Communications of the CCISA
Vol. 31 No. 3 Aug. 2025

基於大型語言模型之網路釣魚郵件偵測研究

蔡尚恩**、郭俊安[2]、孫偉誠[3]
[13]長榮大學資訊工程系、[2]台北科技大學資訊安全碩士學位學程
[1] sean@mail.cjcu.edu.tw、[2] t114c75018@ntut.edu.tw、[3]sweichen1014@gmail.com

摘要

隨著人工智慧技術的發展，利用大型語言模型 (LLMs) 生成的高度逼真網路釣魚郵件已成為嚴峻的資安挑戰。為應對此威脅，本研究提出並實作一個基於大型語言模型的網路釣魚郵件偵測系統。本研究的核心方法在於採用一種資源效率更佳的「提示詞工程」(Prompt Engineering) 策略，在不進行模型微調 (Fine-Tuning) 的前提下，透過零樣本 (Zero-shot) 學習來發揮模型的內在分類能力。我們從郵件中提取文本、網址、圖片光學字元辨識 (OCR) 及附件檔名等多維度特徵，建構結構化提示詞以引導模型進行判斷。本研究特別針對兩種主流的開源輕量級模型—Mistral-7B 與 LLaMA3-8B—進行了實證比較。實驗結果顯示，Mistral-7B 在準確率、精確率、召回率與 F1 分數上均顯著優於 LLaMA3-8B。我們進一步分析，此性能差異可能源於 Mistral-7B 為效率而生的精簡架構，使其在目標明確的分類任務中表現更為穩定；相較之下，LLaMA3-8B 的架構複雜性與龐大詞彙庫，在零樣本情境下可能導致對特徵的過度解讀，從而產生較高的誤報率。本研究不僅驗證了利用提示詞工程偵測釣魚郵件的可行性，也為特定資安應用場景下的模型選擇提供了具體的實證依據與洞見。

關鍵字：網路釣魚郵件、大型語言模型、提示詞工程、資訊安全、Mistral-7B、LLaMA3-8B

*通訊作者 (Corresponding author.)

Regular Paper

Communications of the CCISA
Vol. 31 No. 3 Aug. 2025

Phishing Email Detection Based on Large Language Models

Tsai, Shang-En[1*], Kuo, Chun-An[2], Sun, Wei-cheng[3]
[13] Department of Computer Science and Information Engineering, Chang Jung Christian University, [2]Taipei University of Technology, Master's Program of Information Security
[1] sean@mail.cjcu.edu.tw, [2] t114c75018@ntut.edu.tw, [3]sweichen1014@gmail.com

Abstract

As artificial intelligence technology advances, highly realistic phishing emails generated by Large Language Models (LLMs) have become a severe cybersecurity challenge. To counter this threat, this study proposes and implements a phishing email detection system based on LLMs. The core methodology of this research lies in adopting a resource-efficient "Prompt Engineering" strategy, leveraging the models' intrinsic classification capabilities through zero-shot learning without the need for fine-tuning. We extract multi-dimensional features from emails—including text, URLs, Optical Character Recognition (OCR) from images, and attachment filenames—to construct structured prompts that guide the model's judgment. This study specifically conducts a comparative performance analysis of two prominent open-source, lightweight models: Mistral-7B and LLaMA3-8B. Experimental results demonstrate that Mistral-7B significantly outperforms LLaMA3-8B across all metrics, including accuracy, precision, recall, and F1-score. We further analyze that this performance disparity may stem from Mistral-7B's streamlined, efficiency-oriented architecture, which allows for more stable performance in well-defined classification tasks. In contrast, the architectural complexity and larger vocabulary of LLaMA3-8B might lead to an over-interpretation of features in a zero-shot context, resulting in a higher false positive rate. This research not only validates the feasibility of using prompt engineering for phishing detection but also provides concrete empirical evidence and insights for model selection in specific cybersecurity application scenarios.

