・新聞學研究・　第一六七期
2026 年 4 月　頁 123-180　研究論文

橋接統計與詮釋：大語言模型輔助內容分析用於語料庫輔助論述分析的概念驗證

柯鐘晏*

投稿日期：2025 年 6 月 29 日；接受刊登日期：2025 年 12 月 8 日。

* 柯鐘晏為國立政治大學傳播學院博士，e-mail: duress.ko@gmail.com，ORCID: 0009-0004-6065-7749。

本文引述格式：
柯鐘晏（2026）。〈橋接統計與詮釋：大語言模型輔助內容分析用於語料庫輔助論述分析的概念驗證〉，《新聞學研究》，167，123-180。https://doi.org/10.30386/MCR.202604.0008

《摘要》

語料庫輔助論述分析研究（Corpus-assisted Discourse Studies, CADS）容易落入「統計編故事」的方法論陷阱，研究者看到統計模式就提出詮釋，卻未檢驗用來支持詮釋的樣本是否真的具有統計代表性。

本研究基於概念驗證（proof of concept），提出以大語言模型輔助的內容分析（Large Language Model-assisted Content Analysis, LACA）作為避開此陷阱的輔助方案，旨在評估 LACA 是否能透過批次檢查具統計模式樣本的實際語義，更有效整合語料庫分析與論述分析。

本研究建立了一套人機協作的 LACA 程序，透過實驗檢驗其作為 CADS 橋接機制的可行性。實驗結果顯示，此程序不僅能達到高編碼信度（Cohen's Kappa ≥ 0.80），且相較傳統內容分析能夠節省大量時間，使原本因成本限制而「理論上該做但實務上做不到」的檢驗變得可操作。

研究過程中進一步發現，LACA 超越批次語義檢驗工具的功能，除確認統計模式的語義基礎外，亦能協助研究者發現未預期的論述模式。本文亦論證 LACA 作為詮釋型資訊工具，研究者的詮釋邏輯能透過提示詞嵌入 LLM 的批次處理流程，實現厚描的批次化，橋接統計與詮釋，為混合方法研究的量質整合提供可操作的方法論框架。

關鍵詞：人機協作、大語言模型輔助內容分析、混合方法研究、詮釋型資訊工具、概念驗證、語料庫輔助論述分析

壹、研究背景與目的

「謊言有三種：謊言、該死的謊言、統計數字」，[1] 這句警語在語料庫輔助的論述分析（corpus-assisted discourse studies, CADS）[2] 中尤其值得注意。

理想上，呈現語料庫整體統計模式的語料庫分析（corpus analysis），能減少論述分析（discourse analysis）中人類普遍的認知偏見與研究者的主觀偏見（Baker, 2006, pp. 10-14）。然而，語料庫分析本質上只能得到非語義的詞頻統計模式（Gries, 2016, p. 11）。而在實務操作上，CADS 研究者雖然會透過關鍵詞脈絡（keywords in context, KWIC）來檢驗統計模式背後的實際語義，並根據 KWIC 挑選出來的樣本對所發現的模式進行論述詮釋；但這種檢驗往往沒有考量這些樣本的統計代表性，因此仍然落入研究者任意選擇支持其詮釋假設的樣本作為證據的「挑櫻桃」（cherry-picking）老路。這種情況體現了 Creswell & Clark（2017）所指出混合方法研究中的普遍問題：兩種方法的結果往往缺乏有效整合。CADS 一旦缺乏語料庫分析與論述分析之間的語義連結的系統檢驗，不但無法實現減少偏見的預期效果，反而可能落入曲解統計數字的詮釋陷阱。本文將這種陷阱稱作「統計編故事」（statistical storytelling）。

因此，CADS 需要處理一個根本問題：如何確保語料庫分析所得的統計模式確實反映了有意義的論述策略。以收錄 CADS 經驗研究的 *The*

[1] 此句原出處存疑，但通常被歸於 Mark Twain。

[2] 另一個常見的術語是基於語料庫的論述分析（Corpus-Based Discourse Analysis, CBDA）。

Routledge Handbook of Corpus Approaches to Discourse Analysis 一書為例，其導論即明確指出，研究者必須「盡可能準確且一致地」詮釋語料庫分析的結果。然而，該如何確保詮釋確實對齊了語料庫分析的結果？該文卻缺乏具體說明（Friginal & Hardy, 2020）。而且在這本手冊中，也收錄了缺乏這種對齊檢驗的研究，例如 Vine（2020）雖然聲稱整合了量化與質化分析來探討職場論述中 eh 標記（紐西蘭英語中的一個特殊語用標記）的使用，但即使拿掉量化部分，其質化分析也能單獨成立；Williams（2020）在其對美國國會英語統一法案辯論語料庫的論述研究中展示其分析過程時，直接從語料庫的關鍵詞與共現分析跳到論述分析的意識形態批判，缺少明確檢驗過程來確保兩者之間的語義連結。這些案例顯示，即使在進階參考書所收錄的經驗研究中，語料庫分析與論述分析之間的對齊檢驗仍會被省略，從而削弱了這些研究的信度，即使 CADS 的方法論設計原本是為了補強信度。

另一方面，當然也有許多研究精心設計各種方式，檢驗語料庫分析與論述分析之間是否對齊，但整體而言，這些方式並未形成標準。例如 Gabrielatos & Baker（2008）透過研究人員對該研究核心關鍵詞的 1.6 萬條 KWIC 進行逐行檢視，以確定這些關鍵詞被量化的頻率，使這種量化檢視與研究中用於分析語義與論述韻律的質化檢視（手動檢視數百條）區分開來；除此之外，還比較主流媒體新聞語料庫中的詞彙模式與這些詞彙的字典定義以及新聞編輯原則的差異。然而，前一種檢驗法缺乏樣本量如何決定的說明，難以評估其充分與否；而後一種檢驗法，僅適用於主流媒體新聞這樣規範化的語料庫，至於非規範的語料庫，像是充斥大量諷刺、幽默、次文化與新興詞彙的社交媒介評論，就會需要一種能處理語料複雜意義且標準化的檢驗法。

內容分析就是一個成熟的解決方案：它透過系統化的編碼程序與信

度檢驗機制，將詮釋過程標準化，從而能規模化處理相對複雜且非規範的語料。此外，內容分析將信度視為有效詮釋的基本條件，[3]這個方法論立場也與 CADS 相呼應。相較於個別研究者透過 KWIC 任意挑選樣本作為論述詮釋的證據，內容分析可作為使語料庫分析與論述分析更有效整合（integration）的中介取徑，正如 Creswell & Clark（2017）的混合方法論指出，整合是混合方法的核心特徵，而質化編碼與主題分析的結果與量化資料的結合，是實現有效整合的關鍵步驟之一。

CADS 的一般操作程序是：首先，透過語料庫分析挑出高頻關鍵詞，或與某個檢索詞之間具有顯著統計關係的共現詞，其次，研究者透過 KWIC 檢視這些關鍵詞在原本脈絡中的意義，並挑選一些合適的樣本作為某個論述詮釋的範例（郭文平，2020）。然而如前所述，這種作法有「統計編故事」風險。若在透過 KWIC 使語料庫分析的統計模式能夠有效整合論述分析的詮釋時，針對 KWIC 樣本進行基於隨機抽樣的內容分析，檢驗這些樣本的實際語義，可以有效降低這種風險。然而內容分析成本很高（Holsti, 1969），往往是小規模研究的 CADS 難以採用。

這個問題有望隨著預訓練大語言模型（pretrained large language models, LLMs）服務的普及而解決——讓 LLM 批次執行內容分析，而且已經有研究驗證了 LLM 作為內容分析編碼員的能力（Chew et al., 2023; Dunivin, 2025）。傳統內容分析的成本主要來自編碼標準的建立過程，需要至少兩位專家約定時間討論、達成共識、處理不一致案例，每次修改都需要重新協調和部分重新編碼，過程耗時。相較之下，LACA 提供編碼標準的即時迭代：修改提示詞後可立即批次編碼樣本並檢驗結果。對 CADS 研究者而言，LACA 提供一個低成本的檢驗方法，能夠確

[3] 參見 Krippendorff（2019）關於內容分析的詮釋本質的討論。

認具有語料庫統計模式的所有樣本的實際語義，從而確保用來支持詮釋的樣本具有統計代表性，降低「統計編故事」的風險。

據此，本研究提出大語言模型輔助的內容分析（LLM-assisted content analysis, LACA）作為 CADS 的批次語義檢驗機制。基於概念驗證（proof of concept）的研究設計，其中包含以下三個層次：一、在研究方法（method）層次，LACA 是一套可操作的程序，本研究透過實驗檢驗其技術可行性。二、在研究取徑（approach）層次，LACA 是整合語料庫分析與論述分析的中介取徑，本研究檢驗 LACA 應用於 CADS 時，如何能降低「統計編故事」風險。三、在方法論（methodology）層次，本研究討論此中介取徑的方法論意義。

貳、文獻回顧

如前所述，CADS 有「統計編故事」風險。傳統內容分析雖有助於研究者降低此風險，但成本太高。而本研究提出使用 LLM 作為解決此問題的方案。以下檢視相關文獻說明之。

近年來，LLM 在語義判斷任務上的表現突飛猛進，學界開始探索其在研究方法中的應用潛力。現有文獻主要呈現三條不同取徑。作為量化方法論的延伸，第一條取徑關注 LLM 執行批次任務的表現，如使用 LLM 模擬不同人類子群體對大型社會調查的反應（Argyle et al., 2023）、執行規模化的情感分析（Zhang et al., 2024）與 LLM 輔助內容分析（LLM-assisted content analysis, LACA, Chew et al., 2023）等。

第二條取徑關注 LLM 在語料詮釋任務中可扮演的角色與限制。例如將 LLM 定位為編碼助手（Chubb, 2023）、語用分析工具（Yu et al., 2024）或激發新詮釋角度的對話夥伴（Hayes, 2025），但都指出需要注

意 LLM 處理文化脈絡、隱喻等複雜語用現象的能力邊界。

第三條取徑關注 LLM 的批次與詮釋能力的整合，如 Dunivin（2025）的 LACA（該研究稱作 LLM 質化編碼，LLMs for qualitative coding）強調結合 LLM 的批次能力與傳統內容分析編碼的詮釋厚度，提出一個人機協作的編碼流程，由研究者根據對語料的深入詮釋，迭代設計提示詞，以供 LLM 批次編碼。

前述文獻支持了 LLM 作為批次語義編碼工具的可行性（Chew et al., 2023）以及提示詞設計對編碼能力的提升效果（Dunivin, 2025），但關於 LACA 如何用作 CADS 的檢驗機制，尚未得到探索。本研究嘗試回應這一空缺，採取概念驗證的研究設計，在控制條件下檢驗 LACA 應用於 CADS 的可行性。

參、研究問題

本研究以流行歌曲音樂影片〈玻璃心〉（*Fragile*）的二十餘萬則 YouTube 評論作為語料庫的 CADS 作為範例，檢驗 LLM 在特定條件下能否有效判斷具有語料庫分析統計模式的樣本的實際語義，使語料庫分析能夠與後續的論述分析更有效整合。

2021 年，華裔馬來西亞歌手黃明志（Namewee）與華裔澳大利亞歌手陳芳語（Kimberley Chen），在 YouTube 上發布了一首諷刺激進中國國族主義網民「小粉紅」的官話（Mandarin）流行歌曲音樂影片〈玻璃心〉，短時間內即獲得超過 3 千萬次播放和 20 萬則留言。初步觀察可以看出，這些評論反映該事件吸引了全球不同背景的參與者，其評論內容大多圍繞華裔身分認同（中國人、香港人、大馬華人、臺灣人……等等）進行討論。

