國立台北大學資訊工程學系專題報告

基於深度學習與大型語言模型之網路攻防框架

Deep Learning and LLM-driven Cyber Attack and Defense Framework

專題組員:郭庭均、陳華侑、陳胤圻、粘芸瑄、吳翟

專題編號: PRJ-NTPUCSIE-114-012
執行期間:2025 年 7 月 至 2026 年 5 月

1. 摘要

隨著網頁攻擊技術快速演進，傳統網頁應用程式防火牆高度依賴人工分析與規則更新，面對 Zero-Day 攻擊與新型繞過手法時常出現防禦滯後的問題。本研究提出一套基於深度學習與大型語言模型(LLM)驅動之網路攻防演化框架，旨在建立半自動化的攻擊與防禦閉環機制，並提供可套用於不同 WAF 之通用攻防測試架構。於攻擊端，利用 LLM 生成具潛在繞過能力之 XSS-Payload，並結 Double Deep Q-Network(DDQN)進行多輪變異與強化學習，以持續產生高繞過率之攻擊樣本，並透過瀏覽器動態驗證機制確認攻擊有效性。於防禦端，系統針對成功繞過之攻擊樣本進行特徵分析，並利用 LLM 自動生成對應防禦規則，同時結合遞迴解碼與雙路徑檢測機制，以提升對多重編碼與混淆攻擊之偵測能力。實驗結果顯示，本研究所提出之框架能有效發掘現有 WAF 規則缺口，並透過持續性的攻防迴圈強化防禦能力，將傳統被動式防禦模式轉變為主動式演化防禦模式。

2. 簡介

近年來，隨著 Web 應用程式快速普及，各類網頁攻擊手法也持續演化，其中跨網站腳本攻擊(Cross-Site Scripting,XSS)仍為最常見且具高度威脅性的攻擊類型之一。攻擊者可透過惡意腳本竊取使用者資訊、劫持工作階段(Session)甚至進一步控制受害者瀏覽器，對企業與使用者造成嚴重資安風險。

目前多數網頁應用程式防火牆(Web Application Firewall,WAF)主要依賴專家分析攻擊樣本並撰寫規則進行防禦。然而，當新的繞過技術或 Zero-Day 攻擊出現時，防禦規則往往無法即時更新，使系統在規則修補完成前存在防護空窗期。

此外，現有研究大多聚焦於攻擊生成或防禦強化的單一方向，缺乏能夠持續驗證與更新防禦能力的攻防閉環機制，導致防禦效果難以跟上攻擊技術的演進速度。此外，傳統自動化生成 XSS（如使用 Fuzzing 模糊測試）雖然快速，但盲目變異常會破壞 JavaScript 語法結構，導致產出的語句雖然繞過了 WAF，但在瀏覽器中根本無法執行（語義失效）。要製造出既具備對抗性又具備執行有效性的最新型 Payload，過去高度依賴資安專家的手動構造，使得漏洞挖掘成本極高且無法規模化。

因此，本研究希望建立一套可持續演化的攻防框架，使系統能夠自動發掘現有規則缺口，並透過攻擊與防禦的反覆驗證機制持續提升防護能力，進而降低人工維護成本，提升 WAF 面對未知威脅與新型攻擊手法之防禦韌性。

3. 專題進行方式

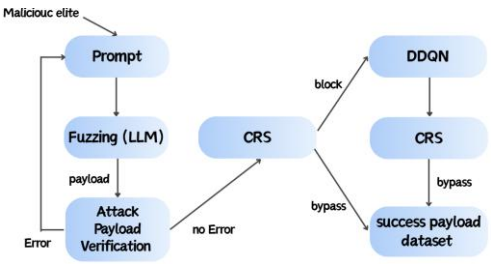
本研究提出一套基於深度學習與大型語言模型之網路攻防演化框架，並建立攻擊生成、攻擊驗證、規則生成與規則強化之閉環機制。

3.1 系統架構

本研究提出一套基於深度學習與大型語言模型之網路攻防演化框架，整體系統由攻擊端與防禦端兩大模組組成。攻擊端負責生成具潛在繞過能力之攻擊樣本，並透過瀏覽器驗證機制確認攻擊有效性；防禦端則針對成功繞過之攻擊樣本進行特徵分析與規則生成，以持續強化防禦能力。透過攻擊生成、攻擊驗證、規則生成與規則更新之循環流程，建立攻擊與防禦共同演化之閉環機制。

3.2 攻擊流程

攻擊流程首先蒐集具有高繞過潛力之攻擊特徵，並將其作為輸入提供大型語言模型生成候選攻擊樣本。生成後之攻擊樣本將透過瀏覽器動態驗證機制檢測其實際執行能力，確認攻擊有效性後再送入防火牆環境進行測試。若成功繞過防火牆則納入攻擊資料集；若遭到攔截，則送入 DDQN 模組進行進一步變異與優化，直到成功繞過或達到最大迭代次數為止。