Keywords: Phishing Email, Large Language Models (LLMs), Prompt Engineering, Cybersecurity, Mistral-7B, LLaMA3-8B

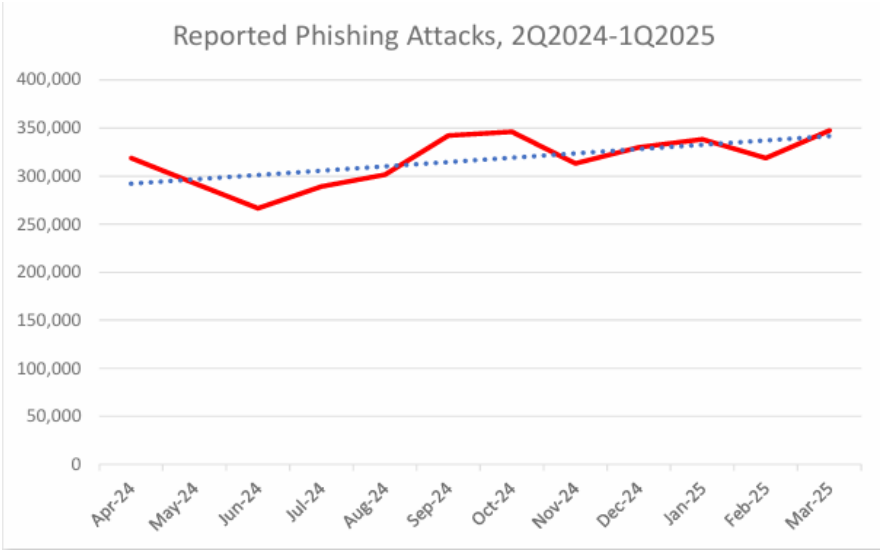
壹、 研究動機與研究問題

自西元 2020 年新冠肺炎爆發以來，全球各地陸續採取遠端工作，顯著提升了網路使用量，而電子郵件更成為重要的聯絡與溝通管道之一。根據 Radicati 最新發布的年度電子郵件統計報告 [10]，至 2024 年全球電子郵件使用者已達 44.81 億人，預計至 2028 年底將進一步增長至 49.7 億人，等同於全球超過一半人口皆為使用者。同年度，每日發送與接收的電子郵件總量已突破 3616 億封，並預估將以年均成長率約 4%的速度增長，將於 2028 年達到 4242 億封，見圖一。

	2024	2025	2026	2027	2028
Worldwide Email Users* (M)	4,481	4,594	4,730	4,849	4,970
% Growth	3%	3%	3%	3%	3%

圖一：2024 至 2028 年全球電子郵件使用者預測 [10]。

隨著電子郵件使用者數持續上升，網路釣魚攻擊事件也逐年攀升。根據 Anti-Phishing Working Group (APWG) 發布的最新報告 [2]，2025 年第一季共觀察到 1,070,000 起釣魚攻擊事件，高於 2024 年第三季的 932,923 起與第四季的 989,123 起，顯示電子郵件安全威脅仍不容忽視，參見圖二。



圖二：2024 年第二季度至 2025 第一季度網路釣魚攻擊趨勢

近幾年，網路釣魚的攻擊手法不斷的進步，也隨著人工智慧 (Artificial Intelligence, AI) 的興起，開始有許多企業與研究員開始嘗試利用了機器學習(Machine Learning, ML)

和深度學習 (Deep Learning) 的技術，開發能夠偵測網路釣魚電子郵件的系統。能夠藉由人工智慧協助我們企業單位避免受到網路釣魚電子郵件的攻擊，這也表示攻擊者亦能夠利用人工智慧來幫助他們製作出可信度、真實度更高的網路釣魚電子郵件。

根據牛津大學在 2023 年所發表之研究指出 [12]，隨著大型語言模型快速演進，攻擊者已能透過 GPT-3.5 與 GPT-4 等模型，系統性地生成語氣自然、內容精準的魚叉式網路釣魚郵件。該研究以 600 多位英國國會議員為樣本，透過公開資料蒐集個人背景，再結合提示詞設計 (Prompt Engineering)，由 LLM 自動撰寫具針對性的詐騙信件，每封僅需數秒生成且具有高度說服力，顯示釣魚攻擊已走向低成本、大規模化。此研究亦證實透過基本提示詞繞過技術 (Jailbreaking)，可突破 LLM 的安全限制 (如 Reinforcement Learning with Human Feedback)，誘導模型產出完整釣魚內容與惡意附件。此結果突顯生成式 AI 已成為攻擊者進行社交工程的高效資安工具，具備高度實務威脅。

Roy 等人於 2024 年所提出的 PhishBots 框架指出 [13]，網路釣魚郵件的生成與偵測正進入一場持續升級的攻防演化。該研究設計了一個多代理人互動架構，讓攻擊者代理人 (Attack Bots) 可透過提示詞工程 (Prompt Engineering) 動態調整輸入，生成更具欺騙性的釣魚信件；而防禦者代理人 (Defense Bots) 則透過語言模型進行判斷與攔截，模擬雙方策略對抗的真實過程。結果發現，攻擊方持續優化提示詞後，能顯著降低被偵測的機率，顯示即便是最先進的偵測模型，仍難以全面掌握語言攻擊的邏輯與風格變化。