為了確認「這個語料庫包含大量華裔身分協商內容」的假設，研究者首先可以進行詞頻分析。在去除數字、標點與介詞後，前 50 高頻詞彙，依序排列如表 1：

表 1：語料庫中前 50 高頻詞彙

#	詞彙	詞頻	#	詞彙	詞頻	#	詞彙	詞頻	#	詞彙	詞頻
01	你[a]	68,377	14	user	10,801	27	还	6,453	40	知道	5,044
02	我[a]	51,454	15	小粉紅[a]	9,386	28	没有	6,368	41	大陆[a]	5,037
03	人	24,488	16	聽	9,194	29	沒有	6,315	42	年	5,020
04	歌	18,154	17	台灣[a]	9,179	30	我們[a]	6,263	43	吃	4,870
05	首	14,368	18	你们[a]	8,387	31	来	6,139	44	很多	4,837
06	自己	13,961	19	想	7,911	32	听	5,786	45	个	4,808
07	中國[a]	13,754	20	中共[a]	7,305	33	没	5,766	46	洪水	4,726
08	看	13,101	21	台湾[a]	7,186	34	碎	5,683	47	一个	4,693
09	玻璃心	12,761	22	哈哈哈	7,042	35	做	5,657	48	自由	4,670
10	说	12,686	23	沒	6,729	36	我们[a]	5,619	49	爆炸	4,662
11	你們[a]	12,250	24	万	6,699	37	会	5,537	50	国家	4,636
12	中国[a]	11,896	25	the	6,668	38	中國人[a]	5,325	—	—	—
13	說	11,580	26	用	6,634	39	讓	5,144	—	—	—

註：詞彙依順序排列，[a]標示人稱代詞與華裔身分相關詞彙。
資料來源：本研究統計整理。

分析結果顯示，與華裔身分協商相關的指標詞彙高頻出現，包括人稱代詞（我、你、我們、你們）、諸華裔相關的政治實體與地理身分（中國、中国、中國人、中共、大陸、台灣、台湾）、相關的諷刺身分標籤（小粉紅）等。

由於人稱代詞是說話者用來表達身分與社會關係的有效語言指標（Davies & Harré, 1990/1999），本研究首先針對語料庫中以第一人稱代詞的自我身分宣告「我是」作為檢索詞進行共現分析，前 50 高機率（以 z-score 為準）詞彙，依順序排列如表 2：

表 2：在語料庫中與檢索詞「我是」共現的前 50 個詞彙

#	共現詞	Z-Score	#	共現詞	Z-Score	#	共現詞	Z-Score
01	我	20.112	18	你們	8.557	35	支持	6.516
02	人	20.011	19	大陆 [a]	8.201	36	自己	6.479
03	你	17.409	20	歌	7.960	37	越南人 [a]	6.479
04	中國台灣省 [a]	14.211	21	不知道	7.650	38	一个	6.428
05	说	12.416	22	覺得	7.584	39	台灣 [a]	6.377
06	說	12.341	23	玻璃心	7.559	40	华人 [a]	6.304
07	大陆人 [a]	11.780	24	觉得	7.451	41	馬來人 [a]	6.162
08	中國人 [a]	11.323	25	想	7.406	42	首	6.158
09	中国人 [a]	10.976	26	知道	7.229	43	小粉紅 [a]	6.135
10	user	10.219	27	蒙古人 [a]	6.998	44	中國海南省 [a]	5.999
11	台灣人 [a]	9.931	28	沒	6.930	45	听	5.999
12	看	9.853	29	香港人 [a]	6.915	46	来	5.844
13	說	9.835	30	但是	6.884	47	句	5.840
14	中國 [a]	9.327	31	爲	6.836	48	看到	5.794
15	高山族 [a]	9.110	32	你们	6.766	49	沒有	5.780
16	马来西亚人 [a]	9.051	33	个	6.656	50	出生	5.752
17	小粉红 [a]	8.739	34	中国 [a]	6.654	—	—	—

註：[a] 標示華裔身分相關詞彙。
資料來源：本研究統計整理。

結果顯示，與身分相關的政治實體、族裔與地域詞彙大量出現。然而高機率的詞彙共現只是語義關係的指標，並不等於語義關係。例如「反正賺了錢拿回台灣花我是」這樣的句子，在語料庫中也會提高「我是」與「台灣」共現的機率。CADS 研究者的通常作法是，利用 KWIC 回到這些詞彙原本的脈絡中找證據，例如「反正賺了錢拿回台灣花我是」這樣的句子果然是少數，而「我是美国华人，看不下去」這樣的句子是多數，就「證實」了在語料庫中，「我是」與「身分詞彙」的共現，確實是說話者在陳述自身身分的有效指標，基於此，進一步可以推

論其高頻出現（見表 1）「證實」語料庫中包含大量華裔身分協商相關內容的假設。

然而接下來的問題是，研究者如何在這個基礎上進行下一階段：針對語料庫中的華裔身分協商的論述模式與其意義進行論述分析。一般作法採用立意抽樣，研究者會透過 KWIC 精心挑選若干高品質樣本進行分析。然而如前所述，共現詞之間並不直接具有語義關係，基於共現分析的 KWIC 立意抽樣，有落入統計編故事的風險。而本研究在這裡故意選擇了「我是」這個與身分詞之間高機率具有語義連結的檢索詞，更容易使人忽視風險。以表 3 為例，由分析工具自動排序而非研究者立意選出的「我是」的左三右七的關鍵詞脈絡（KWIC），具體呈現「語料庫包含大量華裔身分協商相關內容」。

表 3：語料庫中「我是」的左三右七的關鍵詞脈絡條目示範

#	左三	檢索詞	右七
01		我是	94 年 出生 的 。
02	实际上 ，	我是	认识 他们 的 。
03	坦白 說	我是	同情 粉紅[a]的 ， 希望 有朝一日 自由民主
04	我 覺 的	我是	玻璃心 的 感覺 ， 網路 上 很多
05	， 恰恰相反 ，	我是	爱国者。
06	@ @ linallen7067	我是	吃 冷凍 菜 的 , @ @
07	@ @ stackerliew	我是	说 赌 场 , @ @ hugowinging
08	, "	我是	中国人[a] ， 只 觉得 你们 很 无聊
09	过 一样 ，	我是	不同意 ， 你 看 我 贴 的
10	， 当然 ，	我是	中国大陆人[a] 。

註：[a]標示與華裔身分相關的詞彙。
資料來源：本研究統計整理。

表 3 雖然只展示 10 條由語料庫工具自動選取的樣本，但其中已可見「我是」與「華裔相關身分」的語義關聯（如#08、#10）。如前所

述，許多研究會跳過隨機抽樣步驟，立意抽取這類樣本進行後續的論述分析，這正是本研究關注的方法論陷阱：在語料庫分析中，共現的詞彙如「我是 台灣」只是統計關係，不一定有語義關係，如「反正賺了錢拿回台灣花我是」這句話，「我是」與「台灣」的共現與「我」的身分無關。LACA 可協助研究者檢驗這些共現在實際脈絡中是否有語義關係，區分「我是台灣人」與「反正賺了錢拿回台灣花我是」，從而更有效地對齊語料庫分析與論述分析的結果。

據此，本研究採取概念驗證的研究設計，旨在檢驗 LACA 能否有效整合 CADS 中語料庫分析與論述分析。概念驗證的目的，在於建構並展示一個原型（prototype），以檢驗研究構想在特定條件下能否達成預期效果。在操作面向上，概念驗證的重點在於展示「技術可行性」或「理論可能性」，而非實現完整功能，並據此判斷是否值得進一步發展（Elliott, 2021）；而在方法論面向上，概念驗證的意義在於透過系統檢驗，闡明原型所帶來新的知識論範疇，這些範疇決定了未來研究的問題設定、方法選擇與知識生產方式（Kendig, 2015）。如前所述，本研究在驗證 LACA 在技術上能否降低 CADS 方法論風險的同時，也會探討研究過程中湧現的方法論意義。

具體而言，本研究選擇了一個語料中統計模式與實際語義高度對應的語言現象，並透過語料預處理與二元編碼設計，把任務聚焦在語義判斷上，以便在控制變因的情況下，專注評估大語言模型輔助的內容分析（LACA）使語料庫分析與論述分析更有效整合的能力。[4] 研究問題如下：

RQ1：LACA 能否有效識別語料庫中共現詞彙之間的語義關係？

[4] 二元分類可擴充為多層次檢驗，透過遞進的二元分類處理更複雜的編碼任務。關鍵在如何維持逐層篩選後樣本的統計代表性，而這在技術上是可行的。

RQ2：不同 LLM 模型與提示詞如何影響 LACA 的編碼效果？

RQ3：LLM 編碼與研究者標準的一致性如何？

RQ4：本研究提出的 LACA 檢驗機制能否有效橋接 CADS 中統計模式與論述詮釋，避免落入「統計編故事」陷阱？

肆、研究方法

本研究旨在評估 LACA 能否在特定條件下，有效整合 CADS 的語料庫分析與論述分析。近期文獻如 Chew et al.（2023）也報告了 LACA 實驗，但該研究聚焦於開發通用的 LACA。而且與本研究實驗所處理的非規範的 YouTube 評論語料不同，該研究實驗所處理的是相對規範的語料，包括美國總統推文、主流媒體 BBC 報導、部落格貼文與政府報告等。研究目的的差異也體現在提示詞設計（prompt engineering）上。提示詞設計是為 LLM 編寫明確的任務指令與判斷標準，使其能按照研究者的編碼邏輯執行批次任務。該研究的提示詞較為簡潔，僅要求 LLM 依預設的分類原則執行編碼（Chew et al., 2023, pp. 18-22）；相對而言，本研究所設計的提示詞更精細，不僅須明確說明編碼任務，更重要的是，本研究示範了如何透過提示詞設計，將理論與詮釋判斷嵌入 LLM 的批次任務中。例如本研究的任務涉及語法結構層次，像是否定結構、語序變異、間接指涉等；以及語用詮釋層次，像是識別宣告身分的言說行動，與非宣告的言說行動的差異（見附錄一）。

一、實驗設計

實驗步驟如下：

1. 選擇〈玻璃心〉的評論作為語料庫進行詞頻統計與共現分析。初步結果如前所述，在語料庫中，「我是」這個自我身分宣告指標，與「中國人、香港人、臺灣人」等華裔身分相關詞彙高度共現。接下來對於「我是」KWIC 條目的初步檢查指出，其中確實包含「我是中国人」這樣的自我身分宣告。根據初步分析結果，本研究進一步選擇「我是」、「你是」、「我們」、「你們」等用於自我宣稱與他人指涉的人稱代詞，作為下一階段 KWIC 的檢索詞。
2. 將「我是」等檢索詞在語料庫中的 KWIC 條目全數列出，包括這些檢索詞的左三右七個詞（如表 3），若有條目的所有字元與其他條目完全重複，則去除重複只留下一條。本研究假設，這些重複的條目是複製貼上的灌水評論。
3. 對得到的所有條目進行分層隨機抽樣，設定 95% 信心水準與 3% 抽樣誤差，語義相同的「我們／们」與「你們／们」依比例抽樣後合併。
4. 本研究透過人機協作建立標準編碼：首先，研究者先對所抽取的全部樣本進行人工編碼作為初始標準編碼，同時使用初版提示詞讓 LLM 對樣本進行 5 次重複編碼；其次，透過程式工具檢驗編碼結果的人機編碼與 LLM 內部編碼的一致性，計算 Cohen's κ（兩兩信度），並自動識別爭議樣本（出現任何不一致）；其三，研究者逐一檢視工具指出的爭議樣本，可能發現兩種情況。若是 LLM 編錯，則表示 LLM 未能正確識別語料脈絡，需修改提示詞；若是研究者編錯，則需修改標準編碼；其四，使用修改後的提示詞，讓

LLM 重新進行 5 次獨立編碼，再次透過工具檢驗其與標準編碼的一致性；其五，重複步驟二到四，直到研究者判斷編碼原則足夠明確可操作且 κ 值穩定達標（≥ 0.8）。

