.36$, $p = .001$），但在控制溫度的觀察範圍內，所對應的預測差距僅約3.8個百分點，顯示其影響幅度有限，並非造成MRP變異的主要來源。整體而言，在納入留言層級變異後，量表與溫度對穩定性的影響雖達統計顯著，但效果幅度有限，進一步支持前述發現，即相較於語彙表述，這類模型控制與量表條件對MRP的解釋力較為次要。

在本節立場判斷任務中，模型生成結果並未迴避五點量表兩端的極端選項，有38.11%使用極端選項。此結果與前節立場表述的任務僅0.37%的題項選擇極端選項不同。相較於在前節立場表述任務中所觀察到的中間傾向偏誤（不選擇極端選項），較可能來自模型在立場自我表述情境下對極端立場的抑制，而非模型在立場判讀任務中的一般傾向。立場判斷的五點量表分布可見圖1，模型選擇不支持的極端選項比例，比選擇支持的極端選項比例高。除此之外，不管任何一種模型，均有一定比例以上選擇無明確立場的選項。

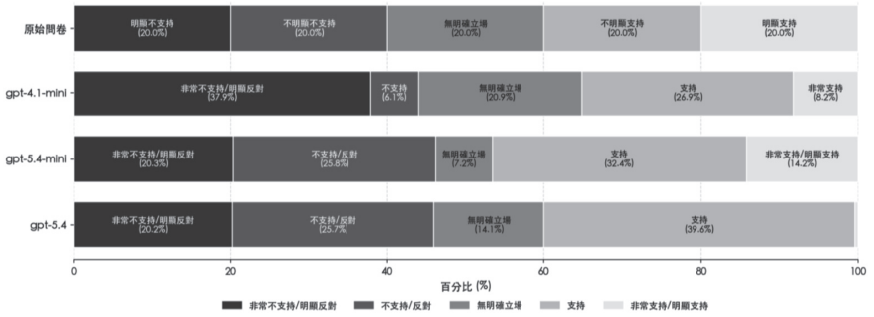

圖1　不同模型在五點量表上的立場分布比較

（二）立場分析的效度比較

接續穩定性分析之後，本研究進一步檢驗模型在立場判斷任務中的效度表現。為降低五點量表極端選項對效度比較的影響，先將五點量表結果對應並合併為三分類，亦即支持、無明確立場與不支持，再以人工預期標籤作為比較基準。結果顯示，三個模型的整體正確率存在顯著差異（$\chi^2(2) = 525.75$, $p < 10^{-114}$）；其中，gpt-5.4-mini的正確率最高（82.1%），其次為gpt-5.4（71.8%）與gpt-4.1-mini（65.3%）。gpt-5.4雖於前述穩定性分析中表現出最高的一致性（平均MRP約為0.990），但其正確率仍明顯低於 gpt-5.4-mini，顯示回應穩定性與分類正確性並不必然一致。

溫度並未顯著影響立場判斷的正確率。各溫度條件下的正確率介於72.3% ~ 74.7%之間，變動幅度有限；以Spearman等級相關檢定溫度與正確率之單調關係，在本研究設定的溫度範圍內，沒有證據支持溫度與效度之間存在系統性關聯（$r_s = 0.000$, $p = 1.000$）。量表設計對效度表現具有顯著影響。四種量表條件下的正確率依序為L5v1（Likert五點量表、語彙版本1）= 77.0%、L3v2 = 74.8%、L5v2 = 73.9%、L3v1 = 68.4%，其差異達統計顯著（$\chi^2(3) = 83.21$, $p < 10^{-17}$）。顯示題項選項的語彙表述與量表長度皆可能影響模型的分類正確率。

立場判斷與說理的先後順序會影響模型的穩定性與效度，且三個模型皆呈現相同方向的結果：「立場先」的表現最佳，「理由先」與「兩階段」則相對較差。在穩定性方面，gpt-4.1-mini的「立場先」高於「理由先」（MRP = 0.9822 vs. 0.9000, Wilcoxon $p = .0418$）；gpt-5.4與gpt-5.4-mini亦呈現相同模式，且三種模式的整體差異均達顯著（gpt-5.4: 0.9978, 0.9600, 0.9156, Friedman $\chi^2 = 11.10$, $p = .0039$; gpt-5.4-mini:

1.0000, 0.8333, 0.8022, $\chi^2 = 14.33$, $p < .001$）。在準確率方面，此一模式同樣成立。gpt-4.1-mini 的「立場先」高於「理由先」（67.33% vs. 57.33%, McNemar $\chi^2 = 43.02$, $p < .001$）；gpt-5.4的三種模式依序為73.56%、66.67%與71.11%，其中「立場先」僅顯著高於「理由先」；gpt-5.4-mini則分別為86.67%、76.89%與67.78%，且三模式差異最為明顯（Cochran's $Q = 98.56$, $p < .001$）。整體而言，先要求模型表明立場、再補充理由，較能維持較高的回應穩定性與判斷準確率。