本研究將使用數種 LLMs，並針對電子郵件內容進行分析，在不對模型進行微調 (Fine-Tuning) 的情況下，結合零樣本 (Zero-shot) 策略，在開源、輕量級的大型語言模型上實現對網路釣魚郵件的有效偵測。方法上，我們聚焦於郵件中具有代表性的潛在特徵—包括郵件內容本身、是否存在短網址與跳轉行為、圖片中辨識出的文字資訊 (Optical Character Recognition, OCR)，以及附件的副檔名，將這些特徵彙整為 prompt 作為模型輸入，讓模型在無需進行額外訓練的情況下進行判斷。最終，我們統整所蒐集的資料集，並針對所採用的模型進行實驗與成效比較，以驗證此方法的可行性與表現。

貳、文獻回顧與探討

大型語言模型 (如 ChatGPT) 的興起在程式碼生成與文本合成等任務中展現巨大潛力，甚至可從公開資訊中生成高度可信的網路釣魚郵件。新加坡政府技術局於 Black Hat USA 2021 的實驗顯示 [4]，其資安團隊向內部人員發送模擬魚叉式網路釣魚電子郵件 (Spear Phishing Email)，其中包含了由人工製作以及由 OpenAI GPT-3 的技術所生成的。其結果顯示，點擊由 AI 生成釣魚郵件連結的人數遠高於點擊人工製作郵件連結的人數。

目前已有初步研究嘗試利用大型語言模型 (LLMs) 來偵測網路釣魚內容，以應對由於濫用 LLMs 導致的攻擊增加趨勢，突顯出採用 LLMs 進行自動化偵測之重要性。例如，Koide 等人 (2023) 提出的 ChatPhishDetector 系統 [8]，透過網頁爬蟲取得網站資訊，並將這些資料轉換為提示詞 (prompt) 提供 LLMs 進行判斷與分析。實驗結果顯示，該系

統使用 GPT-4V 模型，在精確率 (Precision) 與召回率 (Recall) 分別達到 98.7%與 99.6%，證明了 LLMs 在網路釣魚內容偵測方面具備顯著的應用潛力與價值，見表一。

表一：模型表現的比較。URL、HTML 和 Image 表示作為輸入的資料類型。Phishing 和 Non-phishing 表示分類目標。

System	Mode	Model	Precision	Recall	Accuracy	F-measure	URL	HTML	Image	Phishing	Non-phishing
ChatPhishDetector	Vision	GPT-4V	**98.7%**	**99.6%**	**99.2%**	**99.2%**	✓	✓	✓	✓	✓
		Gemini Pro Vision	78.9%	99.1%	89.1%	87.9%	✓	✓	✓	✓	✓
	Normal	GPT-4	98.3%	98.4%	98.4%	98.4%	✓	✓	✓	✓	✓
		GPT-3.5	98.3%	86.7%	92.6%	92.1%	✓	✓	✓	✓	✓
		Llama-2-70B	78.4%	66.4%	74.1%	71.9%	✓	✓	✓	✓	✓
		Gemini Pro	90.5%	95.6%	93.2%	93.0%	✓	✓	✓	✓	✓
	Simple	GPT-4	98.4%	75.5%	87.2%	85.5%	✓			✓	✓
		GPT-3.5	98.6%	77.5%	88.2%	86.8%	✓			✓	✓
dnstwist [4]	-	-	31.3%	-	-	-	✓			✓	
Phishpedia [28]	-	-	26.0%	-	-	-			✓	✓	

近期的研究也開始採取新的多代理 (Multi-Agent) 策略來提升 LLMs 的網路釣魚網站偵測能力。Li 等人 (2025) 提出了 PhishDebate 系統 [6]，透過 URL 分析、HTML 結構、語意內容與品牌仿冒四個專業代理，結合協調者 (Moderator) 與裁決者 (Judge) 進行多輪辯論與評估。該系統透過代理間的多視角辯論降低模型幻覺 (hallucination) 風險，並提高解釋性與準確性。實驗證實，PhishDebate 在真實網路釣魚資料集中，召回率高達 98.2%，整體效能顯著優於單一 LLM 及傳統的鏈式思考 (Chain-of-Thought) 方法。

