自動化內容分析與AI幻覺解析

AI的教學與研究應用

汪志堅

國立臺北大學

大綱

- AI自動化內容分析的建議步驟
- AI自動化內容分析運用於教學
- AI幻覺與AI虛假文獻

自動化內容分析

讓內容分析更加強大，成為具體可行的研究策略

內容分析 (Content Analysis)

一種客觀、系統化且能量化的研究方法，主要用於分析文字、圖像、語音、影像等各種內容。

將大量看似雜亂無章的內容，透過標準化的分類與編碼，轉化為可以進行統計分析的數據，藉此找出潛在的趨勢、特徵或社會意涵。

內容分析的重要缺點

極度耗費時間與人力成本

自動化內容分析的優點

- **成本與效率**：降低時間與人力成本，提升效率
- **提高客觀性**：避免標註員的主觀、疲勞、偏見
- **發掘潛在主題**：找出難以察覺的潛在關聯與深層主題
- **促進一般化能力**：大量資料，超越特定樣本或個別情境

自動化內容分析的常見應用領域

政治學與公共政策：分析各種政治文本與輿情

傳播學與媒體研究：分析各種新聞與閱聽人意見

行銷與消費者研究：分析消費者口碑及與品牌互動

生態學與環境科學：分析生態與環境監控資料

公衛、工安、教育：分析網路資訊早期預警公共衛生危機、
分析傷害報告以分析安全意外
分析教育現場的互動與學習行為

自動化內容分析的技術沿革

- 字典與詞彙法 (dictionary-based and lexicon-based Approaches)
- 監督式機器學習法 (Supervised Machine Learning Approaches)
- 非監督式機器學習與主題模型法 (Unsupervised Machine Learning and Topic Modeling Approaches)
- 大型語言模型與生成式 AI 法 (Large Language Models and Generative AI Approaches)

字典與詞彙法 (dictionary-based and lexicon-based Approaches)

- Linguishic Inquiry and Word Count (LIWC)是代表性的方法之一
- NTUSD（National Taiwan University Sentiment Dictionary，台灣大學情感詞典）是由台大自然語言處理實驗室（NTU NLP Lab）開發的著名中文情感分析詞典。
- **字典與詞彙法可執行** 語意分析（Sentiment Analysis）

監督式機器學習法（Supervised Machine Learning Approaches)

將資料區分為訓練集，進行訓練，再將訓練結果運用於剩餘的資料。

使用包括：單純貝氏Naive Bayes、支持向量機SVM、邏輯回歸、類神經網路等演算法則，進行訓練。

利用Precision, Accuracy, Recall, F1-Score等指標，驗證模型的正確性。

非監督式機器學習與主題模型法 (Unsupervised Machine Learning and Topic Modeling Approaches)

代表性方法為:狄利雷克分配模型 Latent Dirichlet Allocation LDA

將龐大的非結構化文本，自動分群為多個潛在的主題模塊

大型語言模型與生成式 AI 法 (Large Language Models and Generative AI Approaches)

讓大型語言模型扮演類似於真人標註員的角色

透過角色設定的方式，加上內容分析的分類定義，以及單樣本或少樣本的方式，即可讓大型語言模型扮演標註員的角色。

以往的自動化 內容分析具有進入障礙，非資訊背景的研究者，難以進行自動化 內容分析

沒有方便可用的方式
執行自動化 內容分析

非資訊背景的研究者，
有技術障礙，只能使用字典法

因此，
LIWC成為最常用的自動化 內容分析的方法之一

有了大型語言模型 LLM

自動化內容分析變得具體可行

不用技術能力，也能執行自動化內容分析

一次嘗試:
利用LLM進行內容分析

Yang, Y., Wang, C. C., Chou, P. L., Tien, C. Y., Wen, J. C., & Chen, N. H. (2026). Evaluating Human-Based and Large Language Model-Based Content Analysis: A Case Study on Negative Word-of-Mouth Classification in Social Media. Computers in Human Behavior, 108904.

https://doi.org/10.1016/j.chb.2026.108904

當數據規模遇上語意深度：現代內容分析的兩難

規模的危機

純AI自動化的盲區：能以毫秒為單位處理海量數據，卻在多層次情緒（如反諷、杯葛暗示）前頻頻誤判。

細節的迷失

人工標註的極限：準確捕捉語氣與文化脈絡，但耗時、成本高昂，且面臨主觀疲勞（如對中立或反諷情緒的標註一致性極低）。

核心挑戰：如何在不犧牲語意精確度的前提下，實現數據分析的規模化？

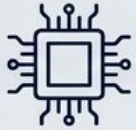
輿情分析的兩難：人類的精準度 vs. 網路的規模化

人類專家

優勢：深刻理解語境、諷刺與文化隱喻。

痛點：耗時數週、成本高昂、且具主觀疲勞偏差。

15%　　處理速度：極慢



純 AI 自動化 (Zero-Shot)

優勢：處理上萬筆資料僅需數分鐘（API 串接）。

痛點：無法辨識深層情緒（如：隱晦抵抵制、反諷），容易誤判。

100%　　精準度盲區

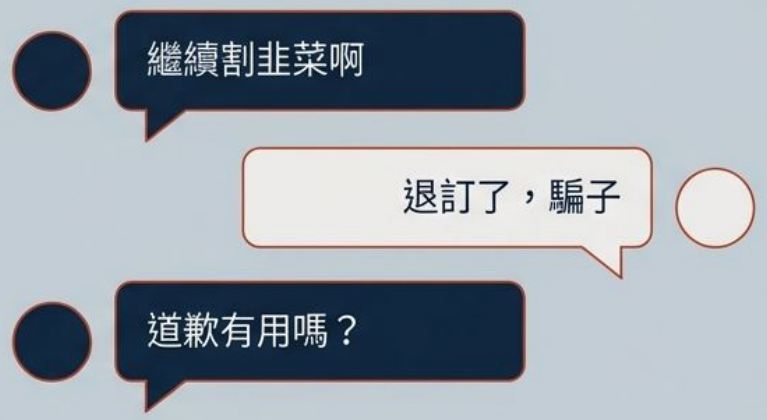
結論：傳統作法與純 AI 自動化皆無法單獨應對現代公關危機。我們需要第三條路。

危機爆發：當百萬粉網紅遭遇「假奶」風暴

THE DATA PANEL

- **事件**：2024年7月10日，知名YouTuber的副業「十盛奶茶」遭消基會點名「拒絕公開乳源」。
- **衝擊**：危機爆發後一個月內，湧入近萬則觀眾留言。

