文字探勘期末專案 - 第 12 組

Prompt 工程對 AI 文本偵測的影響：傳統與深度學習模型的比較分析

B13705003 謝欣佑、B13705007 林采穎、B13705027 黃柏琁、B13705055 陳沐頤、B13608027 吳昕醍

1. 專案背景與研究問題

近年生成式 AI（如 ChatGPT、Gemini）文本的普及，對教育與研究倫理帶來挑戰，但現有 AI 文本偵測研究大多注重模型分類的精度，卻忽略了提示工程（Prompt Engineering）對文本生成風格與模型可偵測性的影響，因此我們想探討以下兩個研究問題：

i. Prompt 類型是否會系統性的改變 AI 文本的語言特徵，進而影響其被模型偵測的難易度？
ii. 不同原理的偵測模型在面對刻意設計的 Prompt 文本時，其性能的穩健性差異為何？

並提出核心假設如下：

i. 模仿自然語氣（如敍述式 Prompt）的 AI 文本將最難被偵測。
ii. 深度學習模型（BERT/LoRA）對抗 Prompt 類型變化的穩健性將優於傳統模型 （XGBoost / SVM）。

2. 研究方法

我們預計採取「資料生成」、「模型優化訓練」與「穩健性實驗」三階段進行研究。

2.1 資料來源與實驗變因（Prompt 類型）

資料類型	目的
基準標註資料	訓練與優化兩種基準偵測器
自建實驗組資料	核心實驗，測試偵測難易度
純人類文本	語言特徵對照組

Prompt 類型	預期風格
A. 指令式	條理分明、易偵測
B. 敍述式	自然、口語化、難偵測
C. 設定角色式	風格化、具情感色彩

2.2 實驗步驟與模型訓練

步驟	內容與核心技術
i. 準備與特徵工程	使用 TF-IDF 進行傳統模型特徵工程，並準備 BERT Tokenization。
ii. 傳統模型訓練	訓練 XGBoost/SVM，使用 Grid Search 進行參數優化，獲得最佳性能。
iii. 深度模型訓練	使用 LoRA 微調 BERT/RoBERTa（learning_rate 等參數）獲得最佳性能。
iv. 核心穩健性測試	將三類 Prompt 文本輸入兩種最佳模型，記錄其「被偵測為 AI」的準確率。

3. 預期成果展現

i. 兩種模型（傳統與深度學習）的訓練與參數優化結果。
ii. 繪製長條圖，量化比較兩模型對三類 Prompt 文本的偵測準確率，驗證模型穩健性。
iii. 基於 LSA 降維與聚類分析（K-means）的語言特徵視覺化圖表。