此外，Lee 等人 (2024) 則探討了多模態 LLMs (Multimodal LLMs) [5] 應用於品牌仿冒網站偵測的可能性。該研究建構一套兩階段偵測架構：首先利用多模態 LLMs 同時分析網站螢幕截圖與 HTML 文字進行品牌辨識；再以第二階段模型進行語意化的網址品牌比對。此法無需事先訓練與維護品牌資料庫，即能有效偵測新型仿冒網站，實驗表明此法能維持 0.90 以上的 F1-score，並對現有的視覺攻擊具有高度的抵抗性，展現出比傳統視覺偵測模型更強的適應力與實用性。

參、研究方法與步驟

3.1 本研究選用模型介紹

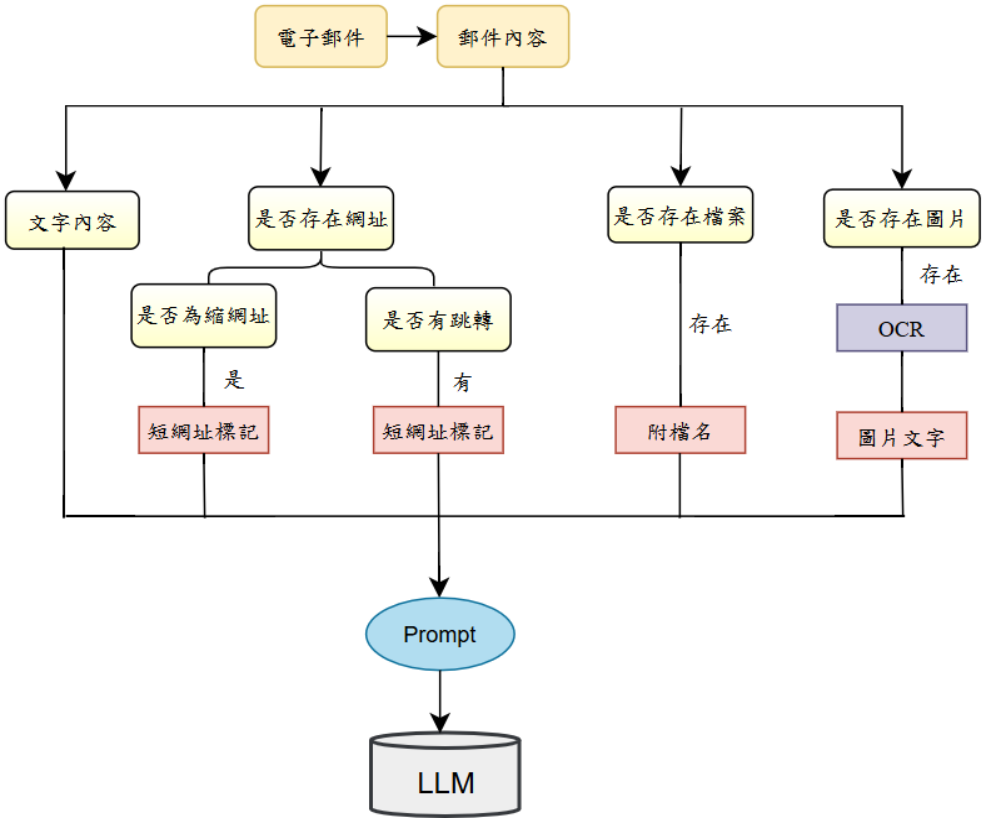
Large Language Model Meta AI (LLaMA) 是 Meta AI 公司發布的大型語言模型，並訓練了各種模型，這些模型參數從 70 億到 650 億不等。在 LLaMA 推出的半年過後，LLaMA2 也跟著問世，其在模型架構方面與 LLaMA1 保持不變，但訓練模型的數據增加了 40%，還有一個不同的地方是，LLaMA2 的所有模型都附帶權重，並且都是免費的。Mistral-7B 是由一間法國科技公司 Mistral AI 所開發的，這間公司也是因應 OpenAI 閉源模式的不同理念而誕生的新創公司；而 Mistral-7B 是一個高效能的基礎模型，擁有 70 億參數，專為在低資源環境中運行而設計，同時，該公司也在西元 2024 年初發表了

Mixtral of Experts，在本文中也提到了 Mistral-7B、Mixral-8x7B 與 LLaMA 各個參數量級的模型進行所有基準測試，表二為基準測試的結果表現。