5. 設計兩組控制實驗，實驗一比較不同 LLM 的編碼能力；實驗二檢驗提示詞精細化對 LLM 編碼效果的影響。實驗程序包括：首先，使用先前建立的標準編碼，對不同模型及提示詞組合進行二元編碼測試，判斷檢索詞（人稱代詞）是否與身分相關詞有語義關係；其次，計算各條件下的編碼一致性（LLM 與人類、LLM 內部），並進行統計比較。

二、語料預處理

本研究採取 2021 年 10 月 15 日至 2022 年 3 月 4 日期間〈玻璃心〉YouTube 官方音樂影片的 209,744 則留言作為語料庫，其中包含 73,760 則初階留言以及 135,984 則二階留言，包含約 230,000 個詞彙。

本研究使用 CORPRO 作為語料庫分析工具（Chueh & Chen, 2016）。語料預處理程序如下：首先，漢字語料有別於拼音文字，詞彙之間不以空格分隔，因此需要透過工具加以「斷詞」（word segmentation），將連續字串切分為語義單位。此外，本研究透過 CORPRO 的「自建辭典」功能，將與研究主題相關的新詞（如「小粉紅」）納入詞典，以提升斷詞準確度。

其次，完成斷詞後，透過關鍵詞脈絡（KWIC），以左三右七的窗口大小，提取包含「我是、你是、我們／们、你們／们」等檢索詞的所有 KWIC 條目。窗口大小依據呈現語義關係所需的脈絡範圍設定。

最後，為確保資料品質，若發現哪些條目的所有字元與其他條目完全重複，就只保留一條。本研究假設這些重複條目是複製貼上的灌水評

論。

三、抽樣方法

採用隨機抽樣，設定 95%信心水準與 3% 抽樣誤差。由於語義相同，「我們／们」與「你們／们」先按比例分層抽樣，再合併用於編碼，以確保繁簡體在樣本中的比例與語料庫中的原始比例一致。抽樣不僅能確保樣本的統計代表性，還能確保人機協作的可行性；研究者需要檢視樣本以建立標準編碼、迭代提示詞設計、對爭議案例進行人工判斷，而抽樣能讓樣本數控制在人工完全能處理的範圍內。[5] 結果見表 4：

表 4：關鍵詞脈絡（KWIC）語料條目抽樣統計

檢索詞	原始條目數	去重後條目數	分層樣本數	分層比例	合併樣本數
我是	3,241	2,911	781	100%	781
你是	3,155	3,086	793	100%	793
我們	6,263	4,780	497	49.84%	980
我们	5,619	4,821	483	50.26%	
你們	12,250	11,135	585	57.81%	1,013
你们	8,387	8,147	428	42.29%	
總計	38,915	34,880	3,567	—	3,567

註：依 95% 信心水準與 3% 抽樣誤差分層隨機抽樣。
資料來源：本研究統計整理。

[5] 在當前 LLM 對電力需求劇增的背景下，採取隨機抽樣亦具有環境倫理的意義。相較於對 20 萬則評論進行暴力計算所需的電力消耗，抽樣（3,567 條樣本僅佔原始 20 萬則評論數的 1.7%）能大幅減少耗能需求。

四、人機協作的內容分析標準編碼

本研究採用人機協作的迭代提示詞設計建立標準編碼，而非傳統內容分析由雙人共識建立的黃金標準——由兩位獨立編碼者分別編碼後，透過討論達成共識所建立——作為評估編碼信度的基準。這個選擇基於以下考量：除了傳統黃金標準的建立面臨高成本的限制外，更根本的問題是，黃金標準編碼本無法直接當作提示詞使用。正如 Dunivin（2025）實驗所顯示，必須把黃金標準編碼本「轉換」成針對 LLM 優化的提示詞，才能得到夠高的編碼信度。然而這種轉換本身就是對黃金標準的再詮釋，如此一來「轉換後的標準」與「直接針對 LLM 設計的標準」，其實質差異為何？若編碼結果有效，其有效性的真正來源是黃金標準的忠實轉換，還是符合了 LLM 的處理邏輯，難以單獨歸因。另一方面，專為 LLM 設計的人機協作標準，可結合研究者的詮釋判斷與 LLM 的批次處理能力，透過快速迭代實現提示詞的優化，同時避免傳統黃金標準的固有限制，與編碼本轉換所帶來的再詮釋問題。此外，相較於傳統雙人黃金標準的協商過程，提示詞版本的迭代本身即為可檢驗的決策記錄，使得人機協作的方法透明度更高。

五、模型與提示詞設計

本研究設計的實驗包含兩項變因，模型效能與提示詞設計。在模型方面，選用 Anthropic Claude 的 Haiku 3.5 與 Claude Sonnet 4——前者處理速度快且成本低，後者具備較強的語義理解與推理能力——以比較不同模型對編碼效果的影響。

在提示詞設計方面，本研究設計了簡易提示詞與精細提示詞（見附錄一），統一由 Sonnet 4 進行編碼，以檢驗不同提示詞設計對編碼效果的影響。

為提高操作效率，本研究開發了輔助工具，用於檢驗編碼效果，並評估實驗變因是否顯著影響編碼效果（見附錄二）。此外，所有 LLM 批次編碼作業均透過網頁介面完成，無須調用 API。[6]

六、信度與統計分析

本研究的信度評估是以 Cohen's κ 衡量模型與「標準編碼」之間的一致性。標準編碼為前述人機協作迭代並由研究者最終決定的過程所建立。模型內部一致性則讓同一模型對相同語料獨立編碼 5 次，計算各次編碼結果間的兩兩 κ 值，取平均作為一致性指標。由於本研究選用的模型在每次編碼時不會保留前次編碼的記憶，[7]因此重複編碼能反映模型的穩定性。

為提高估算可靠性，本研究採 bootstrap 進行 1,000 次重抽，並對 κ 值作 Fisher's z 轉換以符合常態假設。統計分析包括：（1）轉換後的獨立樣本 t 檢驗，比較兩組提示詞下模型與標準編碼的 κ 值差異；（2）Z

[6] 本實驗將 KWIC 樣本，以「我是」的 781 條為例，切割為適當大小如 100 條一個的純文字檔（txt），透過 Claude 網頁介面（能夠處理上傳的文字檔）依序進行批次編碼。配合本研究開發的編碼品質檢驗工具（見附錄二），整個流程（5 次編碼）可在一小時左右完成。在本實驗技術條件中，這種處理方式更能確保 LLM 在執行大批次任務時維持編碼標準的一致性。

[7] 2026 年 4 月，主流模型如 ChatGPT、Gemini、Claude 都已經支援跨對話的長期記憶，而且在付費狀況下，大多預設開啟跨對話記憶模式。像 Claude 這樣預設關閉、需要使用者主動啟用的屬於少數。研究者在選用模型進行 LACA 時需要注意這一點。

檢驗，評估兩組不一致率（任何模型編碼與標準編碼不一致的樣本數除以樣本總數）的標準化比例差。Z 檢驗專用於不一致率比較，以補充 t 檢驗的穩健性。多重比較以 Benjamini–Hochberg 法進行 FDR 校正，顯著標準為 $p < 0.05$（*）與 $p < 0.001$（**）。

為降低技術門檻並提高方法的可複製性，本研究開發了輔助工具來執行上述統計分析。研究者可根據不同階段需求選用（見附錄二）。

伍、研究發現

以下為研究發現，包括展示實驗結果，界定 LACA 的適當配置條件，並說明選擇依據；在應用示範部分，展示經 LACA 檢驗的樣本如何能夠支持後續論述分析，降低「統計編故事」風險。

本研究在 2025 年 5 月進行以下實驗：（1）評估兩個模型（Haiku 3.5 與 Sonnet 4）在使用相同提示詞（精細）識別檢索詞「我是、你是、我們／们、你們／们」與身分相關詞彙的語義關係的編碼效果差異；（2）檢驗同一模型（Sonnet 4）使用不同提示詞設計（簡易與精細）的編碼效果差異。

具體編碼任務為，將人稱代詞指涉特定身分的樣本編為 A（如「台灣人，你們讓人喜歡」），否則編為 B。此任務的關鍵在於區分單純的詞彙共現與實際的語義指涉關係。因此，編碼標準不僅要求將無身分詞共現的樣本（如「你們讓人喜歡」）編為 B，更要求將人稱代詞與身分詞共現但未指涉該身分詞的偽陽性樣本（如「你們喜歡台灣人」）編為 B。編碼結果的信度高低將驗證 LACA 機制能否有效處理此類判斷，以保證 A 樣本的語義有效性。

實驗樣本取自語料庫中各檢索詞 KWIC 左三右七條目的抽樣結果

（95% 信心水準，3% 抽樣誤差），各檢索詞樣本數介於 781 至 1,013 條。信度分析採用 Cohen's κ 值，分別計算 LLM 與標準編碼間的一致性、LLM 重複編碼的內部一致性，以及各條件下的 LLM 與標準編碼間的不一致率（任何編碼與標準編碼不一致的樣本數量除以樣本總數），並使用 Fisher's z 轉換後的獨立樣本 t 檢驗和 Z 檢驗評估比較條件間表現差異的顯著程度。

一、實驗一：模型選擇對編碼效果的影響

實驗結果顯示兩個模型在編碼表現上有顯著差異（見表 5）。Sonnet 4 在四個檢索詞 KWIC 樣本的編碼上達到高度一致的水準（κ = 0.869-0.979），顯示該模型能可靠識別樣本語義。

相較之下，Haiku 3.5 的表現變異較大，且變異程度隨語料複雜度而不同。在語義關係最清晰的「我是」樣本上，達到幾乎完美一致性（κ = 0.947），但在最模糊的「你們／们」樣本上，大幅下降至未達內容分析所需的信度標準（κ = 0.380）。

雖然模型間 κ 值比較的結果多數未達顯著水準，但所有不一致率比較均達到高度顯著（z = 4.38-21.12, 所有 $p < 0.001$），驗證兩個模型在編碼品質上的差異（見表 7）。

這些結果表明，LLM 能夠有效輔助內容分析工作，但並非所有模型都適用——如本研究實驗中的 Haiku 3.5，在處理複雜語料時編碼變異性顯著偏高。因此，研究者需要根據語料特徵選擇適當的模型。需要強調的是，所謂「語料特徵」是模型與語料之間的關係屬性，研究者基於對語料的初步理解所做的模型選擇，仍須透過實際檢驗信度，以確認適用與否。

二、實驗二：提示詞設計對編碼效果的影響

由於用來編碼語料的提示詞必須根據語料的不同而調整，本實驗比較同一模型（Sonnet 4）使用簡易或精細提示詞進行編碼的效果。為簡化實驗，本研究選取編碼難度最簡單的「我是」與最複雜的「你們／们」進行比較，最大程度地凸顯提示詞設計對編碼效果的影響。

實驗結果顯示，精細提示詞能夠改善編碼品質，這種效果在不同編碼難度的語料中都得到驗證（見表 6）。在較簡單的「我是」KWIC 樣本上，使用簡易提示詞已能達到優秀的一致性水準（$\kappa = 0.924$），精細提示詞將其進一步提升到幾乎完美的水準（$\kappa = 0.979$），這種改善在內部一致性上達到顯著差異（$p = 0.032$）。

提示詞的影響在複雜語料上更為顯著。在「你們／们」樣本上，精細提示詞將一致性從 $\kappa = 0.705$ 提升至 $\kappa = 0.869$。雖然提示詞間 κ 值比較多未達顯著差異，但不一致率的比較均達高度顯著（「我是」$z = 6.05, p < 0.001$；「你們／们」$z = 4.08, p < 0.001$）（見表 7），驗證了提示詞精細化的效果。

這些結果表明，即使是表現優異的 LLM（如 Sonnet 4），仍可透過提示詞設計進一步提升其編碼結果。特別是在處理複雜語料時，提示詞設計對於確保編碼品質更為重要。同樣需要強調，提示詞的簡單與精細在描述層次可透過判斷條件維度（如語法、語義、語用維度）與條件數量的多寡來區分（見附錄一），但哪種「更合適」仍取決於特定的語料與模型組合，並透過 κ 值來實際評估。