再按立場類別檢視精確率（Precision）、召回（Recall）與F1 score，模型在不同立場類別上的分類效度存在明顯差異。對「支持」類別而言，Precision = 0.814、Recall = 0.825、F1 = 0.819；對「不支持」類別而言，Precision = 0.773、Recall = 0.842、F1 = 0.806；然而，「無明確立場」類別的對應指標僅為 Precision = 0.430、Recall = 0.342、F1 = 0.381。整體而言，模型對兩端立場類別的辨識表現明顯優於中立類別，顯示效度不足主要集中在「無明確立場」的判定上。

若進一步觀察判斷錯誤的留言題項，效度最低者集中於需要語用推論、反諷辨識或政治脈絡推斷的句子。例如「不在意電費了，別再用核能撕裂臺灣了好嗎」在人工標註中屬於無明確立場，但1,080筆預測中100%被判為反對側，代表模型傾向將「別再用核能撕裂臺灣」直接解讀為反核表態，而未能辨識其實際上較接近「停止爭論」的語用功能。另一例是「自稱最愛臺灣的政黨卻為臺灣帶來戰爭時能源短缺的嚴重風險」，人工編碼將其歸為不明顯支持，但模型表現出明顯分裂：gpt-5.4為100%判為支持，gpt-5.4-mini為94.2%，而gpt-4.1-mini則有66.9%判為反對。整體而言，模型較容易誤判的文本特徵，主要包括中立但帶有負向詞面、反諷語氣、政黨攻擊導向，以及需經由能源風險或政治批評反推出立場的句型。這也說明，在立場標註任務中，僅有高穩定性仍不足以保證高效度，特別是面對隱晦、間接或多層語用結構的文本時，人工校準與複核仍不可或缺。

五、框架判斷的穩定性與詮釋

接續分析高詮釋需求的框架判斷之穩定性及其影響因素，特別關注人類標註一致性是否會對模型穩定性產生影響。結果顯示，即使在人類標註者不一致的題項中，模型的回應並未出現較高的不穩定性。本研究先讓兩位人類編碼者對90個與核能議題相關的段落進行編碼，兩人初步達成共識的段落有59個（65.6%）、不一致的段落有31個（34.4%）。經過LLM標註後，按人類編碼者一致與不一致的編碼結果累計其MRP觀察

穩定性結果如表2。可觀察到不同模型的穩定度雖有差異，均未達統計顯著水準。整體而言，效果量分析顯示人類標註是否一致對模型穩定性的影響屬可忽略至小效果（Cohen's *d*介於0.16 ~ 0.40之間），說明人類編碼分歧並未系統性轉化為模型回應的不穩定來源。

表2

不同模型在框架偵測任務的穩定性

Model	MRP (Agree)	MRP (Disagree)	Δ	*p*-value	Cohen's *d*
gpt-4.1-mini	0.9960	0.9785	-1.76%	.179	+0.402
gpt-5.4	0.9751	0.9871	+1.23%	.370	-0.164
gpt-5.4-mini	0.9695	0.9839	+1.48%	.284	-0.189

註：MRP：眾數比例（modal response proportion）。

進一步在控制人類標註一致性的條件下，以兩位人工編碼者標註一致的59段文本作為效度比較基準，檢視三個模型在高詮釋需求的框架分析任務中的分類表現。結果顯示，三者的準確率由高至低依序為gpt-5.4（80.73%）、gpt-5.4-mini（78.36%）與gpt-4.1-mini（77.80%）。由於三個模型皆針對同一批文本進行標註，因此採用適用於配對資料的Cochran's *Q*檢定比較其整體差異。結果顯示，三個模型之間的整體差異達統計顯著（$Q(2) = 9.89$, $p = .007$）。進一步以 McNemar 檢定進行兩兩比較後發現，gpt-5.4 的準確率顯著高於gpt-4.1-mini（$p = .002$）與gpt-5.4-mini（$p = .028$），而gpt-4.1-mini與gpt-5.4-mini之間則未呈現顯著差異（$p = .602$）。整體而言，在以人工標註一致的題項作為效度基準的條件下，gpt-5.4在框架分析任務中的表現最佳，而gpt-4.1-mini與gpt-5.4-mini的效度表現則相近，此一差異顯示，在高詮釋需求任務中，模型間的差異主要體現在效度層面，而非穩定性本身。進一步而言，gpt-5.4作為推理模型，其效度優勢反映其在高詮釋需求任務中，對語境線索進行整合與套用分析框架的推理能力。

六、跨任務詮釋需求下之穩定性與效度的結構性差異

為回答研究問題二，接續結構性地整理三種不同詮釋需求層級任務的穩定性和效度的差異。在低詮釋需求的立場表述任務中，模型普遍展現極高的穩定性（多數條件下MRP接近1.000），但此種穩定性高度依賴

