生成式 AI 能力評估標準

打造臺灣 AI 落地的關鍵量尺

講者介紹

國科會 TAIDE 專案共同主持人

教育部 Write AI、華師拍檔 專案主持人

中華民國人工智慧學會理事長

台灣數位人文學會理事長

國立中央大學資工系教授

中研院人社中心研究員

臺大語言所合聘教授

蔡宗翰博士

thtsai@csie.ncu.edu.tw

粉絲團：AI界李白

IG/threads: ai.libai

演講大綱

1. **破解迷思**：MMLU 的「文字接龍」本質
2. **典範轉移**：從 Coding 到 **Vibe Coding**
3. **教學現場**：台大文學院的 Benchmark 實作
4. **RAG 實戰**：**TAIHU Benchmark** (台鵠計畫)
5. **臺灣首創**：TAIDE 五大任務導向評測
6. **技術解密**：LLM-as-a-Judge 與自動化評分
7. **結論與展望**

前OpenAI的首席科學家伊利亞（Ilya Sutskever）最近參加了播客節目，這次他已經是自己的創業公司的CEO了。這次談話，他把當前AI領域很多痛點都揭示了出來

伊利亞第一個話題說的就是，研發AI模型的人現在都在自己騙自己，最明顯的表現就是基準測試的成績直接封神，而等你實際用起來就發現也就是個 半獸人

第一節 MMLU 的真相 (1/4)

我們常聽到的 MMLU 到底是什麼？

- **表象**：
  - 涵蓋 57 個學科 (數學、歷史、法律...)
  - 形式看起來像人類的「四選一」選擇題
- **大眾的誤解**：
  - 以為模型像人類一樣，看完題目，思考後「圈選」答案
- **實際的運作**：
  - 它是在算**機率**，不是在做選擇

第一節 MMLU 的真相 (2/4)

科普：MMLU 的算法其實是「文字接龍」

- **Next Token Prediction (下一個字預測)**：
  - 模型讀入題目：「台灣最高的山是？ (A)陽明山 (B)玉山... 答案：」
  - 模型**不會**真的去選 (B)
  - 模型會計算接下來出現字元 A, B, C, D 的**條件機率 (Log-likelihood)**
- **決勝負的方式**：
  - 如果 P(B) = 0.8，P(A) = 0.1
  - 因為 0.8 最高，所以我們說模型「答對了」

第一節 MMLU 的真相 (3/4)

為什麼 MMLU 只適合評測 Base Model？

- **Base Model (基底模型)**：
  - 訓練目標就是「預測下一個字」(Text Completion)
  - 非常適合用 MMLU 這種機率算法來測
- **Chat Model (對話模型)**：
  - 經過指令微調 (SFT) 與人類對齊 (RLHF)
  - 訓練目標是「對話」與「遵循指令」，機率分佈已被改變
- **結論**：
  - 用 MMLU 測 Chat Model，就像用「背誦能力」來考「辯論選手」
  - **這就是為什麼我們需要任務導向評測**

第一節 MMLU 的真相 (4/4)

Benchmark 的真正意義

- **如果只看 MMLU**：
  - 我們只知道模型「背」了多少書
- **產業界需要的是**：
  - 模型能不能「寫公文」？
  - 模型能不能「看財報」？
  - 模型能不能「寫 RAG 系統用的 Query」？
- **這導致了評測典範的轉移**

第二節 領域專家與 Vibe Coding (1/3)

誰來製作 Benchmark？

- **過去的誤區**：
  - 認為做 AI 評測是資工系 (CS) 的事
- **現在的真相**：
  - **資工系不懂法律，不懂公文，不懂醫療**
  - 只有**領域專家**才知道什麼是「好的回答」
- **黃金標準 (Gold Standard) 的來源**：
  - 必須由律師、醫師、文學家、公務員來定義

第二節 領域專家與 Vibe Coding (2/3)