表 5 與 6 呈現兩項實驗的結果，表 7 彙整兩項實驗的顯著性檢驗結果。

表 5：LLMs（Haiku 3.5 與 Sonnet 4）的編碼效果比較

KWIC 檢索詞	模型	提示詞	樣本數	與標準編碼 κ	內部 κ	不一致率
我是	Haiku 3.5	精細	781	.947	.938	5.9%
	Sonnet 4			.979	.987	1.7%
你是	Haiku 3.5		793	.596	.643	37.1%
	Sonnet 4			.883	.876	9.2%
我們／们	Haiku 3.5		980	.662	.652	32.5%
	Sonnet 4			.915	.908	6.2%
你們／们	Haiku 3.5		1,013	.380	.526	51.3%
	Sonnet 4			.869	.860	8.4%

資料來源：本研究統計整理。

表 6：不同提示詞（簡易與精細）的編碼效果比較

KWIC 檢索詞	模型	提示詞	樣本數	與標準編碼 κ	內部 κ	不一致率
我是	Sonnet 4	簡易	781	.924	.920	8.3%
		精細		.979	.987	1.7%
你們／们		簡易	1,013	.705	.703	14.1%
		精細		.869	.860	8.4%

資料來源：本研究統計整理。

表 7：LLMs 與提示詞比較的顯著性檢驗彙整

實驗	KWIC 檢索詞	比較條件	與標準編碼 κ 比較 t（p 校正後）		內部 κ 比較 t（p 校正後）		不一致率比較 z（p 校正後）	
			t(8)	p	t(8)	p	z	p
模型（實驗一）	我是	Haiku 3.5 vs Sonnet 4	1.07	.306	−2.61	.032*	4.38	< .001***
	你是		−0.88	.569	0.58	.570	13.16	< .001***
	我們／们		0.56	.586	−1.14	.418	14.70	< .001***
	你們／们		−0.94	.552	0.19	.856	21.12	< .001***
提示詞（實驗二）	我是	簡易 vs 精細	−0.04	.969	−2.61	.032*	6.05	< .001***
	你們／们		−0.21	.839	0.19	.856	4.08	< .001***

註：κ 比較經 Fisher's z 轉換後進行獨立樣本 t 檢定（df = 8）。不一致率比較採用雙比例 z 檢定。p 值已依實驗分別進行 Benjamini–Hochberg FDR 校正（實驗一，12 個檢定；實驗二，6 個檢定）。*p < .05, ***p < .001（校正後）。實驗一「我是」與實驗二「我是」的內部 κ 比較結果相同（t = −2.61, p = .032），係因兩實驗在該條件下共用 Sonnet 4 與精細提示詞的編碼數據。

資料來源：本研究統計整理。

三、論述分析應用示範

LACA 編碼完成後，本研究進一步分析 A 樣本（人稱代詞明確指涉身分詞）的分布特徵與論述模式。表 8 顯示了不同人稱代詞的身分指涉比例：

表 8：不同人稱代詞 KWIC 樣本的身分指涉分布

檢索詞 KWIC 樣本	具體身分指涉樣本	樣本總數	指涉占比
我是	415	781	53.14%
你是	177	793	22.32%
我們／们	183	980	18.67%
你們／们	162	1,013	15.99%
總數	937	3,567	26.27%

資料來源：本研究統計整理。

從表 8 可觀察到明顯的比例差異：「我是」結構的身分指涉比例最高（53.14%），「你們／们」比例最低（15.99%）。這種分布差異也對應於編碼任務的難易程度——如前所述，「我是」樣本的編碼一致性最高，「你們／们」樣本最低。

這個結果或可歸因於語言中指稱明確性（referential clarity）的差異，第一人稱「我」如「我是」，作為自我導向行動（self-directed action），具有特殊的固定指涉方式，能清楚標示說話者對某身分的主動宣告（Orbán, 2025）。相較於「我」，第二人稱複數的「你們／们」則比較依賴脈絡資訊判定其所指涉的他者群體。在脈絡窗口很小（左三右七）的 KWIC 條目中，這種依賴脈絡的特徵使其身分指涉更為模糊與多義。這種現象也和「自我定位」（self-positioning）與「他者定位」（other-positioning）（Davies & Harré, 1990/1999）的區分相呼應：自我宣告通常語言形式明確，較易識別；相較之下，對他者的分類更依賴脈絡，因此較難達成一致的判定。

為了示範 LACA 能為後續論述分析提供可靠基礎，本研究進一步採用紮根理論的聚焦編碼（focused coding, Charmaz, 2014）分析 937 條 A 樣本。所示範的分析結果，是透過持續比較法（constant comparative method）反覆檢視 A 樣本——包括檢視不限於左三右七窗口的原始脈絡——迭代地建構而得，主要目的在展示這批具有統計代表性的樣本的某部分論述模式，而非展示深入的論述分析。

以下呈現分析結果。A 樣本包含以下幾類論述模式：「自我身分宣告」、「他人身分指稱」、「諷刺或自嘲」、「反串或扮演」、「討論帳號身分」、「討論集體身分」。本研究採用非互斥分類法，每條樣本可同時歸入多個編碼軸（如「我是中國海南省韭菜」[8] 同時歸入「中華

[8] 中國網路用語「韭菜」用來比喻被一再剝削卻不反抗的人民，像韭菜一樣被割了一茬又長一茬。

核心」、「中華邊緣」認同與「自嘲身分」）。各類別的比例計算方式為：包含該軸特徵的樣本數除以總樣本數（937），因此各類別比例總和超過 100%。

其中，中華核心認同的自我身分宣告約27.00%（我是大陆人，我更是中国人[樣本 011]）；中華邊緣認同約 28.07%（我是台灣人不是中國人[樣本 268]），核心與邊緣之間認同約 6.30%（说的好！香港華人，台湾華人，我是新加坡華人[樣本 156]）；諷刺或自嘲身分約 20.81%（你們粉紅[樣本 844]、我是韭菜[樣本 247]）；反串或扮演身分約 8.75%（我是中國台灣省金門縣人[樣本 010][9]）；討論帳號身分（質疑或反質疑評論帳號背後的真實身分）約 16.86%（你是中國人嗎？？別反串啦[樣本 542]）；直述他者身分，約 10.78%（你是少數有獨立思考能力的中國人[樣本 560]）；討論集體身分（質疑、反質疑、假設性描述等）約 35.54%（他们会说你是中国人，你不能说国家不好[樣本 575]））。綜合而言，〈玻璃心〉語料庫呈現出華裔相關身分的評論者之間互不信任地討論何謂華裔身分。

進一步說，LACA 所建構出的統計代表樣本，不僅能協助檢驗既有理論假設，也具備支持探索性發現的潛力。以下舉例說明。

在自我身分宣告中，自嘲身分「椰子」再三出現：從個人自我身分宣稱的「我是中國海南省的椰子」（樣本 284），到集體身分的「我們椰子樹就紮根在海南省島」（樣本 596），再到批評他者的「海南省人長期給我們椰子樹同胞們澆核水」（樣本 649），恰巧構成連貫的敘事鏈。考慮到統計代表性，這些椰子論述在母體中應該有相當數量，才會在隨機樣本中重複出現；再透過 KWIC 追溯原始脈絡可確認，這三個樣

[9] 金門屬於中華民國福建省金門縣，而非中國臺灣省金門縣。會犯這種基本錯誤，表示評論者不是臺灣人的可能性很大。因此本文把該樣本編為「偽裝反串身分」。

本均來自同一帳號的評論。

這個發現的意義不僅在於找到一種新的自嘲隱喻，更在於它的發現方式。這既不是研究者針對預設關鍵詞「椰子」的檢索結果，也不是隨意瀏覽 KWIC 條目的偶然發現，而是在系統化編碼過程中歸納所得。這也是 LACA 相對於傳統 CADS 的另一個優勢所在：後者或者仰賴更高的共現值——在語料庫中與「我們」共現的詞彙中，「椰子」的排行是 228，雖然它的 z-score 4.997 並不低——或者憑隨意瀏覽語料的運氣找到「有趣」案例；而前者即使在運氣不特別好的情況下，仍有較高機率發現具統計代表性的論述模式。

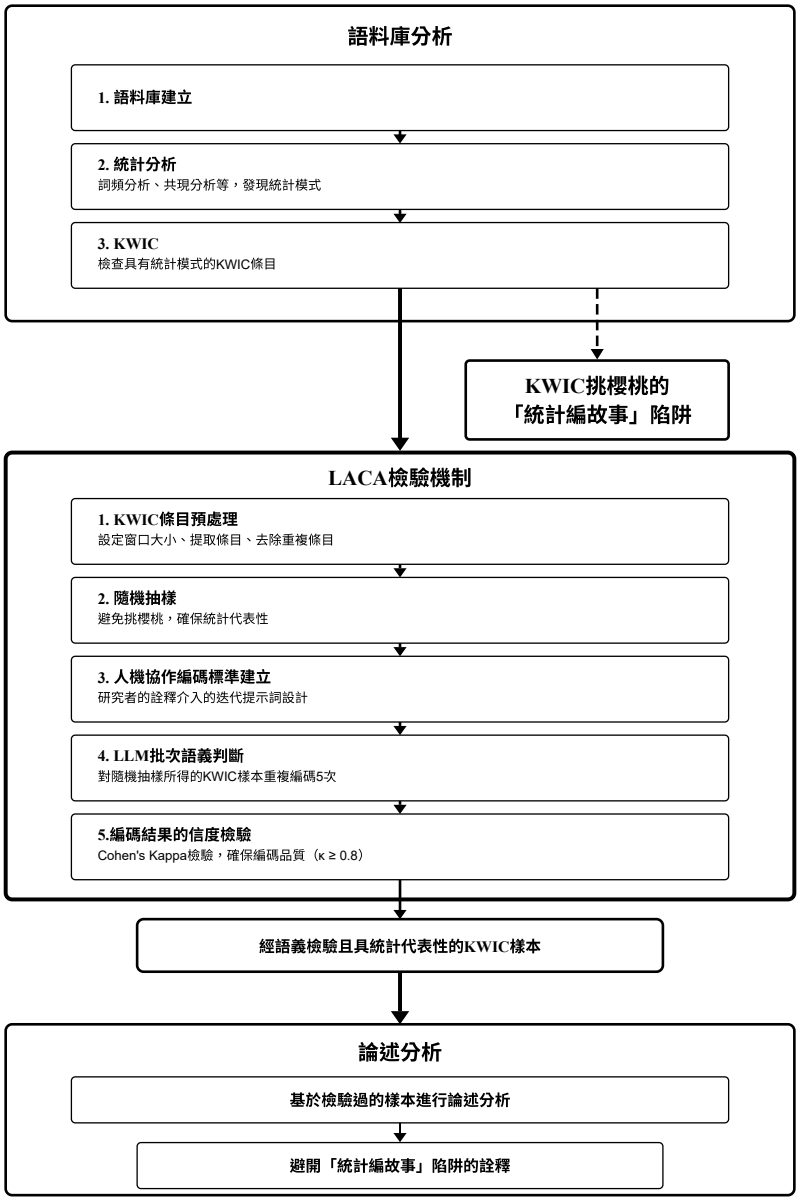
因此，LACA 不僅提供批次語義編碼的具體操作程序，更為 CADS「如何在語料庫中找出有統計意義的論述模式」的工作，提出擴充的方案。

四、小結

本研究示範了在 CADS 中，LACA 檢驗機制能否更有效整合語料庫分析與論述分析。實驗結果顯示，適當配置的 LLM（如 Sonnet 4 配合精細提示詞）能高信度地編碼。這意味著，基於 LACA 篩選過的樣本所得到的詮釋結果，如本研究示範的「〈玻璃心〉語料庫呈現出華裔相關身分的評論者之間互不信任地討論何謂華裔身分」，落入「統計編故事」的風險更低。進一步說，這個檢驗程序還有額外的方法論優勢：透過開放式歸納而非預設詞彙檢索或碰運氣的挑櫻桃，LACA 能更系統地識別研究者未預期的有趣論述模式。身分自嘲的新隱喻「我是椰子」的發現過程展現了這種優勢。

