虛假文獻偵測系統發展

指導老師一：汪志堅老師

指導老師二：賴正育老師

團隊成員：

711436119 資訊碩一 盧信廷

711436112 資訊碩一 黃柏叡

711436113 資訊碩一 陳文澤

目錄

01 研究背景

生成式AI的普及

近期，由大型語言模型（LLMs）驅動的生成式人工智慧（Generative AI）已被廣泛應用於學術寫作與文字精煉。學生與研究人員愈來愈依賴這些工具進行文獻蒐集與初稿撰寫。

虛假引用的出現

儘管 AI 工具功能強大，但經常會產生「虛假引用」，其表現形式包括偽造的參考文獻、真實作者與虛構文章錯誤配對與錯誤或無法解析的 DOI 編號

01 研究背景

學術傳播與誠信風險

這些虛假引用嚴重威脅學術傳播的品質與研究的可重複性。實證研究顯示，在AI輔助的系統性回顧中，引用錯誤與虛假引用的比例顯著偏高。(van Rensburg, 2025)

驗證負擔的增加

隨著虛假文獻增加，期刊編輯、審稿人及學術機構面臨巨大的查核負擔，對現有的學術 Integrity（誠信）與評鑑系統構成了嚴峻挑戰。(Haan, 2025)

02 研究動機與目的

研究目的：從技術實踐到應用評估

識別「AI 幻覺」的關鍵指標

生成式 AI 產生的「虛假引用」本質上是 AI 幻覺的產物。偵測這些虛假文獻不僅是為了糾正錯誤，更是將其視為判別論文是否含有未經人工審核之 AI 生成內容的重要篩選指標。

建構多層次自動化支援體系以提升審查效率

本研究旨在為學生、作者、導師及編輯提供一套自動化的機制，作為預防虛假文獻進入學術體系的防禦策略。該系統不僅能作為研究者的自我檢查工具，降低無意間引用AI幻覺內容的風險，更能透過自動化驗證縮小人工查核範圍。