Model	Active Params	MMLU	HellaS	WinoG	PIQA	Arc-e	Arc-c	NQ	TriQA	HumanE	MBPP	Math	GSM8K
LLaMA 2 7B	7B	44.4%	77.1%	69.5%	77.9%	68.7%	43.2%	17.5%	56.6%	11.6%	26.1%	3.9%	16.0%
LLaMA 2 13B	13B	55.6%	80.7%	72.9%	80.8%	75.2%	48.8%	16.7%	64.0%	18.9%	35.4%	6.0%	34.3%
LLaMA 1 33B	33B	56.8%	83.7%	76.2%	82.2%	79.6%	54.4%	24.1%	68.5%	25.0%	40.9%	8.4%	44.1%
LLaMA 2 70B	70B	69.9%	**85.4%**	**80.4%**	82.6%	79.9%	56.5%	25.4%	**73.0%**	29.3%	49.8%	13.8%	69.6%
Mistral 7B	7B	62.5%	81.0%	74.2%	82.2%	80.5%	54.9%	23.2%	62.5%	26.2%	50.2%	12.7%	50.0%
Mixtral 8x7B	12B	**70.6%**	84.4%	77.2%	**83.6%**	**83.1%**	**59.7%**	**30.6%**	71.5%	**40.2%**	**60.7%**	**28.4%**	**74.4%**

表二：基準測試的結果表現 [1]

考量上述 Mistral AI 與 LLaMA 各個模型基準測試的結果表現，本研究將採用 Mistral-7B 和 LLaMA3-8B 作為模型。圖三為本研究之流程圖，首先對電子郵件資料集進行分析，從電子郵件內容中分別取出寄件者、收件者、文字內容、網址（若存在）、圖片（若存在）和檔案（若存在）之副檔名，並將取出的資料進行預處理，接著將這些資料整理後製作成 Prompt，最後讓 LLMs 讀取 Prompt，偵測資料集並輸出結果，並根據結果的對錯計算出模型評估指標，挑選表現較好之作為系統使用之模型。



圖三：本研究之流程圖

3.2 研究步驟

本研究將研究步驟分為「蒐集資料—特徵截取-Prompt 指示—模型評估—建置網頁」，詳細步驟如下:

3.2.1 蒐集資料

資料集的選擇對於深度學習研究至關重要，高品質且多樣化的資料能有效提升模型偵測能力。針對網路釣魚電子郵件的偵測任務，應特別考量資料集的數量與多樣性，足夠且多元的樣本不僅能強化模型泛化能力，也能幫助模型辨識各種型態的網路釣魚攻擊，進而提升偵測準確性與實務效益。

3.2.2 特徵截取

截取能夠有效分辨網路釣魚電子郵件和一般電子郵件的特徵是非常重要的一件事，對於特徵的處理與說明如下：

- 文字內容：撰寫程式碼將郵件內文字內容取出，作為大型語言模型，能夠更 好的了解內容是否與網路釣魚電子郵件有關聯。
- 網址：檢查郵件內是否存在網址，若存在，則使用 Python 現有的 requests 套件，進行短網址以及是否有跳轉動作的判斷。
- 圖片文字：檢查郵件內是否存在圖片，若存在，則使用影像處理技術，例如： OCR，將圖片內文字取出。一個網路釣魚電子郵件內的圖片也極有可能包含相關的資訊。
- 檔案之副檔名：檢查郵件內是否存在檔案，若存在，將對其副檔名進行分析。據 ASRC 垃圾訊息研究中心在民國 110 年 8 月發表的第二季電子郵件安全觀察中 [3] 提到，有出現不少的惡意檔案都帶有雙重副檔名，然而電腦對於這種檔案的判讀卻是以最後一個副檔名為主。
- Prompt 指示:對於一個大型語言模型，會對其預訓練模型進行微調，讓它可以執行特定的任務，但這也需要耗費大量的時間與電腦資源。因此我們將資料集事先進行分析，並將特徵作為供給系統的 Prompt 文字指示，讓 LLMs 理解這些資料，就能避免時間與資源的大量耗費。

以下為 Prompt 模板介紹:

- 角色定義：明確告訴模型現在要以什麼身分來執行接下來的任務。

```
You are a cybersecurity analyst specializing in email security.
Your task is to determine whether a given email is a phishing email or legitimate.
```

