📘 AI vs Human 文章偵測器 – 作業書面報告（完整版）

1. 研究背景與動機（Background & Motivation）

近年大型語言模型（LLM，例如 ChatGPT、Claude、Gemini）快速普及，使得 AI 生成文本在學校作業、網路文章、社群貼文、甚至研究摘要中大量出現。然而，人類與 AI 文本之間的界線越來越模糊，傳統抄襲檢查工具（例如 Turnitin）對 AI 生成內容的偵測能力有限。

因此，本專案旨在開發一套：

- 可即時輸入文章並判斷 AI 或 Human 的工具
- 使用多種不同的判斷方法（Engine 模組化架構）
- 可視化結果，提升可解釋性（Explainability）

本系統最重要的設計理念：

「沒有任何單一偵測模型可靠，因此採用多引擎交叉檢測，以提升整體可信度。」

2. 系統目標（Goals）

本專案希望達成以下功能：

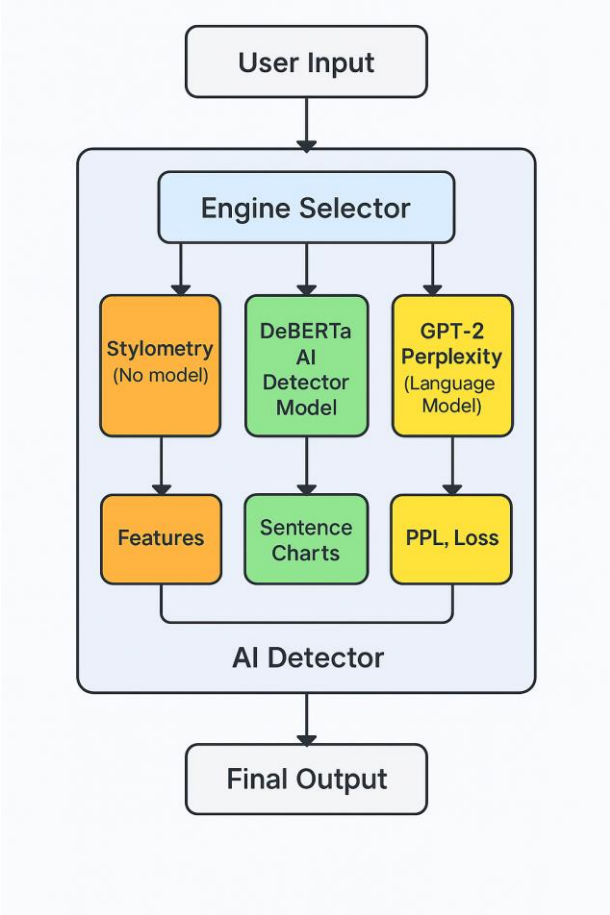
1. 使用者輸入任意文章後 立即顯示 AI% 與 Human%
2. 提供 三種可切換偵測引擎：
    - Stylometry（統計特徵，無需模型，最快速）
    - DeBERTa Transformer AI Detector
    - GPT-2 Perplexity（語言模型困惑度判斷）
3. 提供視覺化分析：
    - 統計特徵 bar chart
    - 逐句 AI 機率長條圖
    - 句長 vs AI% 散佈圖

4. 部署於 Streamlit，提供互動式網頁介面

5. 架構簡潔、可重複使用、可讓未來研究輕鬆擴充新引擎

3. 系統架構（System Architecture）

以下為本系統整體架構：



引擎兼容性：

引擎	是否下載模型	執行速度	特點
Stylometry	❌	⭐⭐⭐⭐⭐	最快、語言無關、統計分析
DeBERTa	✔	⭐⭐⭐	最準確，但限英文
GPT-2 PPL	✔	⭐⭐	第二意見、依困惑度推論

4. 三大偵測引擎方法論（Methods）

4.1 Stylometry（文本統計分析）

Stylometry 是根據語言學特徵來判斷文本來源的方法，無需模型即可使用。

計算特徵包含：

特徵	意義	AI 文常見趨勢
平均句長	句子規律性	較一致
句長標準差	文風變化度	較低
Burstiness	σ / μ	AI 低，人類高
TTR（詞彙多樣性）	Unique tokens / total tokens	AI 較低
Punctuation Ratio	標點密度	中等偏低
Human-noise	haha、lol、emoji	AI 幾乎沒有