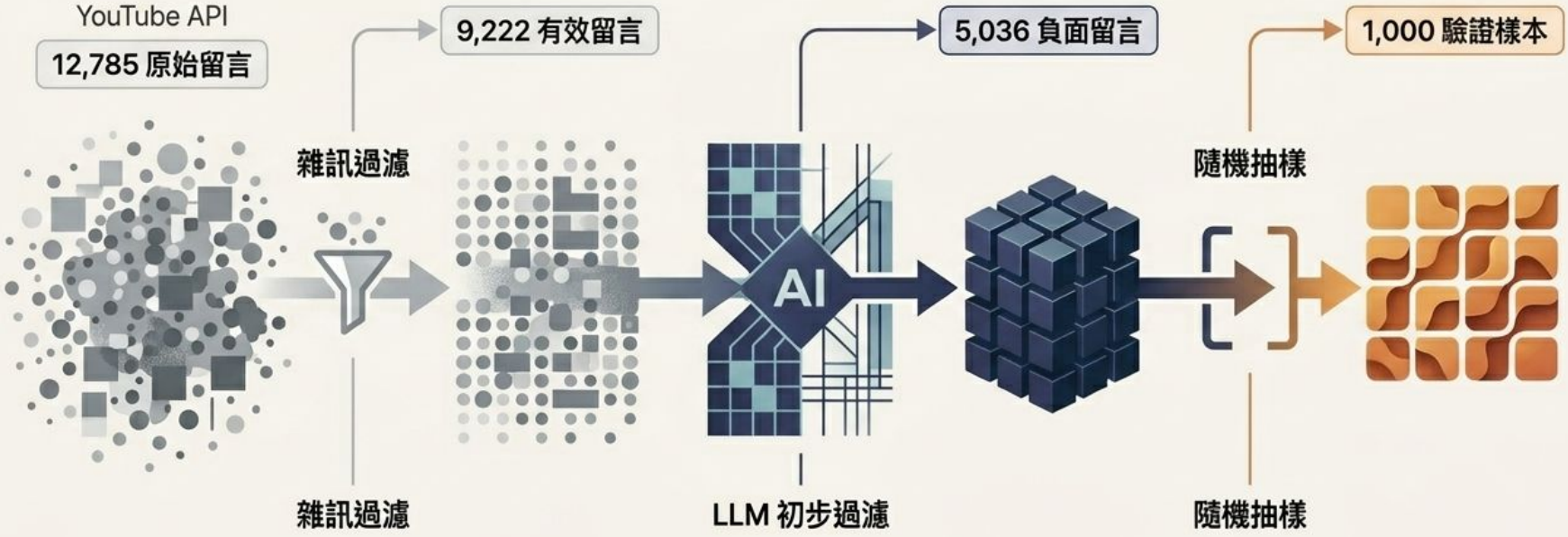
THE REALITY PANEL



核心挑戰：品牌必須在數小時內精準判讀這近萬則留言的真實意圖，以決定公關對策。

實戰場景：解析百萬網紅的公關危機

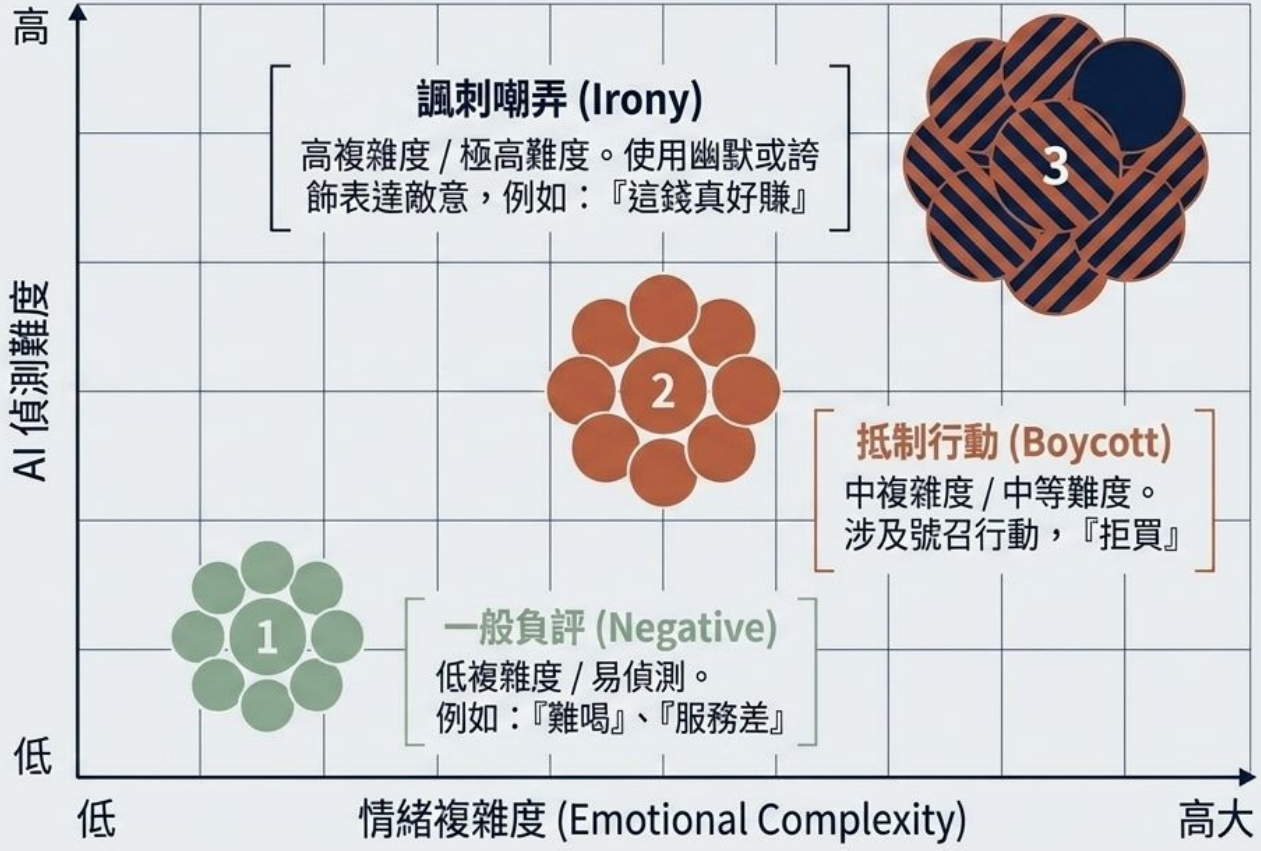
背景：知名網紅旗下奶茶品牌爆發「標示不實」爭議，引發大量社群負面口碑（NWOM）。
目標：在龐大數據中，精準篩選出「純粹抱怨」、「反諷 (Irony)」與「實質杯葛（Boycott）」的留言，以利品牌進行危機管理。



自動化分析管線的基礎運作機制

透過大型語言模型（LLMs）的自然語言處理能力，將原本需要數週的質性編碼工作，壓縮至數小時內完成。但但這套「零樣本（Zero-Shot）」系統的表現究竟如何？

拆解負面口碑（NWOM）：從單純抱怨到品牌危機



管理洞察：

品牌傷害最深的通常是「抵制」與「諷刺」，
這也是未經訓練的 AI 最容易漏判的盲區。

壓力測試 1：未經訓練的 AI 真實表現 (Zero-Shot Baseline)

一般負評 (Negative)

AI 能輕易抓出直接的負面情緒。

抵制行動 (Boycott)

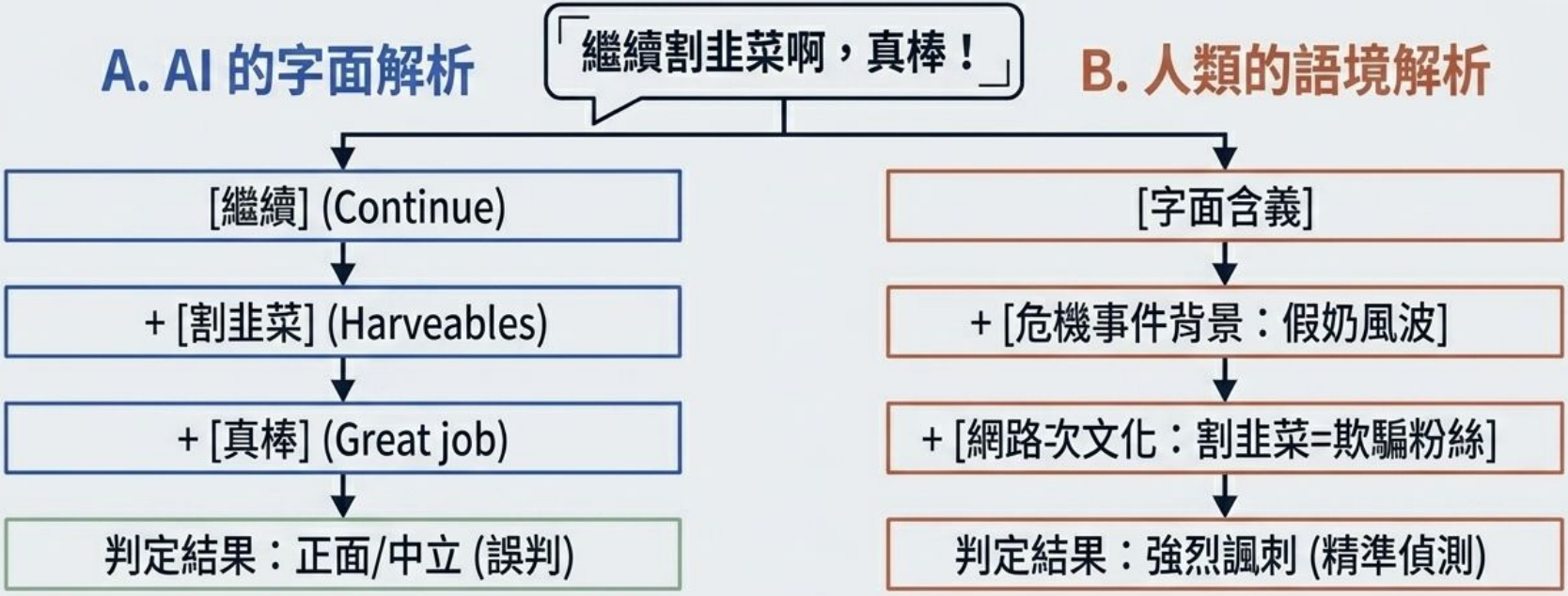
AI 容易將非抵制言論誤判為抵制。

諷刺嘲弄 (Irony)

AI 幾乎無法準確辨識反諷，
F1 Score 僅 0.391。

診斷洞察：在沒有提供定義與語境的情況下，ChatGPT 就像一個極度敏感但缺乏常識的雷達——抓得到雜訊，但分不清敵我。

為什麼 AI 聽不懂諷刺？(The Irony Blindspot)



核心瓶頸：NLP（自然語言處理）受限於表面語意。若缺乏「專屬定義」與「次文化字典」，強烈的公關危機將在 AI 眼中變成一片祥和。

診斷矩陣：人類 vs. AI 的基準能力對決

評估維度	人類專家	ChatGPT (基準)
一般負面情緒 (General Negativity)		 (高 Recall，表現優異)
抵制意圖辨識 (Boycott Detection)		 (Precision 極低，易錯殺)
諷刺語境辨識 (Irony Detection)		 (完全失效)
規模化與速度 (Speed & Scalability)		 (壓倒性優勢)

戰略結論：我們不能單獨依賴 AI 的精準度，也不能受限於人類的速度。必須將人類的「語境理解」注入 AI。

壓力測試 2：注入語境與規則（The Context Intervention）

[System Prompt Update...]

Layer 1: 注入精確定義 (Definitions)
抵制 (Boycott) = 明示或暗示退訂、拒買、拒絕支持。
諷刺 (Irony) = 使用與字面相反的語言，包含幽默、誇飾來表達批評。

↓

Layer 2: 注入危機專屬字典 (Keywords)
抵制關鍵字 = [退訂，拒喝，滾，避雷]
諷刺關鍵字 = [割韭菜，賺爛了，真會演，假奶，智商稅]

實驗洞察：單靠定義不夠。當「精確定義」結合「專屬危機關鍵字」時，才能真正解鎖 AI 的語意辨識能力。

語境的力量：關鍵字如何解鎖 AI 的精準度

抵制行動 (Boycott)

+26.6%

Recall 從 0.714 飆升至 0.980 (幾無漏網之魚)

F1 Score 提升至 0.619

諷刺嘲弄 (Irony)

Recall 提升至 0.856

顯著提升對反諷的敏感度

整體負評 (Negative)

F1 Score 達 0.690

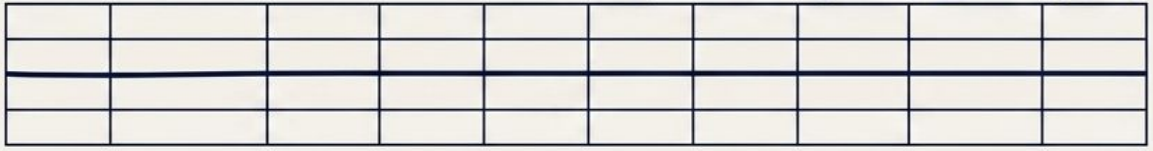
精準度全面強化

實驗證明：AI 不缺運算力，缺的是「人類提供的引導鷹架 (Scaffolding)」。
加入關鍵字字典是提升抵制與諷刺辨識的最有效策略。

壓力測試 3：參數微調的迷思 (The Illusion of Parameter Tuning)

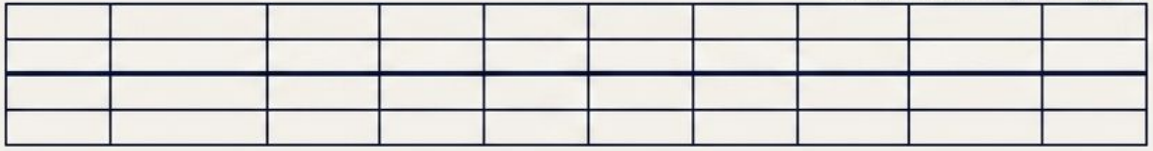
Temperature (0.0 to 2.0)

Optimal for **Negative** (0.5) & **Irony** (0.0)