提示詞中的局部語詞與語言版本，並反映模型既有語料所形成的立場先驗，而非對任務本身的判斷能力。因此，此一任務中的高穩定性，更多對應於「生成慣性」而非「分析效度」。

當任務提升至中度詮釋需求的立場判斷時，模型仍維持整體上高度穩定（MRP約0.930～0.990），但已開始出現隨溫度與語彙表述變動的穩定性下降，顯示模型回應開始受到語用理解與語境整合負荷的影響。更關鍵的是，在此任務中，穩定性與效度之間出現明顯分離：部分模型雖維持高度一致的輸出，卻未必對應較高的分類正確率，顯示模型可能穩定地重複同一判斷偏誤，而非提升分析品質。

進一步在高詮釋需求的框架分析任務中，模型穩定性並未如預期下降，且亦未隨人類標註分歧而系統性變動，顯示詮釋困難度未必直接轉化為生成不穩定。然而，不同模型之間的差異則主要體現在效度層面：具推理能力的模型在框架判斷上呈現較高準確率，而非推理模型則與其速效版本差異有限。換言之，在高詮釋需求任務中，模型能力差異主要反映於「是否能正確套用抽象分類準則」，而非「是否能穩定產生一致答案」。

整體而言，研究結果顯示，LLM在文本標註任務中的表現，並非沿單一維度隨任務難度單調變化，而是呈現一種「穩定性與效度分化」的結構性模式。在低詮釋需求任務中，穩定性高但易受語詞觸發；在中度詮釋需求任務中，穩定性略降且開始與效度分離；在高詮釋需求任務中，穩定性維持但效度差異擴大。此一結果說明，LLM作為文本分析工具時，其方法論評估需同時考量任務詮釋需求、模型能力與研究設計條件，而不能將穩定性或準確率單獨視為模型表現的代表指標。

伍、總結

本研究以LLM在文本標註任務中的回應穩定性與方法論效度為核心，系統性比較模型控制因素、題項設計因素與任務詮釋需求層級如何共塑模型表現。整體研究結果顯示，LLM的輸出並非單純反映模型能力，而是受到提示詞局部語詞、語言版本、模型訓練背景、作答程序安排與任務性質的共同影響。此一結果與近年LLM輔助質性編碼研究的共同觀察一致：模型表現高度依賴提示詞設計、編碼簿設計與任務本身對脈絡理解的要求，而非僅取決於模型是否較新或參數規模是否較大（Chew et al., 2023; Dunivin, 2025; Hou et al., 2024）。

首先，在低詮釋需求的立場表述任務中，本研究發現提示詞中的特

定文字線索會成為觸發模型既有立場先驗的條件，而非中性背景資訊。這表示，研究者若未事先檢驗提示詞中關鍵字或措辭的效果，便可能將提示詞所誘發的生成傾向誤判為模型對公共議題的穩定立場。其次，在模型控制因素上，不同模型版本背後可能連結到不同語料叢集與語言的生成慣性，其影響常大於溫度設定本身。特別是在文字線索較充分的情境下，推理模型gpt-5.4出現明顯的中英語言翻轉，顯示語言不僅是表達媒介，也可能影響模型所調用的知識脈絡及生成方向，因此中英互譯不宜被視為理所當然的等值題項。

再者，在研究題項設計上，本研究顯示影響穩定性的關鍵不在量表長度，而在題項要求是否與模型的能力特性相容。當非推理模型被要求先展開推理再作答，或推理模型被強制置入特定作答格式時，不穩定情形顯著增加。換言之，模型表現的差異不在於推理能力高低，而應理解為研究設計是否與模型回應機制相匹配。溫度在此較像是放大原有不匹配，而非獨立造成不穩定的主因。

當任務由立場表述轉為立場判斷時，模型的作答邏輯也出現明顯變化。在立場表述任務中，模型幾乎不使用五點量表兩端的極端選項；但在立場判斷任務中，模型不再系統性迴避極端選項，而會依文本強度分層使用不同類別。該結果指出，相同模型在「以模仿自身名義表態」與「判讀他人文本立場」兩種不同任務下，其生成的量表選擇策略並不相同。

此外，本研究也顯示，說理要求的影響不能只以「有無理由」粗略區分，而應區分為兩個層次，一是是否要求說理，二是說理與作答的先後順序。在立場相關任務中，直接要求模型做出立場判斷，通常比要求其同時或事前生成理由，具有更高的穩定性與效度；而若必須加入理由，則「先表態、後補理由」通常優於「先說理、再判斷」，也優於將理由生成與分類拆為兩階段處理。這顯示理由生成並不必然提升判斷品質，反而可能成為額外干擾來源。近年研究亦指出，要求模型提供理由雖有時可提升某些編碼任務表現，但效果具有任務依賴性，並不能被視為普遍有效的優化手段（Dunivin, 2025）。