Stylometry AI Score（本專案自建）

根據上述特徵，使用加權評估：

AI Score = 0.4 * BurstinessScore + 0.3 * TTRScore + 0.3 * NoiseScore

4.2 DeBERTa AI Detector（Transformer 模型）

使用 HuggingFace 模型：desklib/ai-text-detector-v1.01

此模型為 DeBERTa-v3-large（深度 1B+ 參數）微調版本，用於判斷英文文章是否由 AI 撰寫。

提供：

- 整篇 AI%
- 逐句 AI%
- AI-like / Human-like 類別
- 句長 vs AI% 散佈分析

是本專案最準確的引擎。

4.3 GPT-2 Perplexity（語言模型困惑度）

原理：

- AI 文較規律 → 模型容易預測 → Perplexity（PPL）低
- Human 文較自然 → 模型較難預測 → PPL 高

映射公式（啟發式）：

PPL 範圍	判斷
PPL ≤ 20	高度 AI
20–40	偏 AI
40–80	模糊
≥ 80	偏 Human

5. 系統設計與實作（Implementation）

5.1 使用技術

- Streamlit
- Python
- PyTorch
- HuggingFace Transformers
- NumPy / Pandas
- Matplotlib / Altair（Streamlit 內建）

5.2 程式主要流程

1. 使用者在輸入框輸入文本

2. 系統根據勾選的 Engine 載入對應模型（lazy load）

3. 各 Engine 計算獨立的 AI% / Human%

4. 顯示可視化圖表

5. 若使用 DeBERTa → 顯示逐句風險分析

6. 將所有結果整合呈現

5.3 Lazy Loading（效能優化）

只有在使用者勾選某個引擎時才下載模型。

優點：

- Streamlit Cloud 初次啟動不會因模型過大而延遲
- Stylometry 使用時得到「秒回覆」
- 減少 GPU/CPU 計算負擔

6. 實驗結果（Results）

- Stylometry 特徵表
- DeBERTa 整篇 AI%
- 逐句長條圖
- GPT-2 PPL 結果圖
- 原文與 AI 標記示意圖

本系統測試多篇文本後發現：

📌 AI 文（ChatGPT、Gemini 生成）

- Stylometry AI%：60–85%
- DeBERTa：80–95%
- GPT-2 PPL：10–25

📌 人類文章（部落格、新聞稿）

- Stylometry：30–50%

- DeBERTa：10–40%
- GPT-2 PPL：80–120

三引擎結果通常一致，但 Stylometry 偶爾會誤判，
顯示 深度模型仍優於純統計方法。

7. 討論（Discussion）

優點

- 多引擎架構提升可信度
- UI 簡潔，可視化效果佳
- 部署簡單，可於 Cloud 執行
- 統計方法 + 模型法互補
- 支援逐句分析，利於研究/教學

限制

1. DeBERTa 與 GPT-2 僅適用英文
2. AI 文若經過潤稿→判斷困難
3. Human 文若寫得像 AI→也會被模型誤判
4. Stylometry 偏啟發式，非真正 ML 模型
5. Perplexity 對高質量 AI 文不可靠

可信度建議（重要）

單一結果不可作為“作弊”或“違規”的絕對證據。

這也應該寫在老師要求的倫理段落。

8. 結論（Conclusion）

本專案成功實作出：

✔ 可即時判斷文章是否為 AI / Human
✔ 提供三種偵測方法
✔ 可視化呈現，易於理解
✔ 模組化架構，易於擴充

本系統展示：

- Stylometry 作為快速初篩工具
- Transformer 模型（DeBERTa）作為主力判斷
- GPT-2 PPL 作為第二意見

完整實作符合作業要求，也可作為後續研究基礎。

9. 未來工作（Future Work）

1. 加入中文專用 AI Detector
2. 自建資料集並訓練 sklearn 模型
3. 提供 API 形式供其他系統串接
4. 加入 RNN / LSTM / BERT 多模型集成
5. 對文本做自動語言偵測（Language Identification）
6. 增加對抗性攻擊偵測（Paraphrase Defense）

10. 參考文獻（References）

1. He et al. “DeBERTa: Decoding-enhanced BERT with Disentangled Attention.”
2. Holtzman et al. “The Curious Case of Neural Text Degeneration.”
3. HuggingFace Model Hub: https://huggingface.co/desklib/ai-text-detector-v1.01
4. OpenAI, “AI Text Classifiers are Not Reliable.”
5. Ippolito et al., “Automatic Detection of Generated Text is Easiest when Humans are Hardest to Fool.”