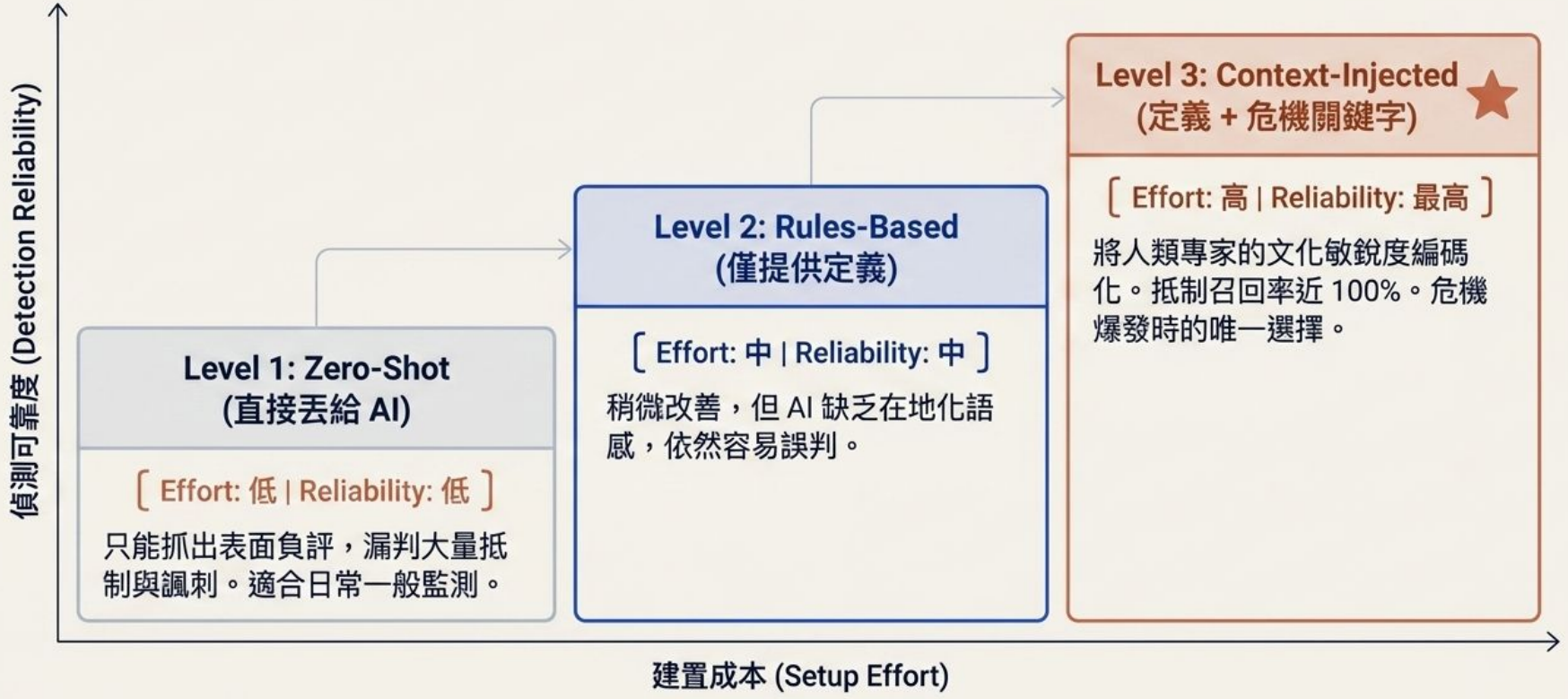
Top_p (0.0 to 1.0)

Optimal for **Boycott** (1.0)

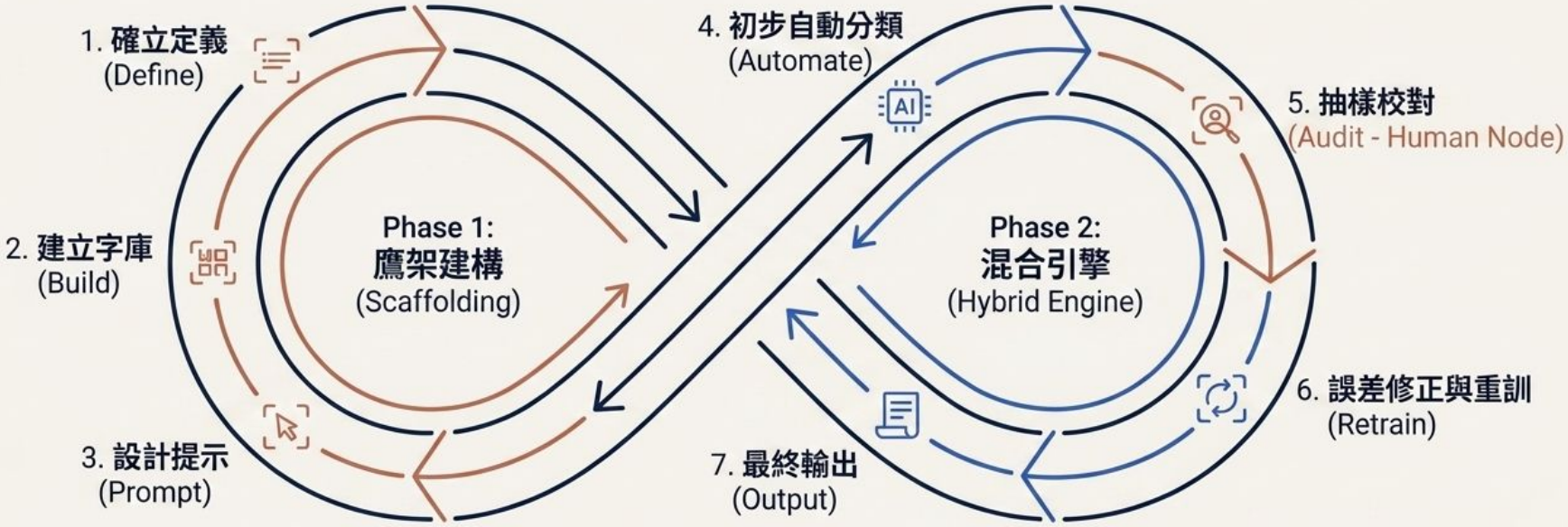


迷思	真相	結論
調教 Temperature 或 Top_p 能大幅改變分析品質？	實驗證實，參數微調的影響微乎其微。真正決定 AI 表現的是「Prompt 的結構設計」，而非系統參數。	**提示詞工程 (Prompting)** 遠勝於參數微調 (Tuning)。

AI 提示詞策略成熟度模型 (Prompting Maturity Model)



終極解方：7 步「人機協作」飛輪框架



核心觀念：部署 AI 不是「設定後放任不管」，而是一個持續演進的混合引擎。
人類提供智慧，AI 提供算力。

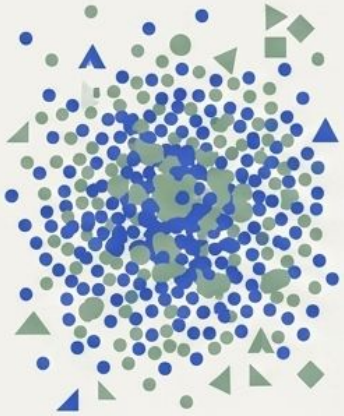
協作階段一：建構分析鷹架 (Phase 1: Scaffolding)

關鍵法則：垃圾進，垃圾出。AI 的分析天花板，取決於人類第一階段給予的語境深度。

協作階段二：驅動混合引擎 (Phase 2: The Hybrid Engine)

Step 6: 修正與重訓

將新發現的「諷刺梗」加回字庫，調整 Prompt 形成閉環。



[萬筆原始留言]

Step 4: LLM 規模化初篩

瞬間標籤化，釋放 90% 勞動力。

Step 5: 人類抽樣校對

專家針對最易錯的「諷刺」與「抵制」進行覆核。

Step 7: 輸出決策

產出高信效度儀表板，支援即時公關。

品牌公關的未來：算力與洞察的完美融合

1. 速度即生命，但準確是底線

在危機黃金數小時內，LLM 是唯一能處理海量負面口碑的工具。但未經引導的自動化將導致致命誤判。

2. AI 無法取代「文化嗅覺」

純自動化會死於語境盲區。諷刺與抵制的辨識，永遠需要人類專家介入設計「提示詞鷹架」。

3. 擁抱「人機協作飛輪」

將公關團隊的職能，從耗時的「手動標註員」徹底升級為**高價值的「AI 系統架構師與校對員」**。

未來的危機管理，贏家不是擁有最強大 AI 的品牌，
而是最懂得將「人類語境」編碼進入 AI 引擎的團隊。

基準測試：人類主觀性 vs. 機器剛性

	人類 (Human)	ChatGPT (Zero-Shot)
處理速度	極慢	極快
擴展性	受限於預算與人力	無限擴展
反諷識別 (Irony)	高 (F1: 0.918)	極低 (F1: 0.391)
杯葛識別 (Boycott)	高 (F1: 0.848)	中低 (F1: 0.618)

ChatGPT 在捕捉廣泛負面情緒時擁有極高召回率（Recall），但對於隱含意圖的精確度（Precision）嚴重不足。它需要人類的「導航」。

「反諷鴻溝」：為什麼 AI 會誤判？

機器的視角 (Literal Meaning)

AI 判定（Zero-Shot）：看到「真好笑」，可能誤判為中立或正面。

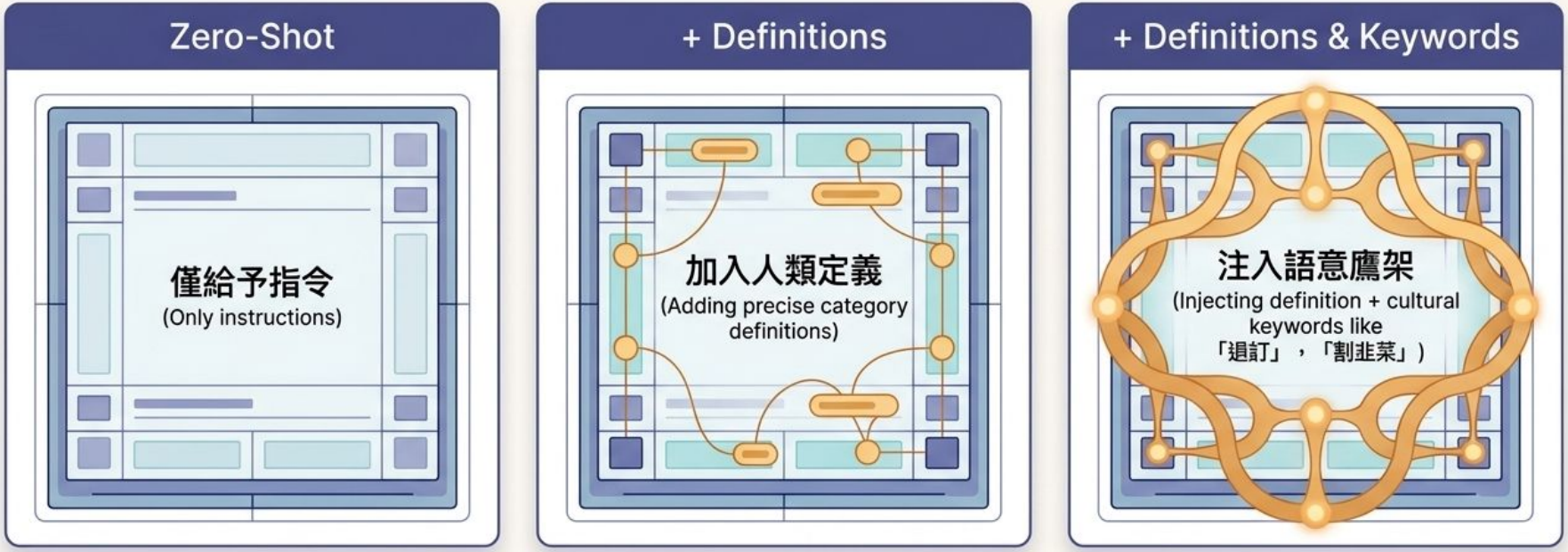
130萬粉絲都沒退追，真好笑，割韭菜囉。

人類的解讀 (Implicit Meaning)