最後，在中高詮釋需求的立場判斷與框架偵測任務中，本研究進一步指出，穩定性與效度需被同時納入分析。模型可以高度穩定，卻只是穩定地重複同一偏誤。尤其在框架偵測任務中，人類編碼者是否一致，並未直接對應模型是否較不穩定；模型差異更明顯地體現在效度，而非穩定性本身。整體而言，本研究說明，LLM用於文本標註時，其方法論評估應同時評估穩定性與準確率指標，而必須同時考慮任務詮釋需求、

提示詞設計、模型特性與人工判準之間的關係。這也呼應既有研究對 LLM 輔助質性分析的共同主張，其價值不在於取代研究者，而在於在經過適當設計與驗證後，成為可被納入研究流程的協作工具（Shin, 2025; Wen et al., 2026）

陸、討論

本研究最重要的方法論意涵，在於說明LLM於文本標註中的生成差異，不宜被理解為單純的「模型強弱」問題，而應被理解為研究設計與模型機制之間的適配問題。既有研究已多次指出，LLM 在質性或演繹式編碼中的表現，會隨提示詞設計、編碼本寫法、是否提供範例、任務是否依賴上下文而顯著波動（Chew et al., 2023; Hou et al., 2024; X. Liu et al., 2025）。本研究進一步系統化地將這個觀點推進到更細的層次，發現即便在相同主題、相同模型家族下，只要改變語言版本、局部措辭、說理要求與作答程序，就足以改變模型的穩定性與效度。這表示，LLM的研究用途不能抽離具體題項條件來談，研究者也需謹慎審視模型生成結果背後的設計脈絡。

其次，本研究結果對當前常見的「要求模型說明理由，以提升透明性與可信度」的做法提出了更審慎的修正。從可解釋性角度來看，要求模型產生理由看似有助於研究者理解其判斷依據；然而，本研究顯示，在立場任務中，理由生成未必提升分類表現，甚至可能引入額外干擾。這一點與部分既有研究形成有條件的一致：有研究發現階段式推理在某些編碼任務中確能改善與人類的一致程度，但這種效果仰賴任務性質與模型控制和研究設計等可操作性因素，並不保證在所有任務上都成立（Dunivin, 2025）。因此，從方法論角度看，說理不是普遍增益機制，而是一項需要被獨立驗證的設計變項。當任務本身已涉及隱含立場、反諷、政治攻擊或語用轉折時，理由生成有可能促使模型以更流暢的語言合理化既有偏誤，而不是更接近人工判準。

第三，本研究有助於重新界定「穩定性」在LLM研究中的地位。許多研究將高一致性或高重複性視為模型可用性的證據，但本研究顯示，穩定性僅能說明模型在相同條件下會反覆產生相似答案，卻不能單獨證成其方法論效度。此一觀察再次強調了效度測試的重要性，模型不僅要能穩定輸出，更必須經過隨機性檢驗、人工複核與與人類編碼結果的比較，才能判斷其輸出是否真正對應研究者欲測量的構念（Chew et al., 2023）。因此，

在立場判斷與框架分析等高詮釋需求任務中，將MRP或其他穩定性指標單獨用作品質保證，可能反而掩蓋模型穩定重複偏誤的風險。

第四，本研究提出的「任務詮釋需求層級」具有理論與方法上的雙重意義。這一概念的作用，不只是把不同文本標註任務依照難易程度加以排序，更重要的是指出，LLM在不同任務中的限制來源其實並不相同。在低詮釋需求任務中，主要風險來自提示詞觸發效果與文字表述的細微偏移；在中度詮釋需求任務中，主要瓶頸則轉向語用推論、隱含立場辨識與脈絡補足；而在高詮釋需求任務中，關鍵限制進一步表現在模型是否能掌握框架邊界、論述組織方式，以及特定社會語境中的意義取捨。這一發現與既有文獻的觀察相互呼應，亦即，越依賴脈絡整合與上下文理解的編碼任務，通常比主要根據文本表面資訊即可完成的編碼任務更困難；同時，LLM在不同質性編碼任務上的表現，也往往呈現明顯不均的情況（Hou et al., 2024; X. Liu et al., 2025）。因此，本研究的貢獻不只是指出哪一個模型在某一任務上表現較佳，而是進一步提出一個更能解釋模型表現差異的方法論框架，即「模型表現必須放在任務詮釋需求與研究設計配置之中理解」。

最後，這些結果也對未來使用LLM進行社會科學文本分析提出較明確的實務原則。首先，研究者不應將提示詞語言視為中性的技術包裝，而應將其視為研究設計的一部分，並加以系統性檢驗。其次，在涉及中高詮釋需求的任務時，穩定性與效度應分開報告，不能以其中一者替代另一者。再次，當任務涉及高語境依賴、高語用負荷或複雜詮釋邊界時，人工複核仍不可取代；LLM較適合被定位為人機協作流程中的輔助編碼者、初步篩選工具或對照判準，而不宜被直接視為可獨立取代研究判斷的研究代理。近年相關研究也普遍指出，LLM較合理的角色，是納入人工監督與判斷之下的協作夥伴，而不是脫離研究者控制、獨立完成判讀的自動分析者（A. Liu & Sun, 2025; Shin, 2025; Wen et al., 2026）。

參考文獻

Abdurahman, S., Ziabari, S. A., Moore, A. K., Bartels, D. M., & Dehghani, M. (2025). A primer for evaluating large language models in social-science research. *Advances in Methods and Practices in Psychological Science*, *8*(2), 25152459251325174. doi:10.1177/25152459251325174

Cardamone, N. C., Olfson, M., Schmutte, T., Ungar, L., Liu, T., Cullen, S.