什麼是 Vibe Coding？

- **Vibe Coding (氛圍編碼/直覺編碼)**：
  - Andrej Karpathy 提出的概念
  - 透過 AI 輔助，寫程式不再是工程師專利
  - 使用者用**自然語言**描述邏輯，AI 負責寫 Code
- **對評測的意義**：
  - **賦權 (Empowerment)**：文學院學生、法務人員現在可以用 Vibe Coding 技術，自己寫出自動化評測腳本
  - **Domain Know-how > Coding Skills**

第二節 領域專家與 Vibe Coding (3/3)

新時代的評測團隊組成

- **不再是**：一群工程師在實驗室跑數據
- **而是**：
  - **Subject Matter Experts (SME)**：出題、制定評分標準 (Rubric)
  - **AI Engineers**：架設評測 Pipeline
  - **LLM as Judge**：執行大量批改

第三節 實踐案例 I：台大文學院 (1/2)

跨領域的教學實驗

- **場景**：台灣大學 數位人文技術與應用課程
- **學員**：文學院學生 (非資工背景)
- **目標**：訓練學生利用自身的人文素養，製作高品質 Benchmark
- **方法**：
  - 根據他們自己提出的任務，去設計評測 AI 的題目與答題標準
  - 利用 Vibe Coding 工具 (如 Colab, AI 輔助)

第三節 實踐案例 I：台大文學院 (2/2)

文科生的逆襲

- **成果**：
  - 學生製作出針對「高中歷史考題」、「指代消解」的深度評測
- **發現**：
  - 文科生設計的題目，就是用戶會下的指令，**比工程師問 AI 得到的題目真實太多了**
- **啟示**：
  - **最好的 Prompt Engineer 和 Benchmark Designer，往往來自人文學科**

第四節 實踐案例 II：TAIHU Benchmark (1/5)

台鵠開示：台灣人文 RAG 評測平台

- **為什麼要做？**
  - 台灣歷史文獻具有高度**非結構化**與**時間敏感性**
  - 通用模型 (如 GPT-4) 容易產生歷史幻覺 (Hallucination)
- **平台定位**：
  - **首個針對台灣歷史文獻**的 RAG (檢索增強生成) 基準測試
  - 整合**明清檔案**、**省議會公報**、**熱蘭遮城日誌**等異質資料
  - 目標是建立人文研究與 AI 技術的對話標準

第四節 實踐案例 II：TAIHU Benchmark (2/5)

核心突破：Benchmark 是怎麼做出來的？

我們經歷了從「全人工」到「半自動化」的演進：

1. **Phase 1: 全人工出題 (Pilot)**
    - 以《熱蘭遮城日誌》為例，由歷史系學生閱讀史料，手寫題目與標準答案。
    - **缺點**：速度慢、成本高，難以規模化。
2. **Phase 2: 半自動化生成 (Scale-up)**
    - **逆向工程**：不先想題目，而是從「史料材料」出發。
    - **LLM 生成**：輸入史料片段，要求 LLM 生成對應的 Q&A。
    - **SME 審核**：**歷史系研究生**介入審核，確認 Query 與 Answer 的邏輯關係與證據力。

第四節 實踐案例 II：TAIHU Benchmark (3/5)

平台機制：雙重排行榜 (Dual Leaderboard)

為了確保學術評測的公正性與防止模型「作弊」：

- **公開排行榜 (Public)**：
  - 開放部分測試資料，讓研究者在開發期獲得即時反饋。
- **私人排行榜 (Private)**：
  - **保留完整測試集**，不對外公開。
  - 用於最終權威評估，防止模型對題目進行 Overfitting (過度擬合)。
- **自動化評分 Worker**：
  - 研究者上傳 CSV，後端自動計算分數並生成分析報告。

第四節 實踐案例 II：TAIHU Benchmark (4/5)

評估標準：擁抱 nDCG

第四節 實踐案例 II：TAIHU Benchmark (5/5)