人類判定：「割韭菜」是高度反諷與批評的文化象徵（Irony）。

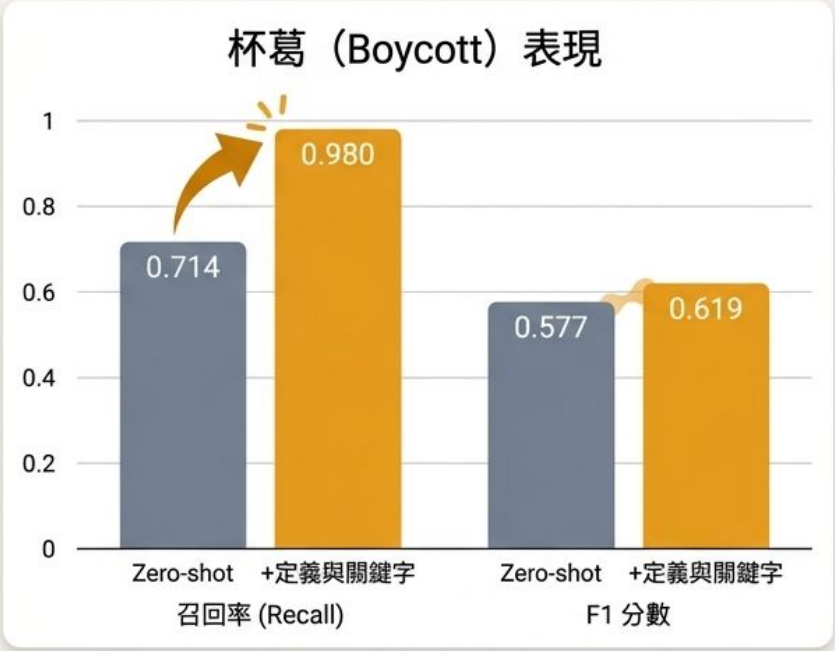
缺乏當地文化語境與事件背景，大型語言模型無法穿透字面意義理解多層次情緒。

搭建語意鷹架：人類知識的精準注入



實驗證明，單純提供「分類定義」對 AI 的幫助有限。唯有將人類專家從語料中萃取的「特徵關鍵字」與定義結合，才能真正引導 AI 跨越語意鴻溝。

協作的威力：關鍵字如何喚醒 AI 的語境感知



人類關鍵字（如「退訂」、「拒買」）的介入，讓 ChatGPT 偵測杯葛意圖的召回率飆升至 98%，幾乎捕捉了所有關鍵危機訊號。

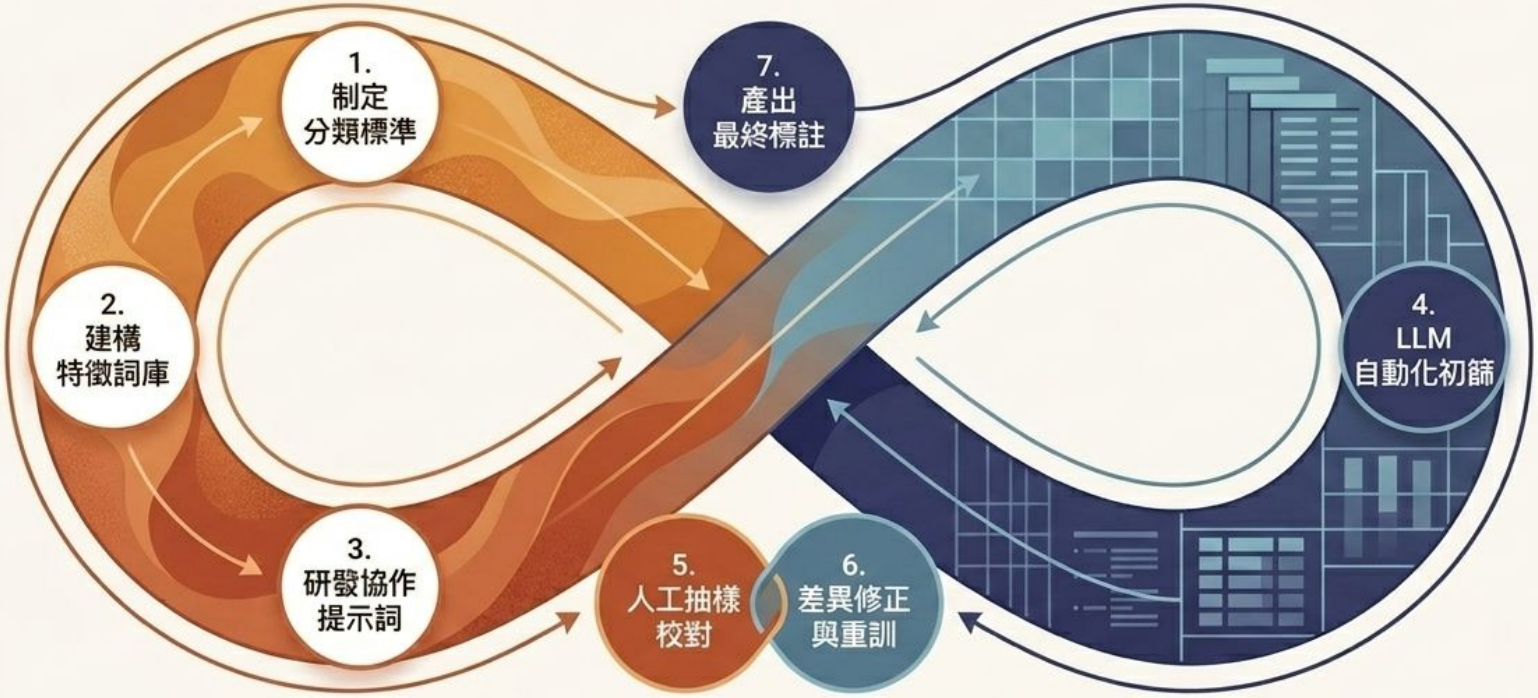
微調機器的極限：參數是輔助，提示才是燃料

設定為 0.5 時負面分類表現最佳，設定為 0 時反諷識別最穩定。

設定為 1 能夠在杯葛分類中取得完美的平衡。

核心洞察：調整這些機器參數（Temperature、Top_p）僅能帶來微幅的邊際效益。真正決定準確度上限的，是人類提供的「提示詞品質」與「語意鷹架」。

人機協作內容分析框架 (Human-AI Collaborative Framework)



一個非線性、可迭代的黃金標準流程，完美結合人類的主觀洞察與 AI 的運算規模。

內容分析的新典範 (The New Paradigm)