W., ... Marcus, S. C. (2025). Classifying unstructured text in electronic health records for mental health prediction models: Large language model evaluation study. *JMIR Medical Informatics*, *13*, e65454. doi:10.2196/65454

Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. (2023). *LLM-assisted content analysis: Using large language models to support deductive coding*. arXiv. doi:10.48550/arXiv.2306.14924

De Paoli, S. (2024). Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review*, *42*(4), 997-1019. doi:10.1177/08944393231220483

Dunivin, Z. O. (2025). Scaling hermeneutics: A guide to qualitative coding with LLMs for reflexive content analysis. *EPJ Data Science*, *14*(1), 28. doi:10.1140/epjds/s13688-025-00548-8

Gao, J., Guo, Y., Lim, G., Zhang, T., Zhang, Z., Li, T. J.-J., & Perrault, S. T. (2024). CollabCoder: A lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. In F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. T. Dugas, & I. Shklovski (Eds.), *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-29). New York, NY: Association for Computing Machinery. doi:10.1145/3613904.3642002

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(30), e2305016120. doi:10.1073/pnas.2305016120

Hila, A., & Hauser, E. (2025). Assessing the reliability of large language models for deductive qualitative coding: A comparative intervention study with ChatGPT. *Proceedings of the Association for Information Science and Technology*, *62*(1), 275-285. doi:10.1002/pra2.1255

Hou, C., Zhu, G., Zheng, J., Zhang, L., Huang, X., Zhong, T., ... Ker, C. L. (2024). Prompt-based and fine-tuned GPT models for context-dependent and-independent deductive coding in social annotation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 518-528). New York, NY: Association for Computing Machinery. doi:10.1145/3636555.3636910

Kim, H., Choi, Y., Yang, T., Lee, H., Park, C., Lee, Y., ... Kim, J. (2024). *Using*

LLMs to investigate correlations of conversational follow-up queries with user satisfaction. arXiv. doi:10.48550/arXiv.2407.13166

Liu, A., & Sun, M. (2025). From voices to validity: Leveraging large language models (LLMs) for textual analysis of policy stakeholder interviews. *AERA Open*, *11*, 23328584251374595. doi:10.1177/23328584251374595

Liu, X., Zambrano, A. F., Baker, R. S., Barany, A., Ocumpaugh, J., Zhang, J., ... Wei, Z. (2025). Qualitative coding with GPT-4: Where it works better. *Journal of Learning Analytics*, *12*(1), 169-185. doi:10.18608/jla.2025.8575

Morgan, D. L. (2023). Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. *International Journal of Qualitative Methods*, *22*, 16094069231211248. doi:10.1177/16094069231211248

Pham, C. M., Hoyle, A., Sun, S., Resnik, P., & Iyyer, M. (2024). TopicGPT: A prompt-based topic modeling framework. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 2956-2984). Stroudsburg, PA: Association for Computational Linguistics. doi:10.18653/v1/2024.naacl-long.164

Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, *11*(6), 887. doi:10.3390/healthcare11060887

Shin, E. (2025). Co-coding classroom dialogue: A single researcher case study of ChatGPT-assisted analysis in science education. *Journal of Computer Assisted Learning*, *41*(4), e70089. doi:10.1111/jcal.70089

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods*, *23*, 16094069241231168. doi:10.1177/16094069241231168

Vossen, M. (2020). Nuclear energy in the context of climate change: A frame analysis of the dutch print media. *Journalism Studies*, *21*(10), 1439-1458. doi:10.1080/1461670X.2020.1760730

Wachinger, J., Bärnighausen, K., Schäfer, L. N., Scott, K., & McMahon, S. A. (2024). Prompts, pearls, imperfections: Comparing ChatGPT and

a human researcher in qualitative data analysis. *Qualitative Health Research*, *35*(9), 951-966. doi:10.1177/10497323241244669

Wen, C., Clough, P., Paton, R., & Middleton, R. (2026). Leveraging large language models for thematic analysis: A case study in the charity sector. *AI & Society*, *41*(1), 731-748. doi:10.1007/s00146-025-02487-4

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P.-Y. (2023). Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces* (pp. 75-78). New York, NY: Association for Computing Machinery. doi:10.1145/3581754.3584136

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, *50*(1), 237-291. doi:10.1162/coli_a_00502