圖四：Prompt 角色定義

- 任務：指定模型要執行的任務，這裡包括分析電子郵件寄件人、收件人、主旨以及內容等，同時也要去辨識寄件人電子信箱的域名是否正常。
- 資料特徵：電子郵件中較有可能辨識出網路釣魚的元素。
- 輸出格式：指定模型輸出結果的格式，方便觀察結果並進行統計。
- 模型評估

```
Analyze:
Review the subject, sender address, headers, body, links, and attachments.
Identify suspicious signs like urgent language, spoofed addresses, misleading URLs, or malicious files.
Identify Brand:
If th                                                                    ltimate domain.
      Email data:

      Sender: {email_data['sender'] if email_data['sender'] else None}
      Receiver: {email_data['receiver'] if email_data['receiver'] else None}          (1)
      Subject: {email_data['subject'] if email_data['subject'] else None}
      Body: {email_data['body'] if email_data['body'] else None}
      Links' info (url, is redirect or not):                                          (2)
      {'\n'.join(email_data['links info']) if email_data['links info'] else None}
      images' text (text):
      {'\n'.join(email_data['images text']) if email_data['images text'] else None}   (3)
Conclude:
Decide if the email is phishing or legitimate.
Clearly explain your reasoning, highlighting key suspicious elements. If uncertain, label as "unknown."

Submit Findings:
Provide your findings in JSON format with the following keys:

phishing_score: An integer indicating the phishing risk on a scale of 0 to 10.
brand: The identified brand name, or "None" if not applicable.
phishing: A boolean value indicating whether the email is a phishing email (true) or a legitimate email (false).
reasoning: A detailed explanation describing the key factors and evidence that led to your conclusion.
```

圖七：Prompt 輸出格式

模型評估是一種個用來檢視模型表現的方法，其中，混淆矩陣 (Confusion Matrix) 以模型得出的預測結果與正解相互對應分成四類，包括 True Positive (TP)、False Positive (FP)、False Negative (FN) 和 True Negative (TN)，接著使用這些元素計算出模型評估指標 (Model Evaluation Index) [7] 中的各項指標如下：

$$\text{Precision} = \frac{TP}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- 建置網頁

考量到目前有多種作業系統，包括 Windows、MacOS、Android 等，因此要針對每一種作業系統進行開發會非常耗費時間，然而，將系統以網頁的方式呈現，不只能解決跨平台的問題，大眾使用上也會方便許多。

本研究之系統前端將會使用超文本標記語言 (HyperText Markup Language, HTML)、階層式樣式表 (Cascading Style Sheets, CSS) 以及 JavaScript，後端則是使用 Python Flask 作為框架進行開發。