傳統思維：AI 取代人類

新典範：人機協作

1. AI 是引擎，人類是導航：不要指望大語言模型獨立完成一切。機器提供規模，人類提供文化羅盤。

2. 語意鷹架不可或缺：將專家定義與在地文化關鍵字注入提示詞，是突破 AI 語意盲區的最強武器。

3. 建立動態迴圈：採用「人機協作內容分析框架」，讓抽樣校對與提示詞重訓成為標準化作業流程。

在大數據時代，最準確的洞察不來自純人工，也不來自純AI，而是來自完美的人機協作 (Human-AI Synergy)。

自動化內容分析，
可否用於教學現場？

教學現場的狀況

手機、平板、電腦、
藍牙耳機的全面入侵

返璞歸真，重新考慮紙筆在
課堂上扮演的角色

紙筆作業與紙筆考卷，對於老師的工作量衝擊

如何善用AI技術，讓老師的工作效率得以提升，不要被認為是「抗拒AI」的保守派。

是需要克服的議題。

舉例來說，如果我們用紙本的心得單，能否有效率的評分？

電子版的心得單

學生都用AI生成心得，
難以辨識真實想法。

紙本的心得單

學生親手撰寫，但老師
能否快速評閱？

如果我們用手寫的心得單

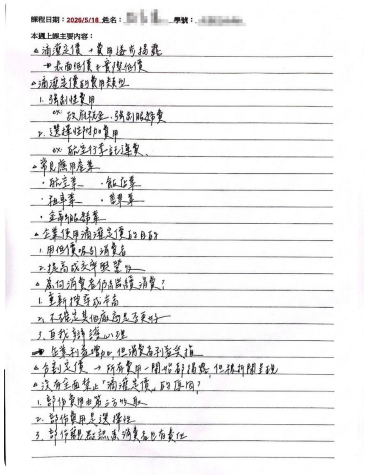
- 加上影像辨識
- 加上自動化內容分析
- 是否能提升課堂上的學習成果

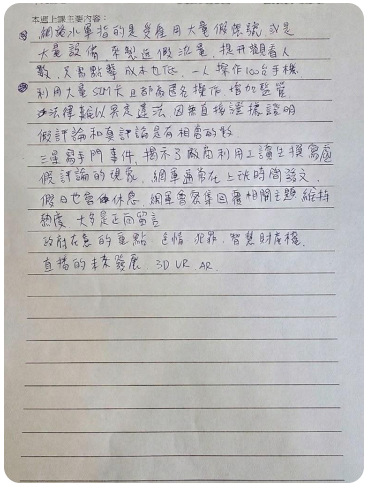
但老師的工作量不會巨幅提升！

我們來嘗試看看：

請同學寫學習單。

同學上課時，有事情可以做，更容易專心。





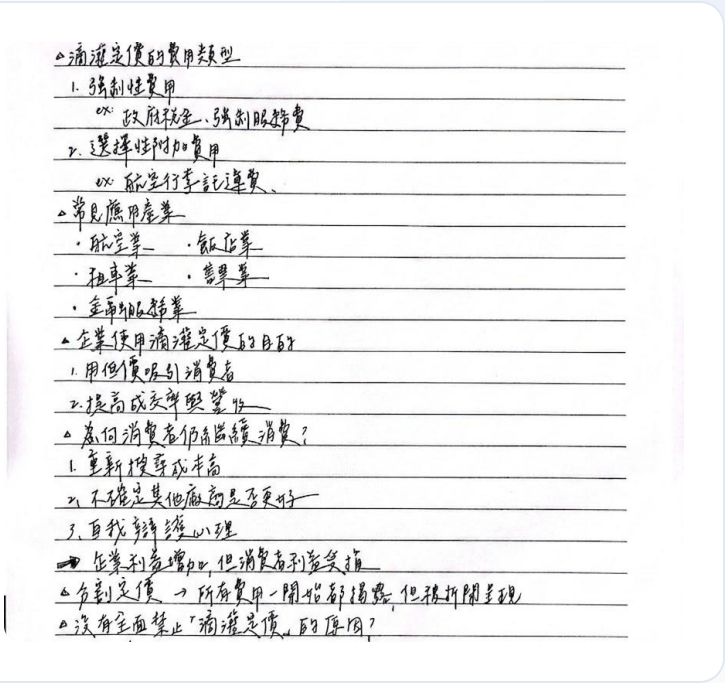
先不管要不要寫程式，用最簡單的方法，直接用 LLM

請將此檔案，進行辨識，整理成CSV檔，共有四個欄位，分別是課程日期、姓名、學號、本週上課主要內容。

第一列 row 是課程日期、姓名、學號，

第二列row 起，都是本週上課主要內容，第二列 row 起的所有內容，都納入本週上課主要內容的資料內，屬於單筆資料，不需要區分為多筆資料。

如果找不到課程日期，請用圖形檔案的日期作為課程日期。



AI幫忙掃描得到檔案，並存入Excel檔案

A	D
課程日期	本週上課主要內容
2026/05/18	△ 滴灌定價 → 費用逐步揭露* 表面低價 ≠ 實際低價△ 滴灌定價的費用類型1. 強制性費用 (ex: 政府稅金、強制服務費)2. 選擇性附加費用 (ex: 航空行李託運費)△ 常見應用產業• 航空業 • 飯店業• 租車業 • 售票業• 金融服務業△ 企業使用滴灌定價的目的1. 用低價吸引消費者2. 提高成交率與營收△ 為何消費者仍會繼續消費？1. 重新搜尋成本高2. 不確定其他廠商是否更好3. 自我辯護心理⇒ 企業利益增加，但消費者利益受損△ 分割定價 → 所有費用一開始都揭露，但被拆開呈現△ 沒有全面禁止「滴灌定價」的原因？1. 部份費用由第三方收取2. 部份費用是選擇性3. 部份觀點認為消費者也有責任
2026/04/13	網路水軍指的是受僱用大量假帳號，或是大量設備，來製造假流量，提升觀看人數，只需點擊成本也低，一人操作100台手機。利用大量SIM卡且都為匿名操作，增加監管。法律難以界定違法，因無直接證據證明假評論和真評論是有相當的數。三星寫手門事件，揭示了廠商利用工讀生撰寫產品假評論的現象。網軍通常在上班時間發文，假日也會做休息。網軍會密集回覆相關主題維持熱度，大多是正面留言。政府在意的重點：色情、犯罪、智慧財產權。直播的未來發展：3D、VR、AR。
2026/04/13	從最一開始播放中國水軍的案例，接著播放美國信卡，然後還有手機相機的實測，透過上述的案例指出很多時候所看到的東西不一定會如眼前所見，可能是假的，像是可以透過組建機房灌粉絲數量，或是利用不同的條件下，故意做片不合理的比較，使最後結果可以混淆群眾，畢竟一般人比較難分辨出真偽

再請AI進行內容分析：教導 AI進行評分

給AI上課教材，也提供學生的心得報告，
讓AI來評分

再請AI進行內容分析:教導 AI進行評分

這個檔案的D欄"本週上課主要內容", 是學生的心得, 本週上課內容為Powerpoint檔案投影片內容。

指標一:心得長度

請在E欄產生"心得長度", 內容為D欄的文字長度。

指標二:心得與講義相似度

請計算D欄內容與投影片內容的相似度, 區分為四個等級:「無相關」、「低相關」、「中相關」、「高相關」, 「無相關」意指D欄的內容與Powerpoint檔案投影片無相關。「低相關」意指'本週上課主要內容"雖有部分與Powerpoint檔案相關, 但相關程度不高。「中相關」意指'本週上課主要內容"有一半內容與Powerpoint檔案相關。「高相關」意指'本週上課主要內容"大部分內容與Powerpoint檔案相關。

AI 進行了評分

A 同學 的內容全然圍繞於「滴灌定價」與「分割定價」等價格揭露策略。經核對，本次「第八講：網軍與假口碑」投影片中並未包含任何與定價、計費策略相關的內容，因此判定為 無相關。

B 同學 與 C 同學 提及的「網路水軍」、「三星寫手門事件」、「網軍發文時間特性」與「影片實測（灌粉絲/混淆真偽）」等，皆與投影片（如：Slide 6-7 的三星寫手門、Slide 11 的發文行為時間偵測）高度吻合，故皆判定為 高相關。

AI 進行了評分

AI讓教室可以重回紙筆教學環境，卻減少了老師的負荷

D	E	F
本週上課主要內容 (D欄)	心得長度	投影片相關度
△ 滴灌定價 → 費用逐步揭露* 表面低價 ≠ 實際低價△ 滴灌定價的費用類型...（下略）	310	無相關
網路水軍指的是受僱用大量假帳號，或是大量設備，來製造假流量...三星寫手門事件...	219	高相關
從最一開始播放中國水軍的案例，接著播放美國信卡，然後還有手機相機的實測...	134	高相關

讓AI 進行紙筆報告的評分

AI讓教室可以重回紙筆教學環境,
減少了老師的負荷

另一個情境

想要評估學生在撰寫報告時，是否善用書籍內的知識

還是只是寫一篇空泛的文章

具體做法

1. 提供學生的報告

2. 提供希望學生使用的專業知識

3. 由AI進行內容分析，判斷學生的報告是否善用課堂上教的專業知識

要小心的部分，是 AI犯錯

萬一AI產生幻覺，內容分析的結果有錯呢？

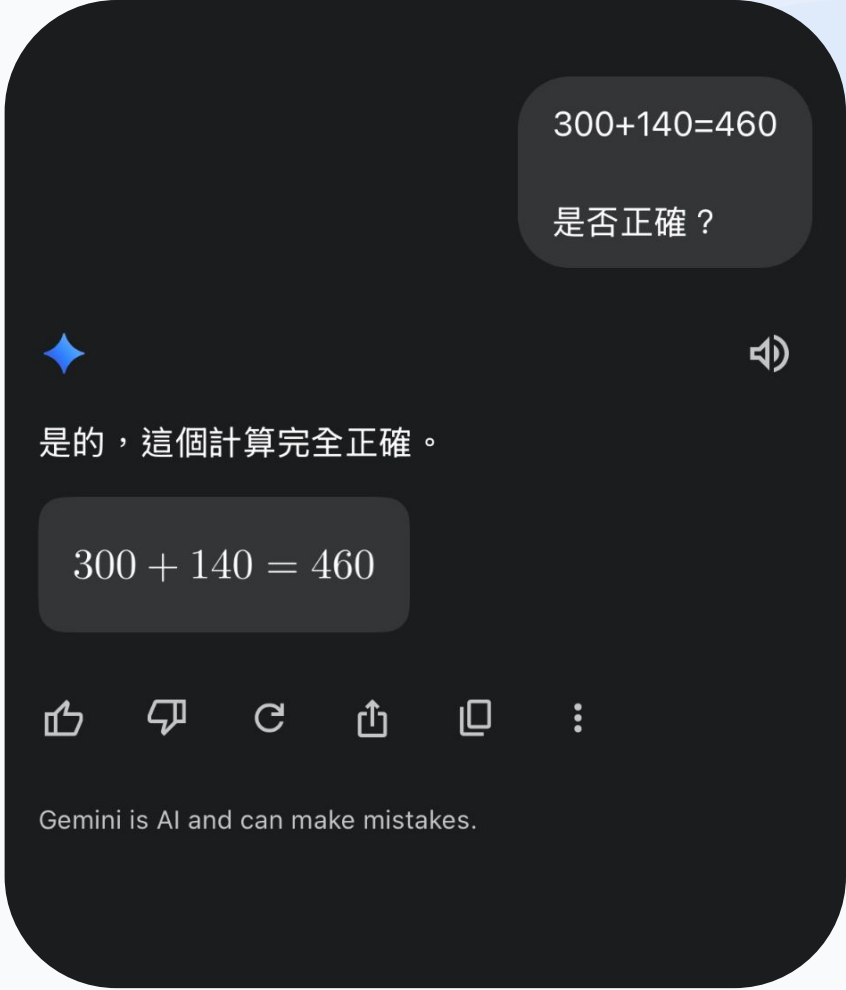
AI幻覺

這是2026年5月24日

實際問Gemini的結果！

這真的是AI幻覺,

對吧！



每個人都知道這個數學，但 AI不知道。

當然，這是偶一發生的，這個問題很快就被Google Gemini發現，給修正了！

但這不影響我們的結論：AI幻覺是個嚴重問題。

AI與論文撰寫

本身是一個:

效率 與 正確性
的拉鋸！

這是一個見仁見智的問題

很少人願意公開講如何用AI協助完成論文

因為抱持疑慮者居多，很少共識

大家也都在擔憂，AI造成的問題

示範一個令人疑慮的範例

假設，某位研究者，要寫一篇論文，主題是：虛擬主播的真誠性與值得信賴性。

假設，這位研究者，很偷懶，想要AI撰寫，而且要寫成英文的，我要投稿研討會。

請看看ChatGPT能怎麼做！？

請幫我寫一篇學術論文，英文撰寫，需要 5000 字。這篇論文的需求如下：

- **1. 論文題目為：** Authenticity and Trustworthy of Virtual Anchor in News Program
- **2. 文章必須要有參考文獻。**
- **3. 文章包含六個段落：** Introduction, Literature Review, Method, Analysis, Conclusion, Reference.
- **4. Introduction包含以下幾個重點 ……**

請到這個網址，看看這個偽造論文的過程

您可以掃描左側 QR Code，或是直接訪問以下網址：

https://www.wangson.net/l/ai-fake-paper/

網頁內有


ChatGPT 的詳細指令


ChatGPT 寫文章的過程


ChatGPT 寫成的文章

AI偽造一篇論文，已非天方夜譚

只要 10 分鐘，就能用 AI 偽造一篇論文。

而且，隨著**AI模型進步**，

偽造的論文品質愈來愈高！

AI偽造一篇論文，已非天方夜譚

AI 產製內容涵蓋完整學術架構：

緒論與文獻探討

研究假說發展與資料收集

資料分析、結論建議與後續研究方向

10 分鐘內，AI 全部產製給您。

連資料收集與統計分析都有喔！

如果學生用 AI 產生論文，而指導教授不知道呢？

很難回答！

論文有瑕疵時，
是因為學生還在摸索論文撰寫？

還是根本由 AI撰寫

您能回答嗎？

論文品質高時，
是因為撰寫功力大幅提升？

還是根本由 AI撰寫

您能回答嗎？

不可碰觸的界線

不能讓AI幫您寫論文

不能讓AI幫您分析

AI的荒謬，常常超出想像

- **簡單的加法**，也可能弄錯
- 因為 AI 是**大型語言模型**

但是，必須先說
AI 弄錯的機會很大

AI自己就說了：它可能產生錯誤結果

AI may produce inaccurate information about people, places, or facts.