02 研究動機與目的

研究目的：從技術實踐到應用評估

自動化驗證系統

開發一套能針對英文學術論文進行自動化參考文獻驗證的系統，並具備自動標記潛在虛假文獻的功能。

評估系統偵測效能

透過實證測試驗證該系統在偵測AI生成虛假引用時的準確性與有效性，評估其作為學術輔助工具的可行性。

探討學術審查與誠信應用潛力

從學術與行政管理的雙重視角，分析該系統在學術審閱流程中的應用潛力

03 系統架構

01 系統環境

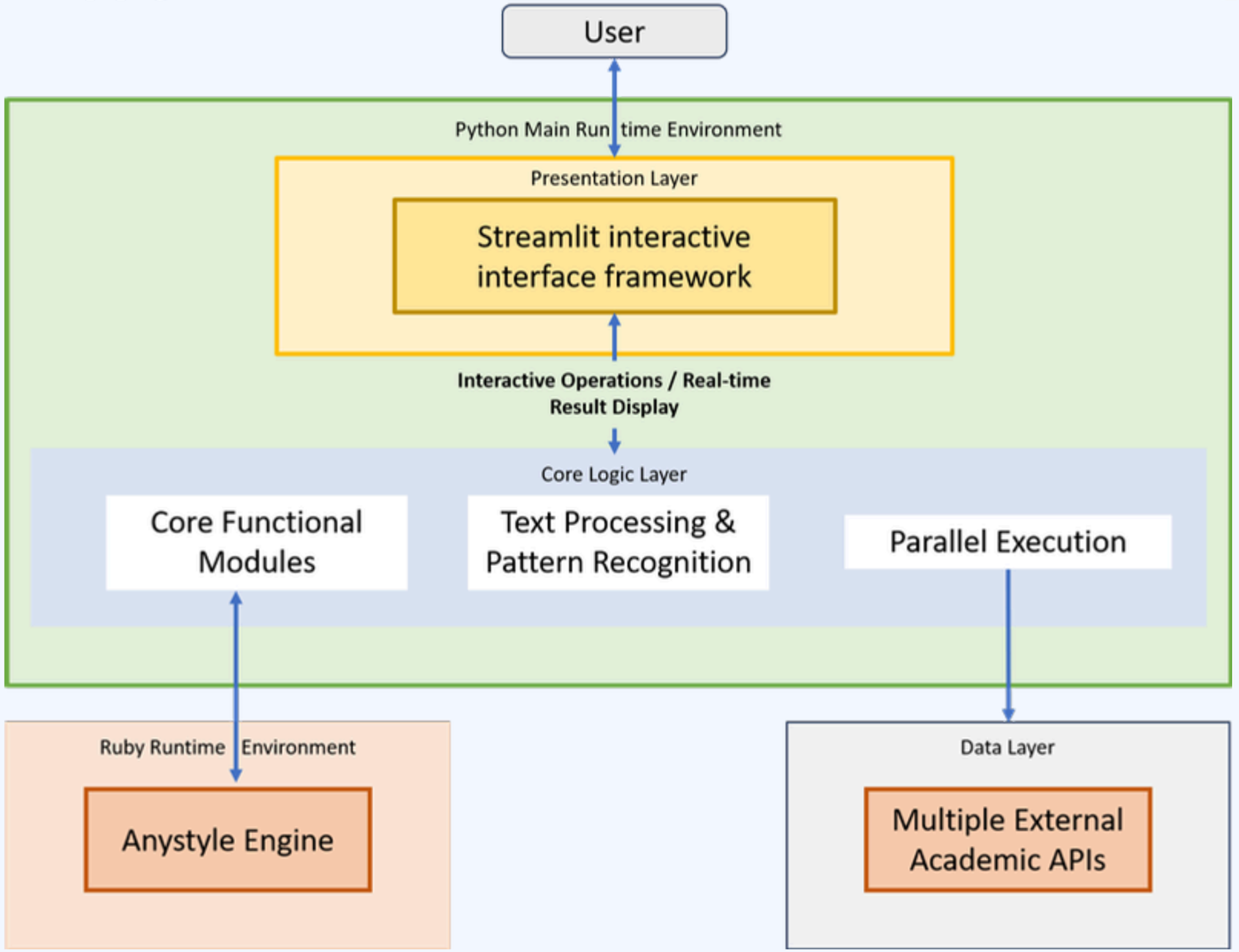


Figure 1: System Environment

03 系統架構

02 系統架構圖

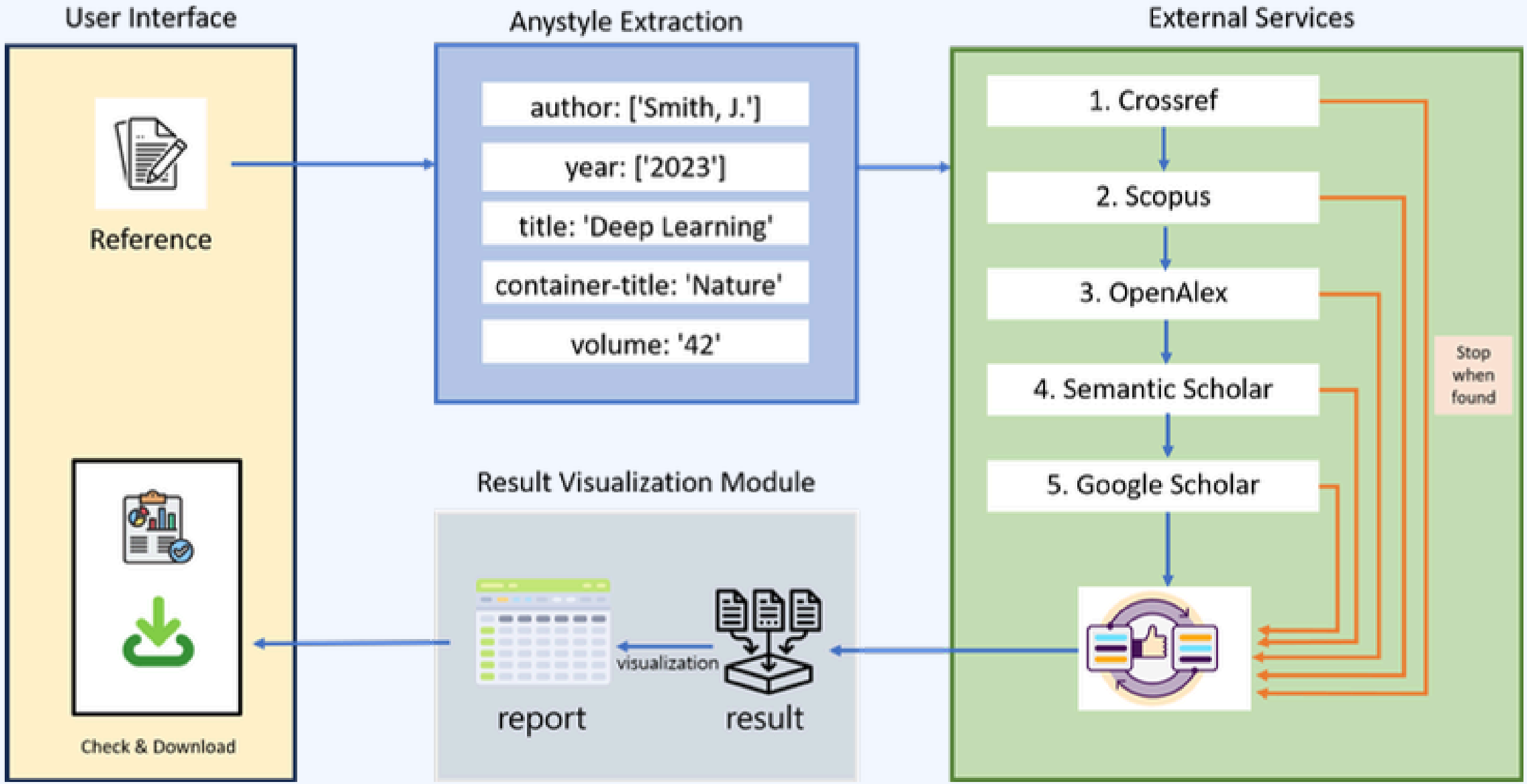


Figure 2: System Architecture

03 系統架構

- **參考文獻解析模組（REFERENCE PARSING MODULE）**
  - 核心引擎：採用高效能 ANYSTYLE 引擎，將非結構化引用文本轉換為結構化資料 。
  - 精確提取：確保作者、標題與 DOI 等關鍵元數據準確，為後續驗證奠定標準化基礎 。

- **引用驗證模組（CITATION VERIFICATION MODULE）**
  - 核心功能：執行跨資料庫的引用真實性驗證，是系統運作的核心 。
  - 多源API整合：串接 CROSSREF, OPENALEX, SEMANTIC SCHOLAR, SCOPUS 及 GOOGLE SCHOLAR 。