圖 1 呈現的工作流程，說明 LACA 如何作為 CADS 的中介檢驗機制，有效整合語料庫分析與論述分析。

圖 1：在 CADS 中的 LACA 檢驗機制



資料來源：本研究繪製。

陸、討論

以下討論本研究在驗證 LACA 技術可行性過程中湧現的方法論發現。如前所述，這些發現並非預設的研究目標，而是在實際驗證過程中逐步闡明的知識論意義，是概念驗證工作在方法論層次的焦點（Kendig, 2015）。

本研究的起始目的是，檢驗在特定條件下，LACA 能否讓 CADS 的語料庫分析與論述分析更有效地整合，降低「統計編故事」風險。

實驗結果指出，Sonnet 4 的編碼信度（Cohen's κ）最高可達 0.979，顯著優於 Haiku 3.5 搭配簡易提示詞的表現。這顯示在適當的模型選擇與提示詞設計下，LLM 能穩定模擬研究者的編碼判斷。

本研究進一步基於 937 個經過 LACA 語義檢驗且具統計代表性的樣本，歸結出「〈玻璃心〉語料庫呈現出華裔相關身分的評論者之間互不信任地討論何謂華裔身分」的詮釋結果，示範 LACA 如何能有效降低「統計編故事」的風險。

本研究進一步指出，LACA 的價值不僅限於「低成本批次語義檢查工具」。如果只是追求基礎分類，以目前 LLM 的能力（如 Sonnet 4），使用簡易提示詞即可。而透過進一步的提示詞設計——這個工作需要研究者投入更多心力理解與詮釋語料——LACA 還能發揮「系統化樣本篩選工具」的功能，將「透過 KWIC 挑選樣本以檢視理論假設」的論述分析過程，從原本的「挑櫻桃」轉變為「系統化識別」。

從成本效益面向來看，以本研究 3,567 條樣本的二元編碼為例，傳統方法需要建立完整的標準編碼本、訓練編碼員、執行及品管編碼流程，約 650-700 工作小時的人力投入。相較之下，LACA 只需要研究者

自行投入約 2 週進行提示詞優化，並支付少許服務費，即可完成同等規模的任務，大幅節省研究成本。

以下進一步討論這些研究發現的方法論意義。

一、模型選擇與臨床驅動的提示詞設計

實驗結果顯示，LLM 在處理不同檢索詞 KWIC 樣本的表現差異顯著。「我是」樣本最簡單、「你們／们」樣本最難達到高信度。這反映了不同難度的語義判斷：自我身分宣告語義明確，要正確辨識第二人稱複數的指涉則需要更精細的脈絡檢查。

這種差異進一步凸顯了模型選擇（Haiku 3.5 與 Sonnet 4）與提示詞設計（簡易與精細）對編碼品質的顯著影響。以「你們／们」樣本為例：同樣使用精細提示詞，Haiku 3.5 的 κ 值只有 0.416，Sonnet 4 則可達到 $\kappa = 0.930$；另一方面，使用簡易提示詞的 Sonnet 4 降到 $\kappa = 0.869$。

這些數據揭示了兩個方法論條件：LLM 性能門檻與研究者知識的轉化。某些複雜的編碼任務需要特定水準以上的 LLM 才能勝任，也需要研究者能將自身原本內隱的知識轉化為適當提示詞。作為研究者可主動操作因素的後者尤其重要。比較簡易與精細兩種提示詞（見附錄一）可觀察到多處設計差異：研究者在精細版中展現了更複雜的判斷原則，如語法的否定結構、不同言說行動類型的判斷等。

然而研究者其實無法完全確定，真正影響編碼效果是哪些提示詞設計差異。在這裡提供兩個故事加以說明。故事一：本研究在 2024 下半年啟動時還只有 Sonnet 3.5。當時以為，本研究的主要貢獻會是各種提示詞設計，因此研究初期不斷測試「某種提示詞會對編碼有效」的各種假設，甚至到了走火入魔的地步——例如本研究曾懷疑使用繁體或簡體中文的提示詞，會影響 LLM 做出「臺獨或反臺獨傾向」的編碼結果。

然而在 2025 年 5 月 Sonnet 4 推出後，先前這些試出來的「咒語」往往失效。例如，原本效果極差的否定型（不要……非……）提示詞，效果大為提高。故事二：在做完實驗二後，本研究另外做了一個小實驗，嘗試將精細提示詞調整得在邏輯上更一致、「更精細」。出乎意料的是，這個貌似更完善的提示詞版本反而降低編碼效果。這個發現揭示了 LACA 的某種「黑盒子」屬性：提示詞並非越精細越好，有時額外細節反而造成干擾。

這些發現使本研究轉向新的解決方案：「臨床驅動」（clinically-driven）的提示詞設計。借鑒古希臘醫學經驗學派（empirical school）的智慧與現代臨床醫學的測試方法，即使無法完全理解運作機制的內在邏輯，研究者仍可透過系統化實驗來找出有效的治療法。本研究主張 LACA 也應採取這樣的臨床驅動取徑。即使無法完全理解為何精細提示詞優於簡易提示詞、又為何「更精細」反而表現不佳，研究者仍可透過系統化檢查 LLM 編碼出錯之處來調整提示詞，並透過信度檢測評估效果，直到找出最適用的版本。

這種「建構→檢驗→優化」的迭代過程，可以成為 LACA 的標準程序。其具體步驟包括：（1）以任務的語義複雜度為基準，選擇適當的模型並建構初始提示詞版本；（2）執行編碼，聚焦於 LLM 編碼錯誤的樣本；（3）判斷錯誤類型，僅針對特殊錯誤調整提示詞[10]；（4）每次修改僅變動一項判斷條件，確保可以追蹤每次變動的來源；（5）檢視

[10] LLM 編碼出錯的基本情況有二：（1）特殊錯誤：對特定類型案例的編碼結果與研究者預期不符，這通常反映提示詞不夠精確或存在邏輯漏洞。這時需要修改提示詞來澄清編碼邏輯；（2）系統錯誤：即使在黑盒子裡可能有某種系統原因所造成，但對使用者而言與隨機出現無異的錯誤。這在語料模稜兩可的情況常見，在這種情況下 LLM 傾向隨機決定編碼結果。對於這類受限於 LLM 能力的錯誤，研究者只能接受這個技術限制，只要錯誤率在容錯範圍內（$\kappa \geq 0.8$）就好。

輸出結果。如果覺得正確率已經提高到某種程度，進行信度檢驗，並記錄每次修改的提示詞版本與該版本得到的 κ 值；（6）反覆迭代至 κ 值達標（≥ 0.8）為止，並留意過度精細化可能導致信度下降的情況。需要強調的是，由於 LLM 的黑盒子屬性、模型的快速更新，以及對研究者知識轉化能力的依賴，這些步驟構成的是可複製的「操作流程」，但其執行效果依賴研究者的操作技藝。因此，提示詞設計無法標準化為跨任務的普遍原則。LACA 之所以能夠複製，在於透明記錄決策邏輯與檢驗過程，而非套用原則。簡而言之，相較於隨著模型更新而可能失效的具體提示詞設計原則，「臨床驅動」的後設原則才真正能夠一概適用。

或許讀者會有疑問：除了「判斷編碼錯誤類型，僅針對特殊錯誤調整提示詞」之外，如果沒有普遍適用的提示詞修改原則，研究者又如何能優化提示詞？確實，本研究在迭代提示詞設計的過程中，也觀察與分析了 LLM 的編碼錯誤，尋找出錯的可能「原因」，但這只是為了提高眼前的編碼任務效果，不是企圖找出可普遍適用的設計原則。在 LLM 快速演化的環境中，過度依賴設計原則反而可能導致失敗，正如前述故事顯示的那樣，原本的提示詞設計原則隨著模型的進步而失效，或者使用邏輯上更精細的提示詞，反而導致信度下降。

本研究的這個主張與 Dunivin（2025）在操作層次上一致，但在方法論層次上有所差異。該研究的 LACA（該研究稱為質化編碼〔qualitative coding〕）同樣透過迭代方式設計提示詞，但程序上是先以傳統雙人黃金標準制訂編碼本，再轉換為 LLM 提示詞，並主張研究者可透過檢視編碼本定義中對 LLM 而言不清楚之處來修改提示詞。本研究則依循 Morgan（2007）主張的實用主義立場，不依賴 LLM 對研究者編碼原則的理解能力，而是以信度（κ 值）作為迭代的依據。值得注意的是，Dunivin（2025）雖在方法論上主張要先有黃金標準編碼本，但

也承認，轉換過的提示詞終究必須接受臨床檢驗，這實際上支持了本研究主張的「臨床驅動」原則。

二、LACA 作為詮釋型資訊工具的方法論意義

上一小節說明，提示詞設計涉及「研究者知識的轉化」。進一步說，不僅提示詞設計，而是一整個 LACA 的完整程序，包括 KWIC 樣本選取、抽樣設計、LLM 選擇到提示詞設計與迭代，每個環節都體現了研究者的理論判斷與詮釋，並將這些詮釋「嵌入」可批次執行的程序中。這讓本研究設計的 LACA 程序可稱作「詮釋型資訊工具」（劉慧雯、柯籐晏，2016）。以下討論這種嵌入的方法論意義。

（一）厚數據分析：從小樣本厚描到批次詮釋

過往對於大數據的方法論批判，如 Crawford（2013, April 2）與 Parks（2014）建議以小數據研究（small data studies）深入特定脈絡，補足大數據分析的侷限。另一方面，劉慧雯、柯籐晏（2016 年 6 月 17-19 日）的「詮釋型資訊工具」方法論原則，則強調在使用批次編碼工具的規模化分析中保留厚描（thick description, Geertz, 1973）的詮釋厚度。

本研究的實驗設計正是基於「詮釋型資訊工具」原則。CADS 研究者原本就會檢視 KWIC 樣本來尋找潛在論述模式，然後根據假設的模式尋找樣本。本研究則是將這套詮釋學循環（hermeneutic circle, Gadamer, 1960/1975）過程，轉化為交由 LLM 執行的提示詞迭代設計，既保持了研究者對語料的詮釋判斷，又實現了這種判斷的批次執行。例如，本研究在檢視語料時發現非標準語序如「作為臺灣人，我們……」的句型容易被 LLM 誤判為無身分表達，便將這類句型的識別規則納入提示詞（見附錄一）。進一步說，此舉不僅實現詮釋的批次化，也為後續論述

分析提供更有效樣本。

（二）詮釋型資訊工具的嵌入

作為能使 CADS 的語料庫分析與論述分析更有效整合的工具，LACA 具有雙重功能：既有規模化處理資料的批次能力，又能透過提示詞的設計，識別值得深入分析的論述特徵。劉慧雯、柯籙晏（2016 年 6 月 17-19 日）原本主張，詮釋型資訊工具的概念核心在於研究者的早期介入。本研究延伸這個觀點指出，LACA 可以被視作為一個多階段「嵌入」（embedding），而非僅是早期「介入」的詮釋型資訊工具：

1. 理論驅動的語料選取：

在語料預處理階段，選擇「我是、你是、我們／们、你們／们」作為檢索 KWIC 的起點，將研究者的理論判斷轉化為可操作的編碼原則。這個選擇基於自我宣告（self-identification）在身分認同理論中的核心地位。相較於其他可能的詞彙，「我是、你是、我們／们、你們／们」是認同協商論述最有效的指標。