圖一、攻擊流程圖

3.2.1 LLM攻擊生成機制

首先蒐集具有高繞過潛力之攻擊特徵作為知識來源。本研究給定大型語言模型的主要攻擊方向有三種：多重編碼以隱蔽攻擊特徵、針對防禦較鬆的 HTTP Header 欄位進行攻擊，以及進行語法混淆。系統將這些攻擊特徵整理後寫入 Prompt，作為大型語言模型生成攻擊樣本之依據。

LLM 根據輸入之攻擊特徵產生候選 XSS-Payload 後，系統透過瀏覽器動態驗證機制檢測 Payload 是否能夠成功執行。驗證內容包含彈窗攔截、敏感資料存取、流量外連監控、隱蔽框架分析以及協議跳轉檢測等五大指標，以確保生成之 Payload 具備實際攻擊能力，而非僅具有語法上的變異效果。

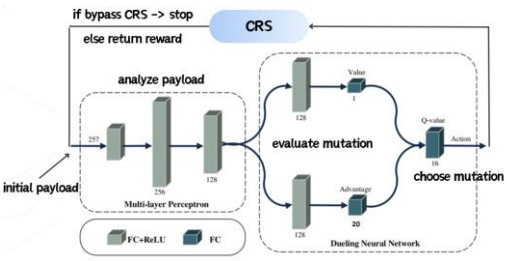
通過驗證之 Payload 將送入 WAF 環境進行測試，若成功繞過防禦機制則直接納入攻擊資料集；若遭到阻擋，則進一步送入 DDQN 模組進行後續強化與變異，以持續提升 Payload 之繞過能力。為確保檢測無盲點，通過驗證之 Payload 將經過三種主要編碼，投放到目標網站的六種 HTTP 區域，包括 Cookie、Header、URL、Args、Body 以及 User-Agent/Referer。

3.2.2 DDQN變異機制

為進一步提升攻擊樣本之繞過能力，本研究採用 Double Deep Q-Network(DDQN)作為 Payload 變異策略選擇模型。DDQN 將目前 Payload 內容視為系統狀態(State)，並從預先定義之變異動作集合(Action Space)中選擇最適合的變異方式進行攻擊優化。

本研究共設計二十種變異方法，包括字串切割、大小寫混淆、多重編碼、特殊字元插入、事件處理函式變形以及上下文轉換等技術。每次變異完成後，Payload 將重新送入 WAF 環境進行測試，並依據測試結果計算獎勵值(Reward)。若成功繞過 WAF 則給予正向獎勵；若遭到攔截則給予負向獎勵，並持續進行下一輪策略搜尋。

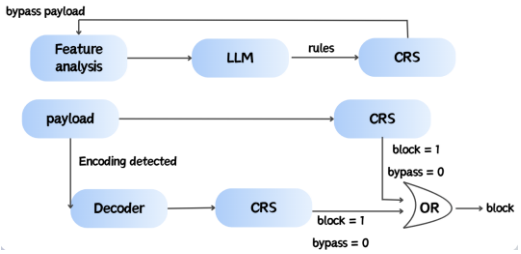
透過不斷更新 Q 值與策略網路，DDQN 能夠學習不同 Payload 狀態下最有效之變異方式，避免僅追求短期效果而忽略長期收益。採用 DDQN 的優點在於模型具備高度『戰略遠見』，不會因短期效果而放棄更有效的變異方法；缺點則為後期訓練策略容易退化，根據本研究測試，將最大迭代次數（Max Round）設定為 10 輪能達到最佳結果。最終，成功繞過防禦機制之 Payload 將匯集成攻擊資料集，作為後續防禦規則生成與防禦能力評估之依據。



圖二、DDQN 變異策略示意圖

3.3 防禦流程

防禦流程以成功繞過防火牆之攻擊樣本作為分析對象，首先進行特徵提取、語境還原與攻擊類型識別，接著利用大型語言模型自動生成對應防禦規則並整合至原有規則集中。此外，系統加入遞迴解碼與雙路徑檢測機制，同時對原始內容與解碼後內容進行檢測，以提升對多重編碼與混淆攻擊之偵測能力。透過持續性的攻防迴圈，系統得以不斷發掘規則缺口並強化防禦能力。



圖三、防禦流程圖

3.3.1 攻擊特徵分析與規則生成

本研究以成功繞過 WAF 之攻擊樣本作為分析對象，透過特徵分析模組擷取攻擊中所使用之繞過技術與關鍵特徵。分析流程包含編碼識別、語境還原、標籤正規化以及關鍵字提取等步驟，以降低攻擊者利用大小寫混淆、字串分割及特殊字元插入等手法所造成之分析困難。