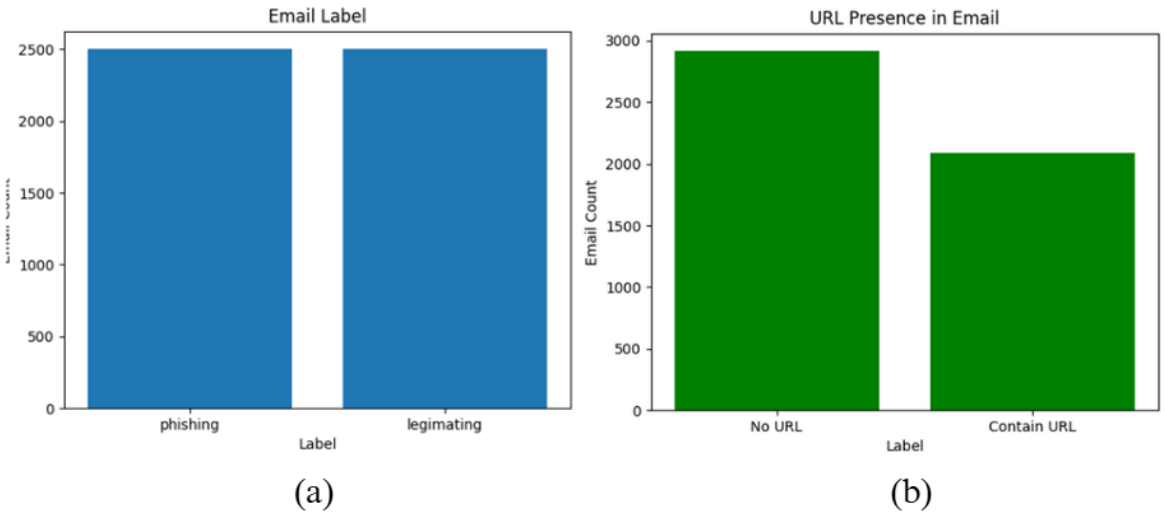
表三：本系統開發所需之工具

系統	工具
前端	HTML、CSS、JavaScript
後端	Python Flask

肆、成果實作

4.1 模型成果評估

本研究蒐集了 Kaggle、Enron [9] [11] 這兩個網站上整理來自真實世界的網路釣魚電子郵件以及一般電子郵件各 2500 筆。接著對這些電子郵件進行資料預處理，包括取得寄件者、收件者、內容、提取圖片文字、檢查檔案副檔名以及網址是否會進行跳轉（見圖八），並製作成 prompt 提供給選定之模型—Mistral-7B、LLaMA3-8B，最後根據模型答對與答錯之數量計算出模型各項評估指標，表四為模型各項評估指標。



(a) (b)

圖八：資料集中釣魚信件與一般電子郵件數量比較 (a)；資料集中包含網址與不包含網址之信件數量比較 (b)

表四：為模型各項評估指標

模型	Precision	Recall	F1-Score	Accuracy
Mistral-7B	82.2%	88.9%	85.4%	84.9%
LLaMA3-8B	75.7%	85.1%	80.1%	78.9%

根據表四的模型評估指標觀察，Mistral-7B 的整體表現優於 LLaMA3-8B，在 Precision (82.2%)、Recall (88.9%)、F1-Score (85.4%) 及 Accuracy (84.9%) 各方面皆有較佳表現，而 LLaMA3-8B 雖然在 Recall (85.1%) 上表現不錯，但 Precision (75.7%) 與 Accuracy (78.9%) 相對較低，導致整體 F1-Score 僅達 80.1%，由於本研究採用的是零樣本提示 (Zero-Shot Prompting) 方法，其效能會高度依賴模型在預訓練階段到所習得之泛化能力的精確性，因此模型的架構的差異將成為影響結果的關鍵因素。

4.1.1 Mistral-7B

Mistral-7B 的設計目標著重於是效率與速度，採用了創新的滑動窗口注意力機制 (Sliding Window Attention, SWA) 與分組查詢注意力機制 (Grouped-Query Attention, GQA)，這使其在處理長序列時能大幅降低運算複雜度和記憶體佔用，實現更快的推論速度。此外，Mistral-7B 使用了較小的 32k 詞彙庫 (Tokenizer with 32k tokens)，整體架構精簡且高效，體現出了「用更少資源做更多事」的原則。

4.1.2 LLaMA3-8B

相較於 Mistral-7B，LLaMA3-8B 雖同樣採用了 GQA 以提升效率，但其一個顯著的架構差異是它有一個極大的 128k 詞彙庫。更大的詞彙庫意味著模型在預訓練時能夠更精細地捕捉語言的細微差別，而其 MLP（多層感知器）層的設計也更為複雜，LLaMA3-8B 的設計概念更重視深度、廣度與生成細膩度。

基於上述架構差異，在一個目標明確的零樣本分類任務中，LLaMA3-8B 的「複雜性」和「龐大詞彙庫」反而可能成為一種負擔。它可能會對提示詞中的特徵進行「過度解讀」，將正常的語言波動或 URL 結構誤判為可疑信號，從而導致更高的誤報率—這正是本研究觀察到的「過度警覺」現象。反之，Mistral-7B 更為精簡、直接的架構，可能使其在執行具體的分類指令時表現得更為強健和穩定，不易受到無關特徵的干擾。

4.2 網頁呈現

考量到目前大眾使用手機的頻率也相當高，瀏覽網站也相當方便，因此本研究開發之系統網頁使用響應式網頁設計 (Responsive Web Design, RWD)，讓使用者不只能在電腦上使用本系統，在手機及平板上也不會因為設備而影響到網頁畫面的呈現，接下來將提供電腦版以及手機、平板的操作方式，參見圖九、十。

- 取得電子郵件

由於本研究以網頁進行開發，無法直接連接使用者之 Gmail，需先將電子郵件下載下來再進行上傳，電腦版只需開啟欲下載之電子郵件，點擊右上角三點圖示，即可看到下載選項；手機、平板使用者則需使用瀏覽器開啟 Gmail，並將瀏覽器調整為電腦版網站，即可依照電腦版的下載流程將電子郵件下載下來。