  - 比對邏輯：依據引用標題與作者資訊查詢書目記錄並計算輸入引用與資料庫檢索條目之間的相似度分數
  - 偵測目標：有效識別不存在的出版物、作者與標題不符、以及 AI 幻覺產生的虛構文獻 。

- 結果視覺化模組 (RESULT VISUALIZATION MODULE)
  - 前端介面：利用 STREAMLIT 框架，建構直觀且使用者友善的操作環境 。
  - 資訊呈現：顯示引用驗證狀態類別並提供錯誤類型摘要報告且列出被標記為高風險參考文獻的詳細資訊
  - 應用效益：透過視覺化統計與分組，協助使用者快速評估整體引用品質並讓審查者聚焦於需手動複查的條目，提升學術審查與自我檢查效率。

04 實驗結果統整


NDLTD IN TAIWAN
臺灣博碩士論文知識加值系統
National Digital Library of Theses and Dissertations in Taiwan
一般民眾
研究人員
校院系所及研究生
(118.150.152.141) 您好！臺灣時間：2026/01/16 00:30
進階查詢
簡易查詢/指令查詢/智慧型選題
資訊管理
系所
and
不限欄位
and
不限欄位
新增查詢欄位 | 移除查詢欄位
Search
查詢模式：精準 模糊 同音 同義詞 漢語拼音 通用拼音
輔助檢索：簡體轉換繁體 拉丁語
縮小查詢範圍
畢業學年度(民國)： 113 至 114
學位類別： 博士 碩士
語言： 中文 英文 日文 其他語文