附錄一：立場判斷任務之題項設計

立場類型	題項內容
明顯支持	綠能對臺灣毫無幫助，便宜穩定的核能才是臺灣發展的重要能源
	反核就是置臺灣於能源斷鏈的危險之中
	核一核二核三本來就可以正常使用
不明顯支持	臺灣廢核說實在就是政治操作而已
	自稱最愛臺灣的政黨卻為臺灣帶來戰爭時能源短缺的嚴重風險
	123不延役 核四總能評估啟用吧
無明確立場	不在意電費了，別再用核能撕裂臺灣了好嗎
	說個笑話，我們還曾經輸出核電經驗給韓國
	只要能夠好好評估，我相信民眾還是可以理智接受的
不明顯不支持	臺灣不缺電，說缺電的都是中共同路人
	國民黨什麼都反對就是要搞爛臺灣對不對
	核廢料就放在重啟核電同意票最高的地方
明顯不支持	我支持非核家園
	核廢料放誰家？
	臺灣持續反核 重啟個屁

附錄二：核能作為能源之九種新聞框架整理表

中文名稱	核心問題定義	道德／價值基礎	典型論述特徵（語言線索）	常見立場傾向
失控科技框架	核能是一種人類無法完全控制、潛藏巨大風險的科技	恐懼、預防原則、對未來世代的責任	災難、輻射、無法挽回、車諾比／福島、惡夢、定時炸彈	反核
進步框架	核能是現代化與科技進步的象徵，可促進經濟與社會發展	科學理性、技術樂觀主義	高科技、安全、效率、專業、創新、下一代反應爐	親核
永續發展框架	核能是否符合環境永續與生態保護	環境倫理、世代正義	乾淨能源、再生能源、核廢料、環境負擔	多為反核
生態現代化框架	為因應氣候變遷，必須依賴高科技（含核能）	氣候責任、科技解方	低碳、氣候危機、現實選擇、反對反核教條	親核
取捨權衡框架	能源政策需比較各種能源的相對成本與風險	相對理性、比較邏輯	比較死亡率、CO_2 排放、土地使用、數據對照	偏親核
成本效益框架	核能在經濟上是否划算	財政理性、公共資源效率	建設成本、補貼、電價、投資回收	立場不一
公共問責框架	政府與企業是否能妥善監管核能	權力制衡、民主監督	監管失靈、隱匿資訊、官商勾結	反核
社會正義框架	核能與能源分配是否涉及不平等	弱勢關懷、全球正義	貧窮、受害者、世代不公、犧牲區域	立場不一
能源自主框架	核能是否有助於國家能源安全	國家安全、主權	擺脫進口、地緣政治、自主供能	親核

資料來源：Vossen（2020）。

註：本表之框架定義與分類依據為Vossen（2020）對Gamson & Modigliani 框架傳統之延伸與更新。

附錄三：提示詞

立場表態提示詞
System Prompt: 你現在是一位臺灣民眾受訪者，正在參與一項關於核能政策的民意調查。 請根據題目設定的身分和情境，提供你的真實想法。 請直接回答問題，不要解釋或說明理由。 只需要從給定的選項中選擇一個答案。 User Prompt: {如果你是屏東縣民，}{退役的核三廠就在屏東縣}{核三廠已經退役}，你會支持重啟核三嗎？支持／不支持／無法判斷 Response Schema: ```json {"answer": "支持" | "無明確立場" | "不支持"} ```
立場判斷提示詞
System Prompt: 你是一位立場分析專家，專門判斷文字內容對特定議題的立場。 你的任務是準確判斷留言者對「重啟核三」的立場。 重要規則： 1. 僅根據留言內容判斷，不要加入自己的推測 2. 注意語氣和用詞的強度 3. 如果留言僅陳述事實或提問，而無明確立場，應判斷為「無明確立場」 4. 反諷或諷刺的語氣要特別注意其真實立場 User Prompt: 請判斷以下留言對「重啟核三」的立場，從以下選項中選擇：非常支持、支持、無明確立場、不支持、非常不支持。 【留言內容】 {item_text} 請根據留言內容判斷立場。 Response Schema: ```json {"stance": "非常支持" | "支持" | "無明確立場" | "不支持" | "非常不支持"} ```

框架偵測提示詞

System Prompt:
你是一位新聞框架分析專家，專門分析核能議題的論述框架。
你的任務是判斷給定的段落最符合哪一個新聞框架。

九大核能新聞框架定義：

1. **失控科技框架**：核能是人類無法完全控制的高風險科技
   - 核心問題：核能是一種人類無法完全控制、潛藏巨大風險的科技
   - 語言線索：災難、輻射、車諾比、福島、定時炸彈、無法挽回

2. **進步框架**：...

3. **永續發展框架**：...

…

9. **能源自主框架**：...