完成特徵分析後，系統將提取之攻擊特徵作為輸入提供大型語言模型，產生對應之防禦規則。生成之規則經測試驗證後整合至 Core Rule Set(CRS)中，以補強現有規則缺口。透過持續蒐集成功繞過之攻擊樣本並產生對應規則，系統能夠建立攻擊與防禦共同演化之閉環機制，使防禦能力隨著攻擊技術的演進而持續增強。

3.3.2 解碼增強與雙路徑檢測機制

為提升系統對多重編碼攻擊之偵測能力，本研究於 CRS 前加入解碼增強模組。當系統偵測到 Payload 存在 URL Encoding、Unicode Encoding、HTML Entity Encoding 或其他編碼形式時，將進行遞迴解碼，以還原攻擊者刻意隱藏之惡意內容。

然而，若直接對所有 Payload 進行解碼，可能導致原本未經編碼之內容被錯誤轉換，進而影響檢測結果。因此，本研究設計雙路徑檢測機制，同時保留原始 Payload 與解碼後 Payload，並將兩者分別送入 CRS 進行檢測。最終透過 OR 邏輯閘進行決策，只要其中任一路徑判定為惡意攻擊，即視為攻擊成立並予以阻擋。

透過解碼增強與雙路徑檢測機制，系統能有效提升對多重編碼、混合編碼及編碼混淆攻擊之偵測能力，同時降低因誤解碼所造成之漏報與誤判問題。

4. 主要成果與評估

4.1 實驗環境建置

為驗證所提出攻防框架之有效性，本專題於虛擬化環境中建置完整測試平台。攻擊端利用大型語言模型生成 XSSPayload，並結合 DDQN 進行攻擊變異，同時透過 Playwright 自動化瀏覽器驗證 Payload 是否能成功執行。防禦端則採用 ModSecurity 搭配 OWASP Core Rule Set(CRS)作為基礎防禦機制，並加入遞迴解碼器與雙路徑檢測架構，以提升對多重編碼攻擊之偵測能力。

本研究之測試資料集來源，良性流量採用開源的 HttpParamsDataset。攻擊流量與變異特徵來源則參考 Libmodsecurity WAF 繞過經驗與開源之 waf-bypass 項目。

CRS 採用異常評分機制(Anomaly Scoring)進行判斷，透過 PARANOIA、BLOCKING_PARANOIA、ANOMALY_INBOUND 及 ANOMALY_OUTBOUND 等參數控制規則嚴格程度與攔截門檻。本專題首先針對不同參數組合進

行測試，以尋找兼顧低誤報率與低漏報率之最佳設定。

4.2 評估指標

為評估攻擊模組與防禦模組之效能，本研究採用誤報率(False Positive Rate, FPR)、漏報率(False Negative Rate, FNR)、攻擊演化增益(Attack Evolution Gain, AEG)及防禦改善分數(Defense Improvement Score, DIS)作為主要評估指標。

4.2.1 誤報率(False Positive Rate, FPR)

誤報率表示正常流量被誤判為惡意攻擊之比例，其定義如下：

$$FPR = \frac{FP}{FP + TN} \times 100\%$$

其中，FP 表示誤判為攻擊之正常流量數量，TN 表示正確判定為正常流量之數量。

4.2.2 漏報率(False Negative Rate, FNR)

漏報率表示惡意攻擊未被成功攔截之比例，其定義如下：

$$FNR = \frac{FN}{TP + FN} \times 100\%$$

其中，FN表示未被偵測之攻擊流量數量，TP表示成功攔截之攻擊流量數量。

4.2.3 攻擊演化增益(Attack Evolution Gain, AEG)

為評估 DDQN 對攻擊樣本繞過能力之提升效果，本研究定義攻擊演化增益(Attack Evolution Gain, AEG)作為評估指標，其定義如下：

$$AEG = \frac{BSR_{DDQN} - BSR_{LLM}}{BSR_{LLM}} \times 100\%$$

其中，BSR_{DDQN} 表示 DDQN 變異後之繞過成功率，BSR_{LLM} 表示原始 LLM 生成 Payload 之繞過成功率。AEG 數值越高，代表 DDQN 對攻擊樣本之優化效果越顯著。

4.2.4 防禦改善分數(Defense Improvement Score, DIS)

為量化改良版 CRS 相較於原始 CRS 之防禦能力提升效果，本研究定義防禦改善分數(Defense Improvement Score, DIS)作為評估指標，其定義如下：

$$DIS = \alpha \left(\frac{DR_{new} - DR_{old}}{DR_{old}} \right) - \beta (FPR_{new} - FPR_{old})$$

其中，DR表示偵測率(Detection Rate)，FPR表示誤報率(False Positive Rate)，而α與β為權重參數。DIS可同時評估