04 實驗結果統整

資訊管理學系

113至114之博士論文

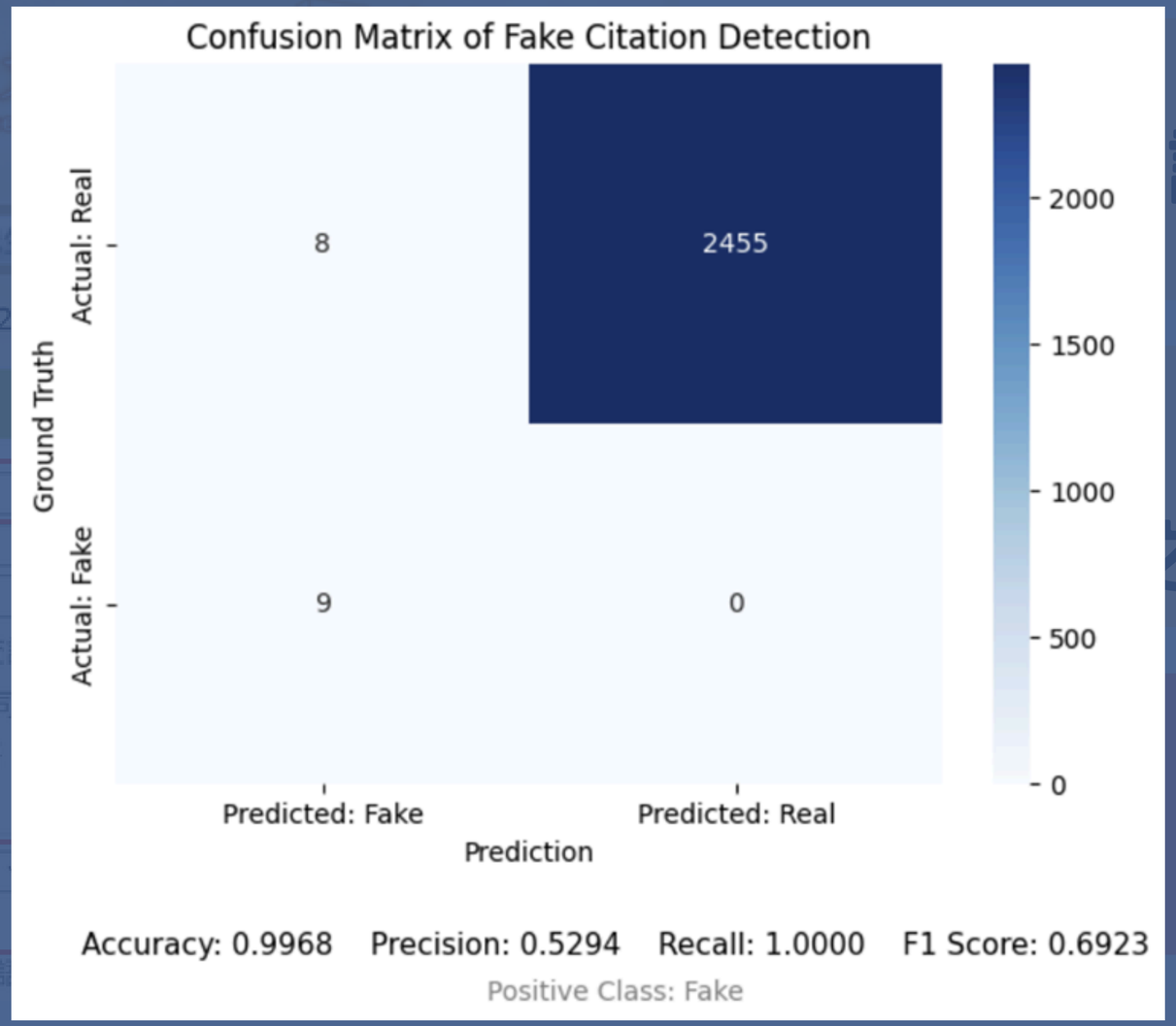
04 實驗結果統整

資訊管理學系

113至114之博士論文

共32篇

04 實驗結果統整


Confusion Matrix of Fake Citation Detection
Ground Truth
Actual: Real
Actual: Fake
8
2455
9
0
2000
1500
1000
500
0
Predicted: Fake
Predicted: Real
Prediction
Accuracy: 0.9968 Precision: 0.5294 Recall: 1.0000 F1 Score: 0.6923
Positive Class: Fake