涵蓋資料集：穿越時空的考驗

我們整合了跨越數百年的異質史料，考驗 AI 的跨時代理解力：

1. **明清檔案**：考驗對古代行政公文、封建制度的理解。
2. **台灣省議會公報**：考驗對近代民主辯論、政策脈絡的掌握。
3. **台灣海防檔**：考驗軍事佈署與外交關係的空間概念。
4. **未來規劃**：納入《漢文台灣日日新報》（媒體視角）與國家文化記憶庫（常民生活）。

平台展示 (1/4)

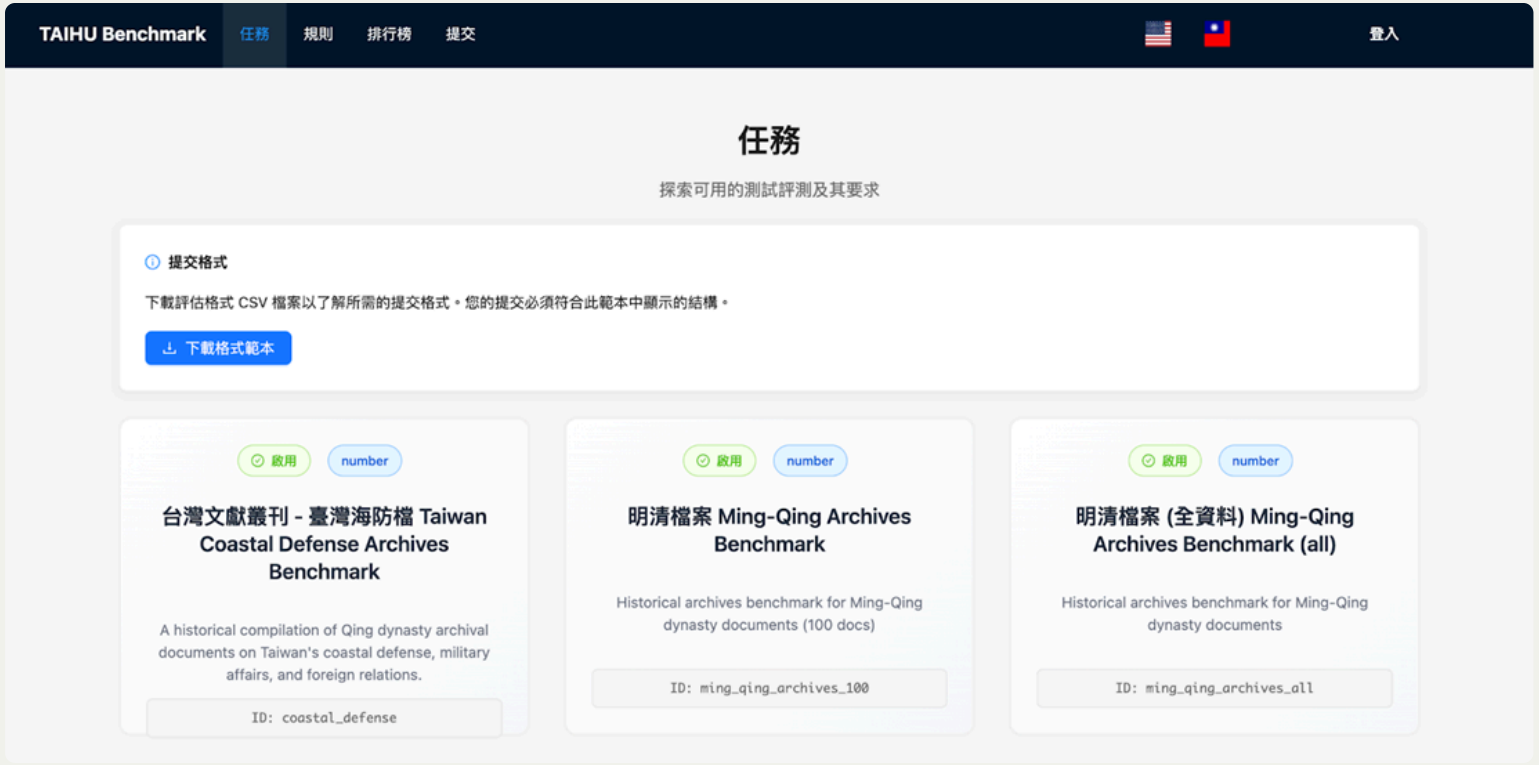
直觀的評測儀表板 (Dashboard)