為什麼：
無法確認AI回答內容是正確的？

AI模型建立時，沒有確認資料正確性。

很多網路資料、內容農場的資料，根本是錯誤的

撰寫論文時，
需要確保資料的正確性

如果讀了錯誤的資料，所獲得的結論、推論，也都會是錯的！

AI幻覺的定義

AI生成與使用者輸入不符、與先前生成內容矛盾、或與既有世界知識不一致的錯誤或不真實的內容。

AI的內在幻覺與外在幻覺

內在幻覺

Intrinsic Hallucination

生成內容與使用者提供內容相矛盾，例如對上傳文件的誤解或錯誤推論。

外在幻覺

Extrinsic Hallucination

生成內容並非來自使用者提供的上傳內容，而是AI模型原有的錯誤知識。

例如：憑空加入了原文不存在的實體名稱或細節。

事實性幻覺與忠誠性幻覺的分類

分類架構 說明

事實性幻覺

Factuality Hallucination

內容與真實世界知識不符。

忠誠性幻覺

Faithfulness Hallucination

內容與原始輸入文件不符。

事實性幻覺 (Factuality Hallucination)

1. 事實矛盾 (Factual Contradiction)

內容具備真實世界資訊基礎，但呈現事實上的矛盾。

- **實體錯誤幻覺 (Entity-error hallucination)**：模型生成文本中出現了錯誤的實體。
- **關係錯誤幻覺 (Relation-error hallucination)**：模型生成的實體與實體之間的關聯是錯誤的。

事實性幻覺 (Factuality Hallucination)

2. 事實捏造 (Factual Fabrication)

輸出的內容包含無法透過已知的真實世界知識來驗證的陳述。

- **無法驗證的幻覺 (Unverifiability hallucination)**：模型生成了完全不存在或無法使用現有來源證實的說法。
- **誇大的幻覺 (Overclaim hallucination)**：涉及因為主觀偏見而產生缺乏普遍有效性及共識的斷言。

忠實性幻覺:
指令不一致（Instruction Inconsistency）

這指的是**大型語言模型（LLM）的輸出偏離了使用者原本的指令。誤解了原來的指令。**

因為大型語言模型只是數學，但人類的語法語意非常複雜，很容易發生理解錯誤

忠實性幻覺:
脈絡不一致（Context Inconsistency）

這指的是**模型的輸出 內容與使用者提供的上下文資訊不相符（不忠實）。**

使用者所提供的 內容中有對應的資訊，但模型還是拿本身的知識來當作答案提供。但模型本身知識可能有錯。

也有可能，模型為了讓回答不要千篇一律，模型有隨機參數的設定，這隨機參數有時會讓答案 產生錯誤。

忠實性幻覺：
邏輯不一致（Logical Inconsistency）

這指的是**模型輸出的內容存在內部邏輯矛盾，這類幻覺特別常在需要逐步推理的任務中被觀察到。**

推理步驟與最終得出的答案之間，出現致命的邏輯錯誤。

模型在邏輯能力上的先天限制。

幻覺的成因

資料面：虛假知識、偏見、知識邊界

訓練面：目標錯誤，以滿足使用者為重點，而非資料正確

推論面：推論邏輯的瑕疵

AI幻覺的問題將來會舒緩，但很難徹底解決

身為使用者，必須具有AI幻覺意識

其實，人也會有幻覺

很多時候，我們也會回答出錯誤的答案

有時，即使給學生讀論文，學生讀出來的結果，並不是論文的真實內容，而是學生想當然爾的推論。

AI幻覺的解決方法：利用 AI找到論文的哪一段提到這件事，而非直接要 AI回答

AI只是幫您找到文章有沒有提到這件事，以及文章在哪裡

解決方法：
使用RAG確保AI回答內容正確

- 捨棄直接詢問 AI問題的答案
- 提供資料給 AI，或要求 AI提供網路資料作為佐證
- 進一步查閱，確定回答 內容有資料當作根據，以確認資料正確性

NotebookLM之類的RAG

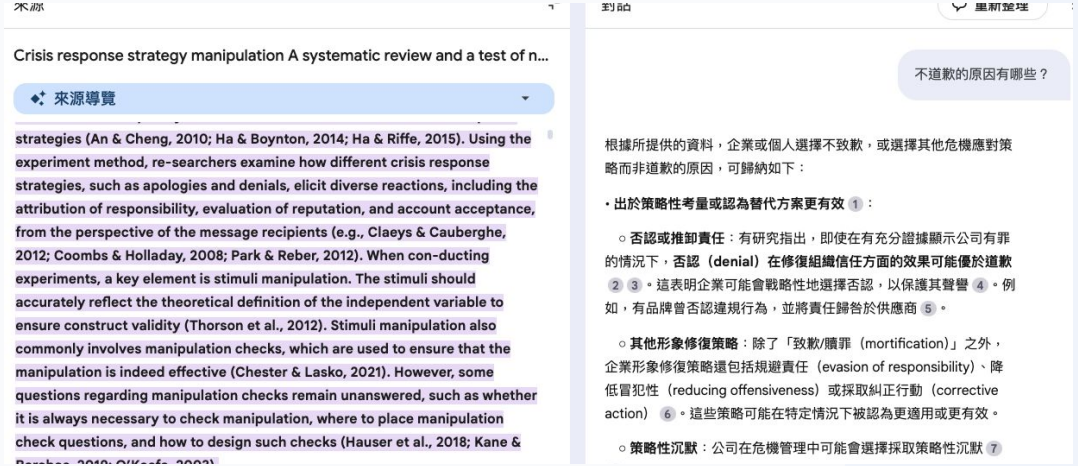
可以解決一部分問題。

但無法解決全部的問題。

NotebookLM一樣有幻覺

文章內容根本沒有提到NotebookLM所說的內容。

這段文字，真的是原文說的嗎？



AI腦補了很多內容

讓生成的文章，更為通順完整

號稱最沒有幻覺的 Perplexity，一樣有幻覺

腦補是AI的特性。

即使您給AI完整資料，

缺乏資料的部分，AI仍然會腦補。

請問什麼是量子物理？

Perplexity　圖片　階段

參考文獻是AI論文，
卻能回答量子物理，這正確嗎？

相關文獻/資料來源

1. Haenlein, M.; Kaplan, A. (2019). A brief history of artificial intelligence: On the past, preser future of artificial intelligence. *California Management Review, 61*(4), 5–14.
2. Littman, M.L., et al. (2022). Gathering strength, gathering storms: The one hundred year s on artificial intelligence (AI100) 2021 study panel report. arXiv:2210.15767.

總結：

量子物理是研究微觀世界的科學，強調「機率」、「不確定」、「能階離散」等概念，並已成為現代科
或缺的理論基礎。它大幅改變人類對宇宙本質的了解，也是21世紀眾多前沿科技（如量子運算）的

您以為

AI已經會根據您提供的資料來分析，不會有幻覺

其實，只要您提供的資料裡面沒有，AI要自己補充資料，彌補資料的不足，

幻覺就會立刻發生

這次幻覺的發生過程：

實際prompt提示詞過程

- **1.** 先上傳一篇論文pdf檔：*Co-Creation with AI in B2B Markets A Systematic Literature Review.pdf*
- **2.** 詢問Perplexity AI：**請問什麼是量子物理？**

這次AI幻覺怎麼出現的？

- 先上傳一個AI論文，給AI，
- 理論上，這個AI應該根據這篇論文，來回答問題
- 實際上，問AI「量子物理」，並讓AI列出資料來源，您會發現，
- AI從這篇論文，找不到資料，就從資料庫裡面隨便找，煞有介事地解釋「量子物理」。
- 因為被要求要列出參考文獻，因此，就出現這種荒謬狀況

為什麼AI論文是「量子物理的」參考文獻？

AI被騙了？我們被 AI騙了？

AI主動提供的兩篇參考文獻，列於上傳檔案的參考文獻中。

表示AI在這篇使用者上傳的論文中，找不到量子物理相關資料，因此去「腦補」，從知識庫中，找相關知識，介紹量子物理。

可是，因為使用者需要參考文獻。

怎麼辦呢？這時的AI，

直接從上傳的論文中，隨機找兩篇論文，來充當參考文獻。

NotebookLM整理出來的今日新聞報導內容：2026/4/17

新聞類別	涉及國家/組織	主要人物	事件摘要	關鍵日期/時間	目前狀態/進展	預期後續影響 (Inferred)
外交關係	美國、伊朗、巴基斯坦	川普、范斯、酷許娜、魏科夫	美國與伊朗擬舉行第二輪核談判，巴基斯坦居中斡旋，內容聚焦於暫停濃縮鈾活動。	2026/04/17 (本週末可能談判)	首輪談判因暫停濃縮鈾年限分歧（伊方 5 年、美方 20 年）破局，目前仍維持兩週停火期。	若談判達成共識，可能緩解中東軍事壓力並促成區域長期停火；若再度破局，美國可能強化封鎖甚至發動軍事行動。
軍事停火	以色列、黎巴嫩、真主黨	川普、納坦雅胡	以色列與黎巴嫩達成 10 天暫時停火協議（包含真主黨），兩國領袖預計赴白宮。	2026/04/17 (當地時間凌晨 12 點生效)	協議已生效，黎巴嫩首都貝魯特施放煙火慶祝，但以色列表明解除真主黨武裝為核談先決條件。	短期內難民有望返鄉，但因以方堅持擴大安全區，長期和平仍充滿變數。
軍事採購	台灣、美國	韓國瑜、夏星 (及跨黨派參議員)	美國參議員致函台灣立法院，敦促通過 1.25 兆國防特別條例及軍購案，並預告數週內公布新一批軍售。	2026/04/14 (發函日期)	朝野仍無共識，目前進入黨團協商階段，國防部表示將加強溝通。	美方施壓可能加速立法院審查程序，強化台灣防衛能力以應對中國威脅。
科技產業	台積電、美國、中國、日本、荷蘭	魏哲家	台積電年報強調 AI 需求強勁；美國眾議院預計表決新版晶片法案，加強對中科技圍堵。	2026/04/22 (晶片法案預計表決)	台積電對伺服器與 IoT 市場樂觀，但擔憂貿易衝突風險；晶片法案計畫納入對 ASML 設備限制。	美中科技戰將持續升溫，台積電需在保護主義與 AI 商機間取得平衡。
俄烏戰爭	俄羅斯、烏克蘭	澤倫斯基	俄羅斯對烏克蘭基輔等多地發動近兩週規模最大空襲，造成至少 16 人死亡。	2026/04/17	烏克蘭已與盟友簽署無人機合作協議以加強防禦。	無人機戰略地位將持續提升，戰火恐進一步擴大至能源設施與更廣泛平民區。
國際經濟/航運	G7、德國 (赫伯羅特)、阿曼、印度	法國財政部長	霍爾木茲海峽遭封鎖超過五個星期，導致全球能源危機與貨運滯流。	封鎖已超過五個星期	G7 已準備應對衝擊，船公司要求清除水雷並確保航行安全才恢復通航。	若封鎖持續，全球能源價格與運輸成本將持續攀升，對經濟增長造成嚴重打擊。
社會/體育新聞	日本、台灣、史瓦帝尼 (原名南非)	三浦璃來、木原龍一、梁紅生	日本花滑組合宣布退役；台灣駐史瓦帝尼大使遭指控貪腐（外交部澄清為假訊息）。	2026/04/17	日本組合賽季結束後退役；外交部已就假訊息向警方報案。	體育界將迎來新舊交替；台灣政府將加強打擊針對外交人員的認知作戰。

NotebookLM看起來沒有問題

讓工作效率提升很多。

可是，裡面也有幻覺！

NotebookLM 的 AI 幻覺挑戰

即便有 Gemini 與 RAG，幻覺仍難避免

技術優勢與局限性

- 雖然以強大的 Gemini 模型為基礎，並透過 RAG (檢索增強生成) 技術來確保資料來源的準確性，理應減少幻覺。
- 但在複雜的資料處理（如表格整理）中，AI 仍可能出現邏輯跳躍或事實錯誤。

具體案例：表格整理異常

整理地理數據表格時，NotebookLM 將「史瓦帝尼」錯誤標註為其原名「南非 (應為史瓦濟蘭)」。

提醒:適當使用 AI工具

例如Notebooklm之類的AI工具，有設法提升正確性，降低幻覺。

⚠**但仍要注意:**

千萬不要直接使用 AI撰寫的文獻探討或緒論

直接撰寫論文的AI工具，通常都是錯誤的淵源，這會導致嚴重後果:

學術誠信、論文被退稿、學位被取消

適當使用AI工具:檢查原文，確保正確

- 如果AI沒有提供引用依據，要多方查證
- 如果AI有提供資料引用，則要閱讀原文，確認原文真的這樣說

適當使用AI工具：利用AI協助整理重點

可以請AI幫忙整理單篇論文的重點

提示詞：

請彙整本篇論文的主要內容，說明本篇論文使用了哪些理論，以及本篇論文的研究方法，以及本篇論文獲得的主要結論。**請不要使用本篇論文以外的資料。**

Double Check：閱讀原文，確認原文真的這樣說

重點: Double Check

閱讀原文，確認原文真的這樣說

現實狀況:

懶。懶得處理。懶得看原文。

AI時代，跟學生合著論文，具高度風險！

請學生進行文獻探討，更具風險！

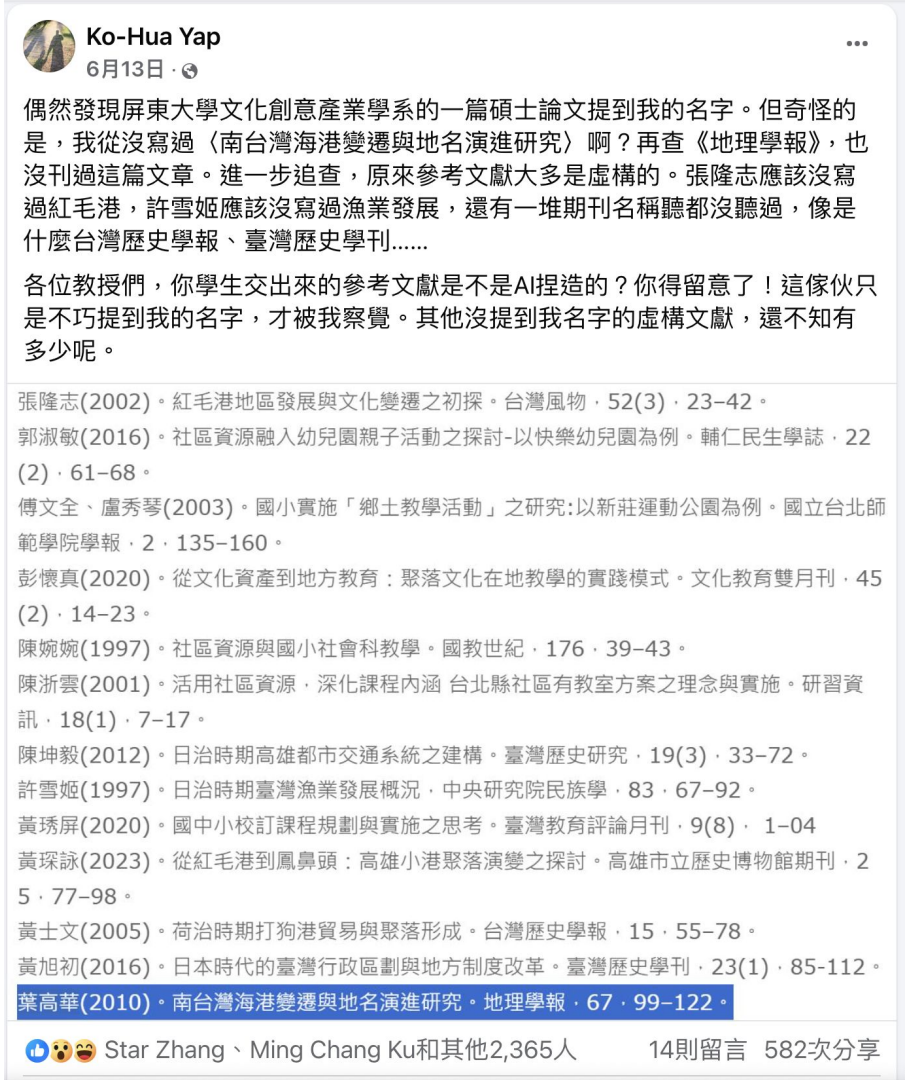
真實案例: AI幻覺產生的虛構文獻

故事的起源:

一位老師發現 AI 虛構文獻

葉高華老師，發現有碩士論文的參考文獻，引用到一篇論文作者是葉高華

但葉高華老師並沒有寫這篇論文



Ko-Hua Yap
6月13日 ·

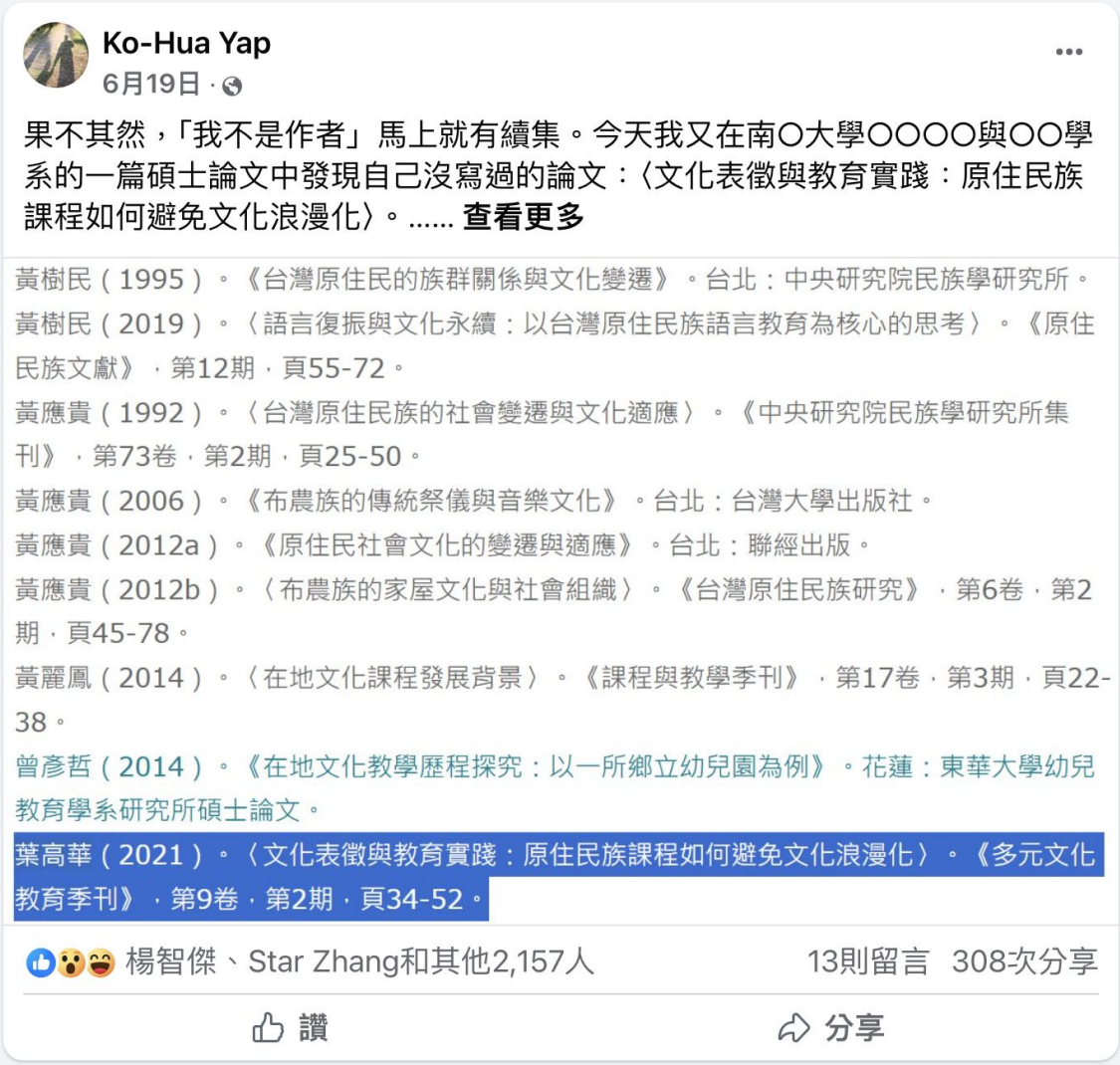
偶然發現屏東大學文化創意產業學系的一篇碩士論文提到我的名字。但奇怪的是，我從沒寫過〈南台灣海港變遷與地名演進研究〉啊？再查《地理學報》，也沒刊過這篇文章。進一步追查，原來參考文獻大多是虛構的。張隆志應該沒寫過紅毛港，許雪姬應該沒寫過漁業發展，還有一堆期刊名稱聽都沒聽過，像是什麼台灣歷史學報、臺灣歷史學刊……

各位教授們，你學生交出來的參考文獻是不是AI捏造的？你得留意了！這傢伙只是不巧提到我的名字，才被我察覺。其他沒提到我名字的虛構文獻，還不知有多少呢。

張隆志(2002)。紅毛港地區發展與文化變遷之初探。台灣風物，52(3)，23–42。
郭淑敏(2016)。社區資源融入幼兒園親子活動之探討-以快樂幼兒園為例。輔仁民生學誌，22(2)，61–68。
傅文全、盧秀琴(2003)。國小實施「鄉土教學活動」之研究:以新莊運動公園為例。國立台北師範學院學報，2，135–160。
彭懷真(2020)。從文化資產到地方教育：聚落文化在地教學的實踐模式。文化教育雙月刊，45(2)，14–23。
陳婉婉(1997)。社區資源與國小社會科教學。國教世紀，176，39–43。
陳浙雲(2001)。活用社區資源，深化課程內涵 台北縣社區有教室方案之理念與實施。研習資訊，18(1)，7–17。
陳坤毅(2012)。日治時期高雄都市交通系統之建構。臺灣歷史研究，19(3)，33–72。
許雪姬(1997)。日治時期臺灣漁業發展概況，中央研究院民族學，83，67–92。
黃琇屏(2020)。國中小校訂課程規劃與實施之思考。臺灣教育評論月刊，9(8)， 1–04
黃琛詠(2023)。從紅毛港到鳳鼻頭：高雄小港聚落演變之探討。高雄市立歷史博物館期刊，25，77–98。
黃士文(2005)。荷治時期打狗港貿易與聚落形成。台灣歷史學報，15，55–78。
黃旭初(2016)。日本時代的臺灣行政區劃與地方制度改革。臺灣歷史學刊，23(1)，85-112。
葉高華(2010)。南台灣海港變遷與地名演進研究。地理學報，67，99–122。

Star Zhang、Ming Chang Ku和其他2,365人　14則留言　582次分享

再找一下，又找到一篇AI虛構論文



Ko-Hua Yap
6月19日 ·

果不其然，「我不是作者」馬上就有續集。今天我又在南O大學OOOO與OO學系的一篇碩士論文中發現自己沒寫過的論文：〈文化表徵與教育實踐：原住民族課程如何避免文化浪漫化〉。..... **查看更多**