2. 方法論考量的抽樣策略：

在抽樣階段，從 34,880 條語料中基於通用的抽樣原則（信心水準 95%，抽樣誤差 3%）抽樣 3,567 條樣本，以隨機抽樣取代挑櫻桃，確保篩選出的樣本具有統計代表性。這個樣本數能夠在資料代表性與人機協作可行性之間取得平衡，在避免挑櫻桃的同時，保有研究者的詮釋空間。

3. 臨床驅動的迭代提示詞設計：

在建立標準編碼階段，透過臨床驅動的迭代提示詞設計與編碼效果

的系統檢驗，將研究者對語料的詮釋，轉化為可批次執行的標準編碼，是整個 LACA 程序最核心的詮釋嵌入點。這個工作是研究者持續進行詮釋判斷、將內隱知識外顯化並加以檢驗，實際上已經是論述分析的前置階段。

三、LACA 的混合方法論品質評估

若以 Creswell & Clark（2017）所提出混合方法研究的六個評估面向，包括推論品質（inference quality）、一致性（consistency）、整合有效性（integration validity/Effectiveness）、理解的擴充（extended understanding）、信度與效度（reliability and validity），以及可行性與實用價值（feasibility and practical value），來評估本研究提出的 LACA 檢驗機制。除了一致性面向，即資料處理的過程與結果與理論的對應程度，因本研究的概念驗證性質而不適用之外，在其他五個面向上表現良好。

在推論品質面向，產生了基於具統計代表性樣本的論述分析結果，說了一個「華裔相關身分的評論者之間諷刺又不信任地討論何謂華裔身分」的故事；在整合有效性面向，建立了可操作的整合程序，能降低 CADS 中量化與質化分析往往未能整合，甚至落入「統計編故事」陷阱的情況；在理解的擴充面向，批次編碼 KWIC 樣本的 LACA 克服了人力難以全面檢驗的限制，使研究能更系統地識別未預期的論述模式；在信度與效度面向，人機編碼的高信度（最佳配置下，κ 值全數 > 0.85）確保後續論述分析的有效性；在可行性與實用價值面向，相較於傳統內容分析，能夠節省大量時間成本。

柒、結論

本研究檢驗了 LACA 用於橋接 CADS 的語料庫分析與論述分析，避開「統計編故事」陷阱的可行性。實驗結果顯示，適當配置的 LLM（如 Claude Sonnet 4 搭配精細提示詞）能達到極高編碼信度（$\kappa = 0.979$），即使處理較複雜的語料，透過迭代提示詞設計，仍能將信度從不可接受提升至優秀水準。

研究結果發現，本研究提出的 LACA 檢驗機制，除了發揮研究原本預期的、大幅降低內容分析成本的「批次語義檢查工具」功能之外，就 CADS 後續的論述分析階段而言，還能發揮「系統化樣本篩選工具」的功能。基於 937 個經過語義檢驗且具統計代表性的樣本，本研究示範地說了一個避開「統計編故事」陷阱的故事。此外，「我是椰子」的案例更展現了 LACA 能協助研究者，以非預設詞彙搜索的方式識別具統計代表性的論述模式。

在方法論貢獻方面，本研究提出的 LACA 檢驗機制為 CADS 提供了可操作的混合方法論整合，以 Creswell & Clark（2017）提出的標準加以評估，在推論品質、整合有效性、理解擴充、信度效度及可行性等面向均表現良好。

本研究採取了幾項概念驗證階段的方法設計選擇。為確保詮釋邏輯與提示詞迭代的緊密銜接，本研究採取單一研究者參與人機協作標準的建立。這是一個「最小可行配置」（minimum viable configuration）的選擇，未來在實際應用時，則可依需求擴充參與者類型，例如納入內部領域專家、外部獨立檢驗者，語言使用者代表等，以涵蓋更多元的詮釋面向。同時，本研究選擇語義關係相對直接的「人稱代詞+身分詞」論述

模式作為檢驗示範，為概念驗證提供了簡單的起點，LACA 在處理更複雜語用現象的應用潛力仍待進一步探索。此外，本研究以 YouTube 評論為實驗對象，研究設計也對應於這類語料的非規範特徵，例如以 KWIC 條目為分析單位。當未來應用於規範化語料如新聞報導時，研究設計需要作出相應的調整，例如以標題、段落或全篇為分析單位。

基於 LACA 檢驗程序的擴充潛力，未來研究可探索 LACA 在不同理論架構與多元語料類型下的應用方式，並建立應用案例資料庫，適時檢驗效果，必要時修改品質判準，以逐步提升這套程序的完整程度與適用範圍。在探索這些可能的應用時，為確認 LACA 是否適用，本研究建議先以兩條判準進行評估：（1）最終 Cohen's κ 是否達到 0.80 以上？（2）相較初始條件（如使用簡易提示詞），κ 值是否顯著提升？若兩者都不成立，則可判定該任務在目前條件下不適合使用這套程序。

進一步說，本研究提出的 LACA 程序，雖然專為檢驗 CADS 而設計，但其中的方法論原則，尤其是「臨床驅動」原則，應可適用於其他內容分析任務，甚至可能用於多模態 LACA——值得注意的是，隨著 LLM，如新版 GPT-4o、Gemini、Claude 等，已具備文字、語音、圖像、影像等多模態辨識能力，多模態 LACA 的技術條件已經具備。當然，與單純的文字語料分析相較，多模態 LACA 的具體程序設計，必然有很大不同。

最後要說明的是，站在論述分析的立場，本研究並不主張統計門檻是詮釋的最終判準。語義錯亂與判斷失序本身，或許是語料中藏有未被發現的論述策略的指標。即便在信度檢驗未達標的情況下，也不應將這類現象單純視為雜訊，反而可作為詮釋與厚描的起點。論述分析當然不需要以 LACA 甚至 CADS 這類工具作為前提。但本研究設計的檢驗程序，能夠在不取代研究者判斷的前提下，為詮釋提供更一致、可檢驗的

語料基礎，讓後續分析能站在更穩固的起點上。

應將 LACA 視作一種品管機制，除此之外別無其他。而最後的結果，到底是詮釋洞見還是編故事，讓時間決定。

參考書目

郭文平（2020）。〈語料庫輔助的媒體論述分析：以台灣平面媒體中國夢報導為語料的實證研究〉，《資訊社會研究》，38，51-92。

劉慧雯、柯籐晏（2016 年 6 月 17-19 日）。〈邁向厚數據：以「詮釋型資訊工具」進行意義分析的概念基礎〉【論文發表】。「中華傳播學會 2016 年會」，嘉義縣，臺灣。

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337-351. https://doi.org/10.1017/pan.2023.2

Baker, P. (2006). *Using corpora in discourse analysis*. Bloomsbury Academic.

Charmaz, K. (2014). *Constructing grounded theory*. Sage.

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

Chubb, L. A. (2023). Me and the machines: Possibilities and pitfalls of using artificial intelligence for qualitative data analysis. *International Journal of Qualitative Methods*, *22*, 16094069231193593. https://doi.org/10.1177/16094069231193593

Chueh, H.-C., & Chen, K.-H. (2016). CORPRO: A Chinese language corpus tool and a case study of media representation of organic agriculture. In J. Xiang (Ed.), *Digital humanities: Between past, present and future* (pp. 285-313). National Taiwan University Press.

Crawford, K. (2013, April 2). The hidden biases in big data. *Harvard Business Review*. https://hbr.org/2013/04/the-hidden-biases-in-big-data

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research*. Sage.

Davies, B., & Harré, R. (1990/1999). Positioning and personhood. In R. Harré & L. v. Langenhove (Eds.), *Positioning theory: Moral contexts of intentional action* (pp. 32-52). Blackwell.

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

Elliott, S. (2021). Proof of concept research. *Philosophy of Science*, *88*(2), 258-280.

Friginal, E., & Hardy, J. A. (2020). Corpus approaches to discourse analysis: Introduction and section overviews. In E. Friginal & J. A. Hardy (Eds.), *The Routledge Handbook of Corpus Approaches to Discourse Analysis* (pp. 1-4). Routledge.

Gabrielatos, C., & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996-2005. *Journal of English Linguistics*, *36*(1), 5-38. https://doi.org/10.1177/0075424207311247

Gadamer, H. G. (1960/1975). *Truth and method*. Seabury Press.

Geertz, C. (1973). Thick Description: Towards an interpretive theory of culture. In *The Interpretation of cultures* (pp. 3-30). Basic Books.

Gries, S. T. (2016). *Quantitative corpus linguistics with R : A practical introduction*. Taylor and Francis.

Hayes, A. S. (2025). "Conversing" with qualitative data: Enhancing qualitative research through large language models (LLMs). *International Journal of Qualitative Methods*, *24*, 16094069251322346. https://doi.org/10.1177/16094069251322346

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Addison-Wesley Publishing Company.

Kendig, C. E. (2015). What is proof of concept research and how does it generate epistemic and ethical categories for future scientific practice? *Sci Eng Ethics*, *22*(3), 735-753. https://doi.org/10.1007/s11948-015-9654-0

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (Fourth Edition ed.). Sage.

Morgan, D. L. (2007). Paradigms lost and pragmatism regained: Methodological implications of combining qualitative and quantitative methods. *Journal of Mixed Methods Research*, *1*(1), 48-76. https://doi.org/10.1177/2345678906292462

Orbán, K. (2025). Self-referring as self-directed action. *Philosophical Studies*, *182*(2), 567-588. https://doi.org/10.1007/s11098-025-02283-2

Parks, M. R. (2014). Big data in communication research: Its contents and discontents. *Journal of Communication*, *64*(2), 355-360.

Vine, B. (2020). Spoken workplace discourse. In E. Friginal & J. A. Hardy (Eds.), *The Routledge handbook of corpus approaches to discourse analysis* (pp. 5-21). Routledge.

Williams, E. A. E. (2020). Critical discourse analysis for language policy and planning. In E. Friginal & J. A. Hardy (Eds.), *The Routledge handbook of corpus approaches to discourse analysis* (pp. 481-498). Routledge.

Yu, D., Li, L., Su, H., & Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apologies. *International Journal of Corpus Linguistics*. https://doi.org/10.1075/ijcl.23087.yu

Zhang, W., Deng, Y., Liu, B., Pan, S., & Bing, L. (2024, June). Sentiment analysis in the era of large language models: A reality check. In K. Duh, H. Gomez, & S. Bethard, *Findings of the association for computational linguistics: NAACL 2024* Mexico City, Mexico.