- **現代化介面**：React 18 + Ant Design 5
- **功能概覽**：GitHub 登入、任務卡片、即時公告

平台展示 (2/4)

任務選擇與規範 (Tasks)

- **明確的規格**：下載 CSV 範本、查看評測 Metric (如 nDCG)
- **資料集多樣性**：包含明清檔案、省議會公報等史料


TAIHU Benchmark
任務
規則
排行榜
提交
登入
任務
探索可用的測試評測及其要求
提交格式
下載評估格式 CSV 檔案以了解所需的提交格式。您的提交必須符合此範本中顯示的結構。
下載格式範本
啟用
number
台灣文獻叢刊 - 臺灣海防檔 Taiwan Coastal Defense Archives Benchmark
A historical compilation of Qing dynasty archival documents on Taiwan's coastal defense, military affairs, and foreign relations.
ID: coastal_defense
啟用
number
明清檔案 Ming-Qing Archives Benchmark
Historical archives benchmark for Ming-Qing dynasty documents (100 docs)
ID: ming_qing_archives_100
啟用
number
明清檔案 (全資料) Ming-Qing Archives Benchmark (all)
Historical archives benchmark for Ming-Qing dynasty documents
ID: ming_qing_archives_all

平台展示 (3/4)

自動化提交與評分 (Submission)

- **拖拽上傳**：User 僅需上傳預測結果 (CSV)
- **即時回饋**：Backend Worker 自動排程計算狀態即時更新


TAIHU Benchmark
任務
規則
排行榜
提交
上傳紀錄
系統管理
我的提交
狀態
檔案名稱
說明
總體分數
提交時間
操作
暫無提交記錄
您還沒有提交任何解決方案，提交您的第一個解決方案開始吧！

平台展示 (4/4)

雙重排行榜 (Leaderboard)

- **Public vs. Private**：防止模型針對考題 Overfitting
- **多維度分析**：可依據 nDCG 分數、提交時間排序

第五節 臺灣首創：TAIDE 專案經驗 (1/2)

國內第一個任務導向 Benchmarking

- **我的角色**：
  - 在 TAIDE 開發初期，我們就意識到 MMLU 對台灣應用無效
  - 我主導設計並實作了**臺灣第一個**針對生成式 AI 的**任務導向評測框架**
- **拒絕空談**：
  - 不看 Perplexity，直接看模型能不能「上工」

第五節 臺灣首創：TAIDE 專案 (2/2)

TAIDE 五大任務評測

1. **自動摘要 (Summarization)**：
    - 長文濃縮、重點提取 (公務應用剛需)
2. **寫信能力 (Letter Writing)**：
    - 公務信函、商務往來、回應民眾陳情
3. **寫文章能力 (Article Writing)**：
    - 起承轉合、論述邏輯、擴寫大綱
4. **中翻英 (Zh-En Translation)**：
    - 精準傳達語意，符合英語慣用法
5. **英翻中 (En-Zh Translation)**：
    - 信達雅，避免翻譯腔，保留在地語感

第六節 Benchmark 實戰：公文撰寫 (1/4)