判斷原則：
1. 仔細分析段落的核心論述與用詞
2. 選擇最符合的單一框架
3. 評估你的判斷信心程度（高/中/低）

信心度評估標準：
- **高**：段落明確包含框架的核心問題、語言線索，且論述清晰
- **中**：段落部分符合框架特徵，但論述較為隱晦或間接
- **低**：段落僅有微弱的框架線索，可能存在其他解讀

User Prompt:

請判斷以下段落最符合哪一個新聞框架：

【段落內容】
{paragraph}

請直接選擇最符合的單一框架，不需說明理由。

Response Schema:

```json
{
  "frame": "失控科技框架" | "進步框架" | "永續發展框架" | "生態現代化框架" | "取捨權衡框架" | "成本效益框架" | "公共問責框架" | "社會正義框架" | "能源自主框架",
  "confidence": "高" | "中" | "低"
}
```

Examining Research Designs for LLM-Based Automated Text Annotation: An Empirical Analysis Centered on Task Interpretive Demands

Ji-Lung Hsieh
Associate Professor
Graduate Institute of Journalism
National Taiwan University

Introduction

Large language models (LLMs) are increasingly used not only for text generation but also for automated text annotation in social science research. Prior studies have shown that LLMs can perform sentiment analysis, stance detection, deductive coding, and document classification with moderate to high agreement relative to human coders in some contexts (Chew, Bollenbacher, Wenger, Speer, & Kim, 2023; Gilardi, Alizadeh, & Kubli, 2023; Hila & Hauser, 2025; A. Liu & Sun, 2025; Wen, Clough, Paton, & Middleton, 2026). Once LLMs are used as analytic instruments, the key methodological question shifts from apparent accuracy to the conditions under which their outputs are stable, interpretable, and methodologically defensible.

Existing evaluations often focus on accuracy, agreement, or prompt engineering, but less often distinguish among tasks with different levels of interpretive demand. Some tasks can be completed through explicit lexical cues, whereas others require contextual inference, implicit stance recognition, or abstract frame classification. Recent research also shows that LLM outputs are sensitive to prompt wording, language choice, instruction order, and uncertainty options such as “cannot answer” or “uncertain”(Dunivin, 2025; Kim et al., 2024; Shin, 2025; Ziems et al., 2024). These findings suggest that prompts are not neutral instructional tools; rather, they constitute an integral component of the research instrument itself.

This study therefore shifts the focus from model benchmarking to methodological conditions. It examines how LLM-based annotation is shaped by three dimensions: model-control factors, item-design factors, and task-specific interpretive demand. Using the public controversy over restarting Taiwan's Third Nuclear Power Plant as a common issue domain, the study compares three tasks—stance expression, stance detection, and frame identification—to clarify when LLM outputs become stable, unstable, valid, or misleadingly consistent.

Method

The study selected the controversy over whether Taiwan's Third Nuclear Power Plant should be restarted because it is socially salient, politically contested, and associated with multiple competing frames. This provided a common substantive setting for comparing tasks that differ in interpretive complexity while holding the issue domain constant.

Three annotation tasks were constructed. The first was a stance-expression task, in which the model was asked to state its own position on whether the plant should be restarted or not. This was treated as a low-interpretation task because it mainly involved generating a stance under a given prompt condition. The second was a stance-detection task, in which the model judged whether social media comments supported or opposed restarting the plant. This was treated as a medium-interpretation task because it required the model to infer another speaker's position rather than produce its own. The third was a frame-identification task, in which the model classified discourse segments into one of nine predefined frames. This was treated as a high-interpretation task because it required deductive application of abstract analytic categories to discourse, similar to prior work on automated coding and frame analysis (Chew et al., 2023; Xiao, Yuan, Liao, Abdelghani, & Oudeyer, 2023).

To examine how design conditions affect outcomes, the study manipulated two classes of explanatory variables. The first consisted of model-control factors, including model version, prompt language, and temperature. The three models compared were gpt-4.1-mini, gpt-5.4, and gpt-5.4-mini. The second consisted of item-design factors, including wording choices, lexical cues, response scales,

whether reasoning was required, and whether reasoning came before or after classification. These manipulations were informed by prior studies showing that even seemingly minor prompt changes can affect coding outcomes (Dunivin, 2025; Hou et al., 2024).

Stability was defined as the extent to which a model produced the same output under fixed input conditions across repeated runs. Validity was examined by comparing model outputs against human coding in the stance-detection and frame-identification tasks. In the frame-identification task, two human coders labeled the discourse segments, and the subset of items on which both coders agreed was used as a stricter benchmark for model accuracy. This design made it possible to distinguish internal consistency from external validity and to test whether these two properties moved together or diverged across tasks of increasing interpretive demand.