附錄一

附錄中各提示詞設計皆為本研究實驗所用版本；括號中所列 κ 值，為該提示詞在相應模型下的信度評估結果。

一、用於「我是」KWIC 樣本編碼的簡易提示詞（Sonnet4, κ = 0.924）

編碼原則：

研究背景：分析《玻璃心》（2021）YouTube 評論中的身份敘事，聚焦自我身份表達。以「我是」為關鍵詞的左三右七 kwic 集合而成的語料庫

A（自我身份表達）：

- 包含「我是」，後直接接身份標記（政治實體、地理身份、族群身份、政治身份、集體標籤），語義上明確宣告自我身份。
- 條件：
  - 「我是」後為名詞或名詞短語（如「我是台灣人」「我是小粉紅」）。
  - 排除動詞中介（如「我是祝福 XX」）。
- 範例：
  - 我是台灣人（地理身份）
  - 我是華人（族群身份）
  - 我是韓粉（政治身份）
  - 我是小粉紅（集體標籤，可能自嘲或認同）

B（非自我身份表達）：

- 包括：

  - 無身份標記（如「我是杀人犯」）。
  - 含身份標記但未以「我是」明確表達自我身份（如「五毛很討厭」「我是祝福中國人民」）。
  - 標籤他人身份（如「你是五毛」「他是台獨分子」）。
  - 含集體標籤但無明確主詞指涉（如「五毛拿好」）。
- 範例：
  - 我是祝福中國人民（行動表達）
  - 你是五毛（標籤他人）
  - 五毛拿好（無明確主詞）
  - 我是哪儿的？（無身份標記）

身份標記定義：
- 政治實體：台灣、中國、中共等（含縮寫：台、中、CCP）。
- 地理身份：大陸、香港、台北人、中國台灣省等。
- 族群身份：華人、漢人、客家人等。
- 政治身份：統派、獨派、韓粉等（正式立場或政黨認同）。
- 集體標籤：五毛、小粉紅、韭菜、塔綠班、台蛙、1450、牆內人等（帶諷刺或貶義，僅在「我是」明確自我宣告時計入 A 類）。
- 變體：縮寫（台、中）、複合形式（台澎金馬）、修飾性複合詞（真正的中國人、緬甸勞工、正港台灣人）、行動動詞（辱華、反共，僅限如「我是反共人士」）。

根據以下格式編碼以下語料：流水號順序 tab 理由 tab 編碼結果（小寫）

二、用於「我是」KWIC 樣本編碼的精細提示詞（Sonnet4, $\kappa = 0.979$）

編碼原則：

研究背景：分析《玻璃心》（2021）YouTube 評論中的身份敘事，聚焦自我身份表達。以「我是」為關鍵詞的左三右七 kwic 集合而成的語料庫

A 類（肯定或探討「我」的身份）：

語句涉及「我是」與顯性身份標記（政治實體、地理身份、族群身份、政治身份、集體標籤）的確認或探討，聚焦「我」的身份定義，涵蓋：

「我是」+ 顯性身份標記（如「我是台灣人」）。

「我是」+ 否定詞（「不是」「不喜歡」「不接受」「不認同」等）+ 顯性身份標記，描述「我是」的身份（如「我是不是小粉紅」「我是不喜歡被說是小粉紅」）。

他人表示「我」是顯性身份標記（如「他們說我是境外勢力」）。

他人表示「我」不是顯性身份標記（如「他們說我不是台灣人」）。

他人質疑「我」的顯性身分標記（如「他們說我是不是台灣人？」）

關鍵：必須包含顯性身份標記（如「台灣人」「小粉紅」），且聚焦「我」的身份確認或探討（不要求認同）。

B 類（不涉及肯定或探討「我」的身份）：

語句完全無顯性身份標記，或身份標記僅作為「我」的行動的對象，如「我是客觀討論，中國會滅亡」「我是討厭共產黨」。

明確規範：若無顯性身份標記，一開始即歸為 B 類，避免推斷隱含身份（如「語言」暗示「華人」，「立場」暗示「統派」）。

身份標記定義

政治實體：

範例：美國、台灣、中國、中共等（含縮寫：US、台、中、ccp）。反例：國家、政府、祖國

地理身份：

範例：歐美、大陸、香港、台北人、中國台灣省、馬來籍、越南籍、牆內、牆國等。反例：國外、海外、地區、地方、山上、島內。

族群身份：

範例：華人、漢人、客家人、少民等。反例：民族、同胞。

政治身份：共產黨、KMT、民進黨、統派、獨派、韓粉、馬列子孫等（正式立場或政黨認同）

集體標籤：

範例：五毛、小粉紅、韭菜、塔綠班、台蛙、1450、牆內人、中華膠、反賊、玻璃心、境外勢力等。反例：老百姓、普通人、底層、群眾、人民。

- 變體：縮寫（台、中）、複合形式（台澎金馬、北上廣深）、修飾性複合詞（真正的中國人、緬甸勞工、正港台灣人）、行動動詞（辱華、反共，僅限如「你們反共人士」）。

根據以下格式編碼以下語料：流水號順序 tab 理由 tab 編碼結果（小寫），範例：001 無顯性身份標記 b、003 含顯性身份標記「馬列同路人」，他人表示「我是」該身份 a、987「我」肯定自我身分「小粉紅」a

三、用於「你是」KWIC 樣本編碼的精細提示詞（Sonnet4, $\kappa = 0.883$）

編碼原則：

研究背景：分析《玻璃心》（2021）YouTube 評論中的身份敘事，聚焦自我身份表達。以「你是」為關鍵詞的左三右七 kwic 集合而成的語料庫

A 類（肯定或探討「你」的身份）：

語句涉及「我是」與顯性身份標記（政治實體、地理身份、族群身份、政治身份、集體標籤）的確認或探討，聚焦「你」的身份定義，涵蓋：

「你是」+ 顯性身份標記（如「你是台灣人」）。

「你是」+ 否定詞（「不是」「不喜歡」「不認同」等）+ 顯性身份標記，描述「你是」的身份（如「你是不是小粉紅」「你是不喜歡被說小粉紅」「你是不認同中國人」）。

他人表示「你」是顯性身份標記（如「他們說你是境外勢力」）。

他人表示「你」不是顯性身份標記（如「他們說你不是台灣人」）。

他人質疑「你」的顯性身分標記（如「他們說你是不是台灣人？」）

關鍵：必須包含顯性身份標記（如「台灣人」「小粉紅」），且聚焦「你」的身份確認或探討（不要求認同）。

B 類（不涉及肯定或探討「你」的身份）：

語句完全無顯性身份標記，或身份標記僅作為「你是」的行動的對象，如「你是認為中國會滅亡？」「你是不是支持台灣？」「你是被中共洗腦」。

明確規範：若無顯性身份標記，一開始即歸為 B 類，避免推斷隱含身份（如「語言」暗示「華人」，「立場」暗示「統

派」）。

身份標記定義

政治實體：

範例：美國、台灣、中國、中共等（含縮寫：US、台、中、ccp）。反例：國家、政府、祖國。

地理身份：

範例：歐美、大陸、香港、台北人、中國台灣省、馬來籍、越南籍、牆內、牆國等。反例：國外、海外、地區、地方、山上、島內。

族群身份：

範例：華人、漢人、客家人、少民等。反例：民族、同胞。

政治身份：共產黨、KMT、民進黨、統派、獨派、韓粉、馬列子孫等（正式立場或政黨認同）。

集體標籤：

範例：五毛、小粉紅、韭菜、塔綠班、台蛙、1450、牆內人、中華膠、反賊、玻璃心、境外勢力等。反例：老百姓、普通人、底層、群眾、人民。

- 變體：縮寫（台、中）、複合形式（台澎金馬、北上廣深）、修飾性複合詞（真正的中國人、緬甸勞工、正港台灣人）、行動動詞（辱華、反共，僅限如「你們反共人士」）。

根據以下格式編碼以下語料：流水號順序 tab 理由 tab 編碼結果（小寫），範例：001 無顯性身份標記 b、003 含顯性身份標記「馬列同路人」，他人表示「你是」該身份 a、987 肯定「你是」顯性身分「小粉紅」a、561「你是」的行動的對象是地理身份「海南人」b

四、用於「我們/们」KWIC 樣本編碼的精細提示詞（Sonnet4, κ = 0.915）

研究背景：分析《玻璃心》（2021）YouTube 評論中的身份敘事，聚焦身份表達。以「我們」為關鍵詞的左三右七 kwic 集合而成的語料庫。

A 類（「我們」肯定或探討自我身份）：

語句涉及「我們」與顯性身份標記（政治實體、地理身份、族群身份、政治身份、集體標籤）的確認或探討，聚焦「我們」的身份定義，涵蓋：

「我們」+ 顯性身份標記（案例「我們台灣人」）。

「我們」+ 否定詞（「不是」「不喜歡」「不接受」「不認同」等）+ 顯性身份標記，描述「我們」的身份（範例「我們不是中國人」「我們不喜歡被說是中國人」）。

他人表示「我們」是顯性身份標記（如「他們說我們是境外勢力」）。

他人表示「我們」不是顯性身份標記（如「你們不要說我們是中國人」）。

非標準語序（如「作為台灣人的我們」）等同於「我們」+ 顯性身份標記，歸為 A 類。

關鍵：必須包含顯性身份標記（如「台灣人」「小粉紅」），且聚焦「我們」的身份肯定或探討（不要求認同）。

B 類（不涉及肯定或探討自我身份）：

語句完全無顯性身份標記，或身份標記僅作為「我們」的行動的對象，範例「我們統一台灣」。

明確規範：若無顯性身份標記，一開始即歸為 B 類，避免推斷隱含身份（如「語言」暗示「華人」，「立場」暗示「統派」）。

身份標記定義

政治實體：

範例：美國、台灣、中國、中共等（含縮寫：US、台、中、ccp）。反例：國家、政府、祖國。

地理身份：

範例：歐美、大陸、香港、台北人、中國台灣省、馬來籍、越南籍、牆內、牆國等。反例：國外、海外、地區、地方、山上、島內。

族群身份：

範例：華人、漢人、客家人、少民等。反例：民族、同胞。

政治身份：共產黨、KMT、民進黨、統派、獨派、韓粉、馬列子孫等（正式立場或政黨認同）。

集體標籤：

範例：五毛、小粉紅、韭菜、塔綠班、台蛙、1450、牆內人、中華膠、反賊、玻璃心、境外勢力等。反例：老百姓、普通人、底層、群眾、人民。

- 變體：縮寫（台、中）、複合形式（台澎金馬、北上廣深）、修飾性複合詞（真正的中國人、緬甸勞工、正港台灣人）、行動動詞（辱華、反共，僅限如「你們反共人士」）。

根據以下格式編碼以下語料：流水號順序 tab 理由 tab 編碼結果（小寫），如：980 包含地理身份「大陸人」，肯定「我們」的身份 a、111 無顯性身份標記 b、151「我們」的行動的對象是地理身份「海

南人」b

五、用於「你們/们」KWIC 樣本編碼的簡易提示詞（Sonnet4, κ = 0.705）

編碼原則：

研究背景：分析《玻璃心》（2021）YouTube 評論中的集體身份標籤，聚焦第二人稱的身份指涉。以「你們」為關鍵詞的左三右七 kwic 集合而成的語料庫

A（指涉對方身分）：

- 包含「你們」，後直接接身份標記（政治實體、地理身份、族群身份、政治身份、集體標籤），語義上明確指涉對方身份。
- 條件：
  - 「你們」後為名詞或名詞短語（如「你們台灣人」「你們小粉紅」）。
  - 排除動詞中介（如「你們支持 XXX」）。
- 範例：
  - 你們中國人（直接標記）
  - 你們台灣人（直接標記）
  - 你們這些五毛（集體標籤）
  - 你們華人（族群身份）

B（不指涉對方身分）：

- 包括：
  - 「你們」無特定身份指涉（如「你們看看」）。
  - 含身份標記定義但有動詞中介（如「你們支持台灣人」）。
  - 無身份標記（如「你們杀人犯」）。

  - 含身份標記但未以「你們」明確指涉對方身分（如「五毛很討厭」）。
  - 含集體標籤但無明確主詞指涉（如「五毛拿好」）。
- 範例：
  - 你們都看過嗎（無身份指涉）
  - 你們欺負台灣人（動詞中介）
  - 你們支持中國人（動詞中介）

身份標記定義：
- 政治實體：台灣、中國、中共等（含縮寫：台、中、CCP）。
- 地理身份：大陸、香港、台北人、中國台灣省等。
- 族群身份：華人、漢人、客家人等。
- 政治身份：統派、獨派、韓粉等（正式立場或政黨認同）。
- 集體標籤：五毛、小粉紅、韭菜、塔綠班、台蛙、1450、牆內人等（帶諷刺或貶義，僅在「你們」明確指涉對方時計入 A 類）。
- 變體：縮寫（台、中）、複合形式（台澎金馬）、修飾性複合詞（真正的中國人、緬甸勞工、正港台灣人）、行動動詞（辱華、反共，僅限如「你們反共人士」）。