為什麼選「公文」作為標竿？

- **高難度與高剛需**：
  - 格式極度嚴格 (錯一個字都不行)
  - 語氣極度講究 (上行、平行、下行文不同)
- **意義**：
  - 這是驗證 LLM「指令遵循」與「在地化」的最佳場域
  - 這需要**公務專家**與 **Vibe Coding** 的結合

第六節 Benchmark 實戰：公文撰寫 (2/4)

Step 1: 專家知識萃取 (The "Vibe")

- **專家 (SME) 的工作**：
  - 告訴我們什麼是「好公文」
  - **顯性規則**：字號位置、分項編號 (一、(一)、1、(1))
  - **隱性規則**：「請 核示」vs「請 鑒核」、避免過度口語化
- **轉化**：
  - 我們將這些「專家直覺」轉化為**評分量表 (Rubric)**

第六節 Benchmark 實戰：公文撰寫 (3/4)

Step 2: LLM-as-a-Judge 技術導入

- **問題**：請公務員改 1000 篇公文太貴、太慢
- **解法**：用 GPT 或 Gemini Pro 扮演評審
- **Prompt 設計 (關鍵)**：
  - *「你是一位資深公文專家，請根據以下 [評分規程]，針對 [模型產出] 進行評分...」*
- **CoT (Chain of Thought)**：
  - 要求 Judge 先寫評語，再打分數，準確度提升 20%

第六節 Benchmark 實戰：公文撰寫 (4/4)

Step 3: 混合評估 (Hybrid Evaluation)

我們建議的**最佳實務**：

1. **Rule-based Filter (Vibe Coding 實作)**：
    - 用簡單程式碼篩掉格式錯誤、禁語
2. **LLM-as-a-Judge**：
    - 進行大規模內容品質評分 (語意、邏輯)
3. **Human Audit**：
    - 隨機抽樣 5% 由專家複核，校準 LLM Judge 的偏差

第七節 臺灣策略：在地化與主權 AI (1/3)

為什麼「在地化 Benchmark」是戰略物資？

- **掌握話語權**：
  - 如果我們用 MMLU，台灣模型永遠比不過國外大模型
  - **我們要定義什麼是「好」的繁體中文 AI**
- **文化護城河**：
  - **例子**：「芒果乾」(政治隱喻)、「拜票」(選舉文化)
  - 只有透過在地化測試，才能確保模型懂臺灣

第七節 臺灣策略：在地化與主權 AI (2/3)

對中小企業 (SME) 的意義

- **降低試錯成本**：
  - 中小企業沒能力自己做 RAG Benchmark
- **公共財 (Public Goods)**：
  - 釋出「TAIHU Benchmark」、「TAIDE 公文評測集」
  - 讓企業知道：*「做客服，用 Model A；寫 RAG，用 Model B」*

第七節 臺灣策略：在地化與主權 AI (3/3)

建立臺灣 AI 評測生態系

- **開放資料**：政府應釋出去識別化的公文、法規
- **人才培育**：
  - 像台大文學院一樣，培育更多懂 AI 的領域專家
  - 推廣 Vibe Coding，讓專家能自己造 Benchmark
- **平台化**：建立 Taiwan Leaderboard，不比 MMLU，比任務能力

第八節 結論

1. **看穿本質**：MMLU 只是測文字接龍，不代表工作能力。
2. **擁抱專家**：Benchmark 的靈魂在領域專家，不在工程師。
3. **善用工具**：透過 **Vibe Coding** 與 **LLM-as-a-Judge**，低成本實現大規模評測。
4. **先行經驗**：從 TAIDE 到 TAIHU Benchmark，台灣已經有能力制定自己的標準。
5. **最終目標**：建立屬於台灣的 AI 評測量尺，掌握數位主權。

學習AI，不分年齡領域，大家一起來！

Before Grade 9

Grade 7-12

博客來傳送門

AI界李白

Q & A

蔡宗翰 博士

thtsai@csie.ncu.edu.tw

粉絲團: AI界李白

IG/threads: ai.libai