圖九：電腦版 Gmail 電子郵件下載方式

圖十：手機、平板之 Gmail 電子郵件下載方式

- 系統網頁操作

本研究開發之系統，將提供前端介面讓使用者上傳電子郵件到後端伺服器進行偵測，最後再將結果回傳至前端供給使用者查看，參圖十一、十二。

圖十一：電腦版上傳（左圖）與偵測結果（右圖）頁面。

圖十二：手機版上傳（左圖）與偵測結果（右圖）

伍、結論與建議

本研究成功實作出一款基於大型語言模型的網路釣魚電子郵件偵測系統，並針對

兩個相較輕量且開源的模型—Mistral-7B、LLaMA3-8B 進行測試，最後經過模型表現評估，我們選擇 Mistral-7B 作為核心偵測模型，在系統設計的部分，使用了 RWD，確保大眾無論是在電腦、手機或是平板上皆能使用，且版面不受設備螢幕大小影響。然而，在開發後的多次使用下，不論是系統偵測表現還是網頁操作上還是有改進空間，條列如下：

- 模型表現：隨著不斷有新的開源模型推出，未來將持續關注並測試不同模型偵測網路釣魚電子郵件的表現。
- 資料集多樣化：目前測試的資料集以公開資料集為主，未來會持續擴展來自不同地區、語言以及產業的網路釣魚電子郵件，以提升模型的適應性。
- 提升使用者體驗：在系統網頁使用上，目前是需要使用者自行上傳電子郵件，對於不常使用電子產品的大眾還是會有些不便，未來會往連接 Gmail 或是其他電子郵件應用程式的方向前進。

研究結果證實，大型語言模型在網路釣魚偵測上的潛力，並透過無需微調的 Prompt 設計與特徵截取來提升偵測準確度。本研究所提出之零樣本架構與彈性 Prompt 特徵設計，未來亦可導入類似攻防模擬機制，以提升系統在真實環境下的整體防護能力，且隨著開源模型的發展、資料集的優化，以及更便捷的使用者介面，本系統可望進一步降低網路釣魚攻擊帶來的風險，並提供更可靠的電子郵件安全保障。

參考文獻

[1] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M. A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix , and W. E. Sayed. “Mixtral of experts,” *arXiv preprint* arXiv:2401.04088, 2024.

[2] *Anti-Phishing Working Group*. “Phishing Activity Trends Report – 1Q 2025,” Available: https://apwg.org/ (2025/7/26)

[3] *ASRC 垃圾訊息研究中心*, “ASRC 2021 年第二季電子郵件安全觀察,” Available: https://www.asrc-global.com/index.html (2025/7/23)

[4] E. Lim, G. Tan, K. H. Tan , and T. L. “Turing in a Box: Applying artificial intelligence as a service to targeted phishing and defending against AI-generated attacks,” *National Science Council Project Report*, presented atBlack Hat, 2021. Available: https://www.blackhat.com/ (2025/7/25).

[5] J. Lee, P. Lim, B. Hooi , and D. M. Divakaran. “Multimodal large language models for phishing webpage detection and identification,” *arXiv preprint* arXiv:2408.05941, Aug. 2024.

[6] W. Li, S. Manickam, Y. W. Chong , and S. Karuppayah. “PhishDebate: An LLM-based multi-agent framework for phishing website detection,” *arXiv preprint* arXiv:2506.15656, Jun. 2025.

[7] S. Wang. “機器學習模型評估指標-confusion matrix, precision, recall, and F1-score,” Available: https://medium.com/ (2025/8/1).

[8] T. Koide, N. Fukushi, H. Nakano , and D. Chiba, “Detecting Phishing Sites Using ChatGPT,” *arXiv preprint* arXiv:2306.05816, 2023.

[9] S. Kandoi, “Phishing emails 2,” Available: https://www.kaggle.com/datasets/shallykandoi/phishing-emails-2 (2025/7/25)

[10] THE RADICATI GROUP, INC. *Email Statistics Report, 2021-2025*. Available: https://www.radicati.com/ (2025/8/5).

[11] W. W. Cohen. “Enron Email Dataset,” Available: https://www.cs.cmu.edu/ (2025/7/25).

[12] J. L. Fellows, M. Rowe, C. Reuter , and R. K. M. Bankston. “Spear phishing with large language models,” *arXiv preprint* arXiv:2310.05734, 2023.

[13] S. Roy, M. Alshahrani, M. S. Haleem , and M. S. Rahman. “PhishBots: LLM-generated adversarial phishing email attack against LLM-based detectors,” *arXiv preprint* arXiv:2403.12472, Mar. 2024.