根據以下格式編碼以下語料：流水號順序 tab 理由 tab 編碼結果（小寫）

六、用於「你們/们」KWIC 樣本編碼的精細提示詞（Sonnet4, $\kappa = 0.869$）

研究背景：分析《玻璃心》（2021）YouTube 評論中的身份敘事，聚焦身份表達。以「你們」為關鍵詞的左三右七 kwic 集合而成的語料庫。

A 類（肯定或探討「你們」的身份）：

語句涉及「你們」與顯性身份標記（政治實體、地理身份、族群身份、政治身份、集體標籤）的確認或探討，聚焦「你們」的集體身份，涵蓋：

「你們」+ 顯性身份標記（案例：「你們中國人」「中國人,你們」「中國人是你們」、反例：去你們新加坡）。

「你們」+ 否定詞（「不是」「不認為」等）+ 顯性身份標記，否定「你們」的身份（如「你們不是中國人」「你們不認為自己是中國人」）

他人表示「你們」是顯性身份標記（如「誰叫你們共產黨」）。

他人表示「你們」不是顯性身份標記（如「我看你們不是中國人」）。

他人質疑「你們」的顯性身分標記（如「你們說我們不是中國人？」）

關鍵：必須包含顯性身份標記（如「台灣人」「小粉紅」），且聚焦「你們」的身份確認或探討。

B 類（不涉及肯定或探討「你們」的身份）：

語句完全無顯性身份標記（「你們深入不到那種地方」），或「你們」的行動的對象是身份標記（案例「你們要武統台灣」反例「你們中國要武統」）。

明確規範：若無顯性身份標記，一開始即歸為 B 類，避免推斷隱含身份（如「語言」暗示「華人」，「立場」暗示「統派」）。

身份標記定義

政治實體：

範例：美國、台灣、中國、中共等（含縮寫：US、台、中、ccp）。反例：國家、政府、祖國。

地理身份：

範例：歐美、大陸、香港、台北人、中國台灣省、馬來籍、越南籍、牆內、牆國等。反例：國外、海外、地區、地方、山上、島內。

族群身份：

範例：華人、漢人、客家人、少民等。反例：民族、同胞。

政治身份：共產黨、KMT、民進黨、統派、獨派、韓粉、馬列子孫等（正式立場或政黨認同）。

集體標籤：

範例：五毛、小粉紅、韭菜、塔綠班、台蛙、1450、牆內人、中華膠、反賊、玻璃心、境外勢力等。反例：老百姓、普通人、底層、群眾、人民。

- 變體：縮寫（台、中）、複合形式（台澎金馬、北上廣深）、修飾性複合詞（真正的中國人、緬甸勞工、正港台灣人）、行動動詞（辱華、反共，僅限如「你們反共人士」）。

根據以下格式編碼以下語料：流水號順序 tab 理由 tab 編碼結果（小寫），如：980 否定「你們」的身份「大陸人」a、111 無顯性身份標記 b、151「你們」的行動的對象是地理身份「海南人」b

附錄二

本研究開發了兩個互補的 LACA 輔助工具，分別用於提示詞迭代設計與編碼結果的統計比較。建議先以編碼品質檢驗工具進行臨床驅動的提示詞迭代設計，再以統計比較分析工具比較不同提示詞條件的編碼信度差異。兩個工具均要求 Python 3.7 以上版本，需安裝 numpy、scipy、pandas、sklearn 等套件，完整的工具套件已開源發布於 GitHub：https://github.com/duress/llm-coding-reliability-analyzer

1. 編碼品質檢驗工具

編碼品質檢驗工具是一個專為 LLM 提示詞迭代設計的工具，便於研究者即時檢視編碼品質並識別問題樣本。該工具接受制表符（Tab）分隔的純文字檔案輸入，每行包含六個編碼值（LLM 五次編碼加一個人類標準編碼），計算 LLM 與標準編碼間信度和 LLM 重複編碼內部的 Cohen's κ 值。

本工具用於計算多項信度指標，包括 LLM-人類不一致率（任何 LLM 編碼與人類標準編碼不一致的樣本比例）、LLM 內部不一致案例（LLM 各次編碼間存在分歧的樣本）以及各編碼者的平均 κ 值。輸出詳細的 TXT 信度報告，包含完整的 κ 值統計數據、不一致樣本明細（顯示樣本流水號、各編碼結果和最終編碼），以及 CSV 結果檔案（含最終編碼、原始編碼結果和一致性分析）。

2. 統計比較分析工具

統計比較分析工具是一個專為檢驗不同條件下 LLM 編碼差異而設計的綜合性工具。該工具整合了完整的統計分析流程，包括 Cohen's κ

計算、Bootstrap 重採樣標準誤估算、Fisher's z 轉換、獨立樣本 t 檢驗、z 檢驗（用於比較不一致率），以及 Benjamini-Hochberg FDR 多重比較校正等統計方法。

程式支援兩個編碼資料集的比較分析，自動計算 LLM 與標準編碼間信度、LLM 重複編碼內部信度，並進行不一致率的統計比較。輸出詳細的 TXT 統計報告，涵蓋兩個條件的 κ 描述性統計（平均值、標準差、各次編碼的個別 κ 值）、統計比較結果（t 值、z 值）、原始與校正後 p 值，以及顯著性標記。

Bridging Statistics and Interpretation: An LLM-assisted Content Analysis Approach to Corpus-assisted Discourse Analysis

Lu-Yen Ko*

ABSTRACT

Corpus-assisted Discourse Studies (CADS) face the methodological pitfall of statistical storytelling. Researchers often use keywords in context (KWIC) to purposively select samples matching statistical patterns from corpus analysis and then conduct discourse analysis based on these samples without systematically verifying their statistical representativeness. While content analysis offers a mature solution to this methodological pitfall, its high cost renders it practically infeasible for any CADS study.

This research proposes a Large Language Model-assisted Content Analysis (LACA) validation mechanism to integrate corpus analysis and content analysis in CADS, rendering previously "theoretically necessary but practically infeasible" semantic verification operationally viable, thereby avoiding the pitfall of "statistical storytelling".

Research Questions

As a proof-of-concept study, this research examines the proposed LACA validation mechanism under a minimum viable configuration, using YouTube

* Lu-Yen Ko holds a Ph.D. from the College of Communication at National Chengchi University. e-mail: duress.ko@gmail.com. ORCID: 0009-0004-6065-7749.

comments on the popular song “Fragile”[11] as the corpus and addressing four questions.

1. Can LACA effectively identify semantic relationships between co-occurring words in the corpus?
2. How do different LLM (Large Language Model) models and prompts affect LACA’s coding performance?
3. What is the consistency between LLM coding and researcher standards?
4. Can the proposed LACA mechanism effectively bridge statistical patterns and discourse interpretation in CADS, avoiding the pitfall of “statistical storytelling”?

Research Methods

First, the study obtains statistically representative KWIC samples through systematic sampling, using personal pronouns 我是/I am, 你是/you are, 我們/们/we, and 你們/们/you [plural] as search terms and establishing a reliable foundation for semantic verification.

Second, it systematically develops a Standard Coded Set through human-machine collaborative iterative prompt construction and refinement. This hermeneutic circle of construct → verification (κ) → refinement involves iteratively examining LLM coding results and refining prompts to improve coding standards’ logic and clarity, until consistency stabilizes at Cohen’s $\kappa \geq 0.8$. This ensures coding judgment principles possess clarity and operability.

Third, the study conducts experiments using the established standard

[11] “Fragile” (玻璃心) is a Mandarin pop music video released on YouTube in 2021 by Malaysian-Chinese singer Namewee (黃明志) and Australian-Chinese singer Kimberley Chen (陳芳語). The song satirizes radical Chinese nationalist netizens known as “little pinks” (小粉紅) and rapidly accumulated over 30 million views and 200,000 comments on YouTube, making it a significant corpus for studying online identity discourse.

coding as an evaluation benchmark, comparing the coding effectiveness and consistency of different LLM configurations (Haiku 3.5 vs. Sonnet 4) and prompt types (simple vs. refined). These experiments verify LACA's feasibility as a bridging mechanism for CADS.

The coding task distinguishes samples where personal pronouns reference specific identities as A (e.g., "臺灣人，你們讓人喜歡" / "Taiwanese people, you are likable") from those that do not. The task's key challenge lies in distinguishing mere lexical collocation from actual semantic reference. Coding standards must identify not only samples lacking identity word collocation (e.g., "你們讓人喜歡" / "you are likable") as B, but also false positives where pronouns collocate with identity words without referring to them (e.g., "你們喜歡臺灣人" / "you like Taiwanese people") as B. Coding reliability validates whether LACA can effectively handle such judgments, ensuring Category A samples' semantic validity.

Research Findings

The findings reveal that model and prompt configurations significantly impact LACA's coding performance. When both models use refined prompts to code KWIC samples from four search terms (我是，你是，我們/们，你們/们), Sonnet 4 significantly outperforms Haiku 3.5. Sonnet 4 achieves nearly perfect consistency across all tasks ($\kappa = 0.869$-0.979). In contrast, Haiku 3.5's performance declines with corpus complexity - for the most ambiguous "你們/们" samples, reliability drops to $\kappa = 0.380$, or below content analysis standards.

For prompt comparison, when Sonnet 4 codes identical KWIC samples, refined prompts significantly outperform simple prompts. For simpler "我是" samples, both prompts achieve excellent consistency, but refined prompts further improve reliability (from $\kappa = 0.924$ to $\kappa = 0.979$). Prompt effects are

more pronounced on complex corpora: for the most challenging "你們/们" samples, refined prompts elevate consistency from acceptable levels (κ = 0.705) to excellent levels (κ = 0.869). Results demonstrate that selecting appropriate model and prompt configurations is critical to ensuring LACA's effectiveness.

Beyond its expected function as a batch semantic verification tool, LACA also serves as a systematic filtering tool, assisting researchers in discovering meaningful discourse patterns from semantically-validated and statistically-representative samples. For instance, this research identifies a novel self-identity metaphor from LACA-verified samples: "I am a coconut."

The research further demonstrates how LACA-verified samples enable identifying discourse patterns in the corpus - specifically, "pervasive irony and distrust toward commenters' self-declarations" - effectively avoiding the pitfall of "statistical storytelling".

Discussion

Building on these findings, the study examines LACA's methodological significance. Results show that its effectiveness depends on two factors: LLM performance thresholds and researchers' ability to transform domain expertise into executable prompts. However, prompt engineering faces a black-box challenge: specific design principles become obsolete as models evolve, and logically more refined prompts may even reduce coding reliability.

To address this challenge, the study proposes the Clinical-Driven principle of prompt engineering, advocating systematic iterative prompt refinement with empirical effectiveness as the optimization standard. This meta-principle ensures LACA's continued applicability as LLMs and prompt strategies evolve. Reproducibility depends on transparently documenting decision logic and verification processes and not on replicating specific prompt principles.

LACA embodies the methodological significance of an interpretive information tool. From theory-driven search term selection and methodologically-informed sampling design to clinically-driven prompt engineering, researchers' theoretical judgments and interpretations are embedded into the CADS process through LACA's mediation at multiple stages, essentially realizing the batch implementation of thick description.

LACA provides a concrete operational framework for integrating quantitative and qualitative approaches. Across five dimensions, LACA performs strongly. In inference quality, it produces discourse analysis results based on statistically- representative samples. In integration effectiveness, it establishes operational integration procedures, reducing CADS's frequent failure to integrate quantitative and qualitative results. In expanding understanding, it enables researchers to systematically identify unanticipated discourse patterns in corpora. High human-machine coding reliability (under optimal configuration, all κ values > 0.85) ensures subsequent discourse analysis validity. In feasibility and practical value, compared to traditional content analysis, it achieves approximately significant reductions in both cost and time.

Research Limitations and Future Directions

This research adopts a proof-of-concept minimum viable configuration. Future applications can expand as needed. The single-researcher design can involve multiple researchers. Selecting the straightforward "personal pronoun + identity word" pattern, LACA's potential for more complex pragmatic phenomena awaits exploration. Future research can explore LACA across different theoretical frameworks and corpus types. Leveraging LLMs' multimodal capabilities, the Clinical-Driven principle can serve as the meta-guide for LACA's methodological expansion towards multimodal applications.

Keywords: Human-AI collaboration, Large Language Model-assisted content analysis, Mixed methods research, Interpretive information tools, Proof of concept, Corpus-assisted discourse studies