04 實驗結果統整

	Predicted: Fake	Predicted: Real
Actual: Fake	9	0
Actual: Real	8	2455

04 實驗結果統整

	Predicted: Fake	Predicted: Real
Actual: Fake	9	0
Actual: Real	8	2455

Accuracy ≈ 99.68%

Precision ≈ 52.94%

Recall = 100%

F1 Score ≈ 69.23%

05 討論與結論-研究成果效能

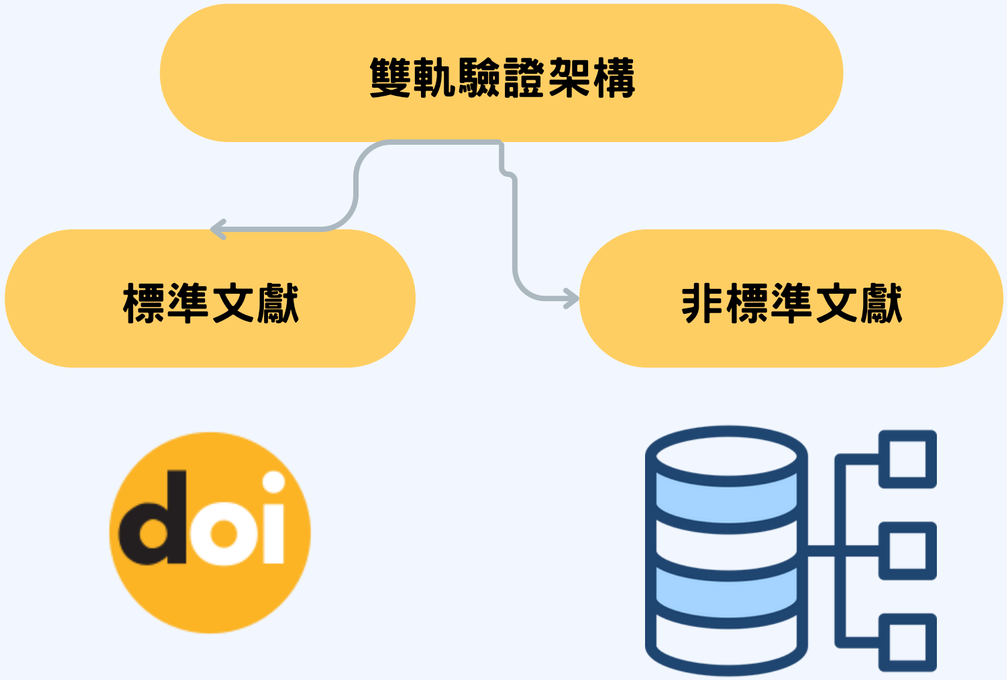
高準確度與強大的攔截能力

- 系統實證評估顯示，分類準確度（Accuracy）高達 99.68%。
- 召回率（Recall）達到 1.0000，代表測試數據集中的所有「虛假文獻」皆被成功偵測。

以「誠信為本」的設計取向

- 精確度（Precision）為 0.5294，反映出系統採取保守的偵測策略，優先將模糊個案標記為待查。
- 透過自動化標記高風險引用，讓審閱者能集中精力於關鍵覆核。

05 討論與結論-核心技術貢獻



應對期刊、會議論文、技術報告及網頁資源等學術文獻的**多元異質性**，提供全面的驗證管道。

05 討論與結論-未來展望與應用整合

01

推動學術流程體制化

將自動化預審機制嵌入**期刊投稿平台**或**機構審查系統**，建立事前防禦機制。

02

技術擴展與升級

探索跨語言知識圖譜(Cross-lingual Knowledge Graph)的整合，提升對全球學術資源的覆蓋力。

03

維護數位時代的學術誠信

透過技術工具的制度化，建立更具韌性的學術工作流，有效應對生成式 AI 對學術誠信帶來的長期挑戰。

THANKS!