Results

The results show that LLM annotation behavior is not simply a reflection of general model quality. Rather, outputs are jointly shaped by prompt wording, prompt language, response procedure, model version, and task type. In the low-interpretation stance-expression task, specific lexical cues embedded in prompts activated pre-existing stance tendencies. Terms related to identity, decommissioning status, or plant location could shift the generated position, indicating that wording itself "materially" influences what the model appears to "behave." This supports prior arguments that prompt phrasing can change annotation outcomes even when the nominal task remains constant (Dunivin, 2025; Ziems et al., 2024).

Among model-control factors, model version and prompt language generally exerted a stronger effect than temperature. In particular, Chinese and English prompts could produce different response tendencies, suggesting that the model drew on partially different discourse associations across languages. Prompt language therefore cannot be treated as a transparent delivery medium. Temperature was less important as an independent source of instability; instead, it can be seen as an amplifier to increase mismatches between the task design and the model's response mechanism.

Item-design factors were equally consequential. The critical issue was not simply whether a scale was longer or shorter, but whether the design fit the model's way of responding. Requiring non-reasoning models to explain before answering, or forcing reasoning models into overly rigid structured outputs, often reduced stability. More broadly, the results challenge the assumption that adding explanations automatically improves annotation. In stance-related tasks, directly requesting a judgment often produced higher stability and, in some cases, higher validity than requiring prior reasoning. When explanations were included, an answer-first, explain-later procedure generally performed better than an explain-first, answer-later procedure, and both typically outperformed split multi-stage procedures. This pattern suggests that generated explanations may function as post hoc rationalizations rather than transparent reasoning traces, a concern also raised by Wachinger, Bärnighausen, Schäfer, Scott, and McMahon (2024).

The results also indicate that models behave differently when expressing a stance versus detecting someone else's stance. In the stance-expression task, models tended to avoid the most extreme points on five-point scales, showing a centrist bias when generating their own position. In contrast, when judging social media comments, the same models did not consistently avoid extreme categories; instead, they used stronger labels when the textual evidence was clearer. This suggests that "stance expression" and "stance detection" should not be treated as interchangeable variants of the same task, because they involve different interpretive operations.

As interpretive demand increased, the relationship between stability and validity became more important. High stability did not guarantee valid annotation. A model could be highly stable while repeatedly making the same misclassification. In the frame-identification task, model differences were more visible in validity than in stability. GPT-5.4 achieved the highest frame-classification accuracy when compared against the subset of discourse segments on which two human coders agreed, outperforming the smaller models. Yet overall differences in repeated-run stability were limited. Moreover, categories that were more difficult for human coders did not map neatly onto lower model stability. Human ambiguity and model instability are therefore related, but not equivalent.

To sum up, these results support a layered interpretation of LLM-based

annotation. In low-interpretation tasks, the main risk lies in lexical triggering and prompt-framing effects. In medium-interpretation tasks, the main bottleneck shifts toward pragmatic inference and implicit stance recognition. In high-interpretation tasks, the challenge becomes the application of abstract coding categories to discourse-level meaning. Accordingly, model performance should not be discussed as a single stable property. It must be interpreted in relation to both the research design and the interpretive burden of the task.

Conclusion

This study demonstrates that the methodological evaluation of LLM-based text annotation should move beyond isolated performance metrics and instead treat model outputs as products of research design. Across stance expression, stance detection, and frame identification, the findings show that stability and validity are jointly shaped by model version, prompt language, lexical wording, response procedures, and task-specific interpretive demand. The main contribution of the study is therefore not simply to identify the strongest model, but to show that the meaning of model performance changes across different annotation tasks.

Several implications follow. First, prompt language and wording should be treated as part of the research instrument rather than as neutral technical packaging. Second, stability and validity should be reported separately, especially in tasks requiring contextual inference or interpretive coding. Third, reasoning should not be presumed beneficial; it should be treated as an empirical design variable whose effects depend on the task. Fourth, in high-interpretation settings such as frame analysis, human oversight remains essential. LLMs are better understood as collaborative tools for preliminary coding, screening, or comparison than as autonomous replacements for researcher judgment, consistent with recent discussions in LLM-assisted qualitative analysis (A. Liu & Sun, 2025; Shin, 2025; Wen et al., 2025).

The study also proposes task interpretive demand as a useful framework for understanding why different annotation tasks fail in different ways. Surface-level tasks are especially vulnerable to lexical triggering and wording shifts; medium-level interpretive tasks are constrained by pragmatic ambiguity and contextual inference; high-level interpretive tasks are limited by

abstract category boundaries and discourse-level meaning construction. This perspective offers a more explanatory methodological account than generic claims about one model being stronger or weaker than another.

Future work should test whether these patterns generalize beyond nuclear-energy discourse, compare deductive coding with more open-ended thematic analysis, and incorporate more systematic treatment of abstention, uncertainty, and coder disagreement. Longitudinal comparisons across model generations would also help determine whether the observed instabilities reflect enduring structural limitations or only temporary features of current model families. More broadly, the study argues that LLM-based automated text annotation can only be evaluated responsibly when model behavior is interpreted through the combined lenses of research design, uncertainty management, and task-specific interpretive demand.