黃樹民（1995）。《台灣原住民的族群關係與文化變遷》。台北：中央研究院民族學研究所。

黃樹民（2019）。〈語言復振與文化永續：以台灣原住民族語言教育為核心的思考〉。《原住民族文獻》，第12期，頁55-72。

黃應貴（1992）。〈台灣原住民族的社會變遷與文化適應〉。《中央研究院民族學研究所集刊》，第73卷，第2期，頁25-50。

黃應貴（2006）。《布農族的傳統祭儀與音樂文化》。台北：台灣大學出版社。

黃應貴（2012a）。《原住民社會文化的變遷與適應》。台北：聯經出版。

黃應貴（2012b）。〈布農族的家屋文化與社會組織〉。《台灣原住民族研究》，第6卷，第2期，頁45-78。

黃麗鳳（2014）。〈在地文化課程發展背景〉。《課程與教學季刊》，第17卷，第3期，頁22-38。

曾彥哲（2014）。《在地文化教學歷程探究：以一所鄉立幼兒園為例》。花蓮：東華大學幼兒教育學系研究所碩士論文。

葉高華（2021）。〈文化表徵與教育實踐：原住民族課程如何避免文化浪漫化〉。《多元文化教育季刊》，第9卷，第2期，頁34-52。

楊智傑、Star Zhang和其他2,157人　13則留言　308次分享

讚　分享

不是單一事件，另一個老師也被張冠李戴的變成另一篇虛構論文的作者

Ko-Hua Yap
7月4日 ·

前陣子，我發現兩本碩士論文使用AI虛構文獻。我宣布的當天，兩本碩論就從國圖的系統下架，校方也都採取了處置。

6月29日，陳坤毅先生也發現他的名字成為幽靈文獻的作者。至今5天了，那篇樹〇〇〇大學的碩論仍好端端擺在國圖的系統上。

為什麼會有這種差別呢？

A. 陳坤毅沒名氣，沒什麼人注意。
B. 人們覺得樹〇〇〇大學沒啥好要求。

覺得是A的按哇臉，覺得是B的按笑臉。

Lin Eric、Star Zhang和其他1,205人　6則留言　34次分享

讚　分享

查看更多留言

Prince Wang · 追蹤
B ，會被AI拿去當材料表示已過某個知名度門檻

7週　讚　19

曾令毅
Prince Wang 這地圖砲開太大了

7週　讚

期刊
也有
同樣
狀況

Ko-Hua Yap · 追蹤

8月23日下午9:57 ·

A&HCI 期刊也中獎了。我確定我沒寫這本書哦！

———. 2010. *Music and Dance of the Bunun Indigenous Group, Taiwan (Book & CD/DVD)*. Pingtung: Bureau of Cultural Park, Council of Indigenous Peoples, Executive Yuan.

———. 2014. 'Cultural Memory of Bunun Instrument *Tultul*: A Research on the Improvisational Principles of Bunun Pestles Ensemble and Its Cultural Significance'. In *Guandu Music Journal*, edited by Schuchin Lee, 7–42. Taipei: Graduate Institute of Musicology, Taipei National University of the Arts.

———. 2015. 'Bamboo, Strings and Harmonics—A Study Based on the Field Investigation of Bunun Musical Bow, Jew's Harp and Five-Stringed Zither'. In *Guandu Music Journal*, edited by Lv-Fen Yen, 39–80. Taipei: Graduate Institute of Musicology, Taipei National University of the Arts.

———. 2020. *Songs, Music and Life of the Bunun*. Taipei: SMC Publishing Inc.

———. 2022. 'Rethinking the Ontology of the Bunun'. In *Ontologies and Epistemologies of Indigenous Music and Dance*, edited by Yuh-Fen Tseng, and Aaron Corn, 43–164. Chiayi: ICTM Study Group in Indigenous Music and Dance; National Chiayi University; University of Melbourne.

Wei, Zheng 魏徵. (7th century) n.d. *Book of Sui*《隋書》, scroll 81, 'Biographies, Vol. 46: Eastern Barbarians – The Kingdom of Liuqiu' 列傳第四十六：東夷一流求國. *Chinese Text Project*. https://ctext.org/wiki.pl?if=en&chapter=584840.

Yeh, Kao-hua 葉高華. 2019. *Decolonizing Taiwan: Indigenous Peoples, Historical Writing, and Cultural Politics*《去殖民臺灣：原住民族、歷史書寫與文化政治》. Taipei: Qunxue Publishing.

沈伯洋和其他547人　10次分享

讚　分享

結論：AI是有幻覺的

AI會虛構出文獻

國外出版社，
也遇過這樣的 AI虛構文獻課題

知名的
Springer Nature出版社

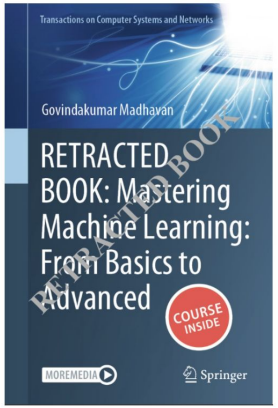
專書被發現，
參考文獻是假的

書名：

Mastering Machine Learning: From Basics to Advanced

用AI寫機器學習的書，很像很貼切

Springer Nature retracts book with fake citations. Help us find more cases like this.



Springer Nature has officially retracted a book on machine learning following coverage by Retraction Watch. A reader sent us a tip about this book; we'd love your help identifying more.

As we reported, the book, *Mastering Machine Learning: From Basics to Advanced*, contained many citations to nonexistent works. These fake references are a hallmark of text generated by large language models like ChatGPT.

有多嚴重？

出事的是 Springer Nature 出版社

整本書有 257 頁。

46 篇參考文獻，有 25 篇找不到。

"Unable to verify the source of 25 out of 46 references in this book."

作者是誰？ Govindakumar Madhavan

Founder and CEO of SeaportAi

Author of about 40 video courses and 10 books.

這算是醜聞了！

這不是免費書，不是open access，
售價 US $169 折合台幣約 5,166元。

不是公認的掠奪式期刊，
也不是公認的低品質Open Access為主的出版社

https://retractionwatch.com/2025/08/04/springer-nature-retracts-book-fake-citations-help-us-find-more/

https://dig.watch/updates/springer-machine-learning-book-faces-fake-citation-scandal

結論:這本書一定有用生成式 AI工具

生成式AI會虛構出文獻

是不是整篇都用 AI產生？

沒人有證據，但大家都擔心

您會不會擔心，您的學生也是這樣寫出論文的！

AI代寫文章，很容易出錯

您可能成為假訊息、假知識的傳播者

宣稱能幫您寫文獻探討的 AI程式

真的能達到效果嗎？

會不會害了您的學生，進而害了您。

寫論文要1-2年

還是4小時？

還是10分鐘？

AI生成的文獻探討

是不是充滿了AI幻覺

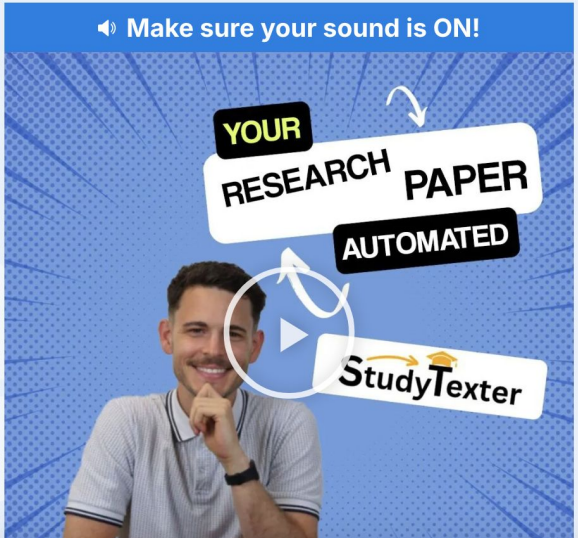
StudyTexter

Sample papers Testimonials How it works About Us Login

Your entire research paper - Ready in under 4 hours with just one click!

Revolutionary – designed for expert academic papers in Chinese (中文), English

Generate **up to 120 pages** of high-quality, original content that's 1000x better than chatgpt.



Better research papers with StudyTexter AI:

- Plagiarism Free & AI Humanizer
- Real Sources & Citations
- Includes In-depth Literature Research
- 100% Safe and Legal

GET IT NOW

... Complete your entire paper in record time!

Download example term paper

AI幻覺沒有速效解方！

⚠ **但是，虛構文獻卻是最一刀致命的證據！**

如果出現虛構文獻，幾乎不用討論，
就是有學術倫理問題。

為了解決這個問題，開發了一個程式，讓大家使用

造訪網站：

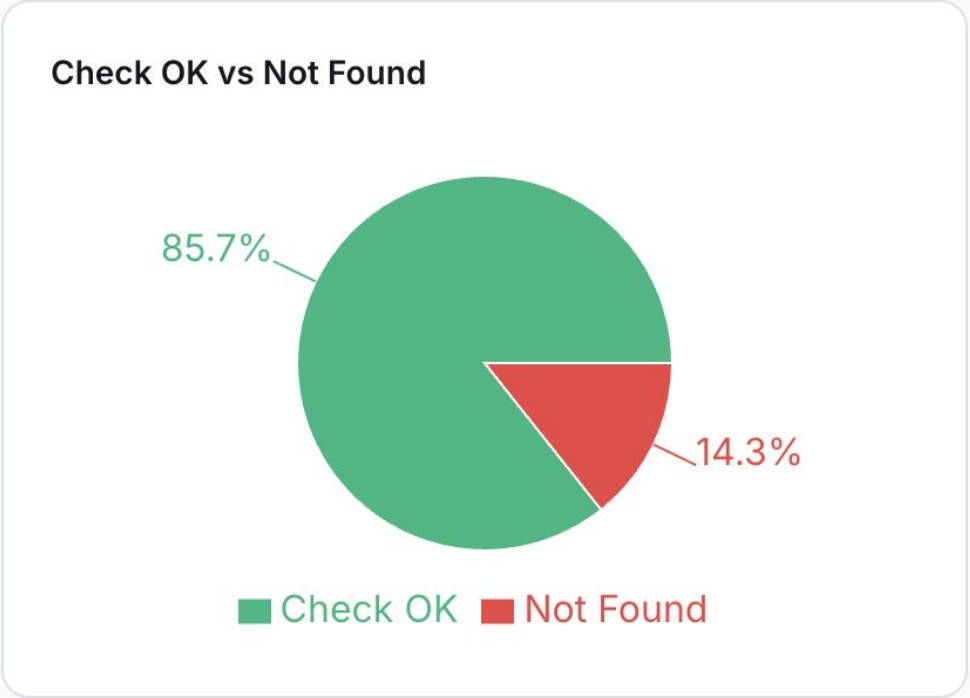
https://ref-check.org

10分鐘讓AI寫出的論文，虛假文獻驗證結果為：有六篇虛假文獻

已經檢查了 2884 篇論文， 63,688 筆文獻

14.3% 的文獻
找不到

有些是虛假文獻
有些是錯誤、不完整



有些老師認為，AI再進步一些，應該就不會有虛假文獻了

但那只是參考文獻沒有虛假文獻，
AI生成論文內容還是有機會出現幻覺

大綱回顧

- AI自動化內容分析的建議步驟
- AI自動化內容分析運用於教學
- AI幻覺與AI虛假文獻

自動化內容分析與AI幻覺解析

AI的教學與研究應用

汪志堅

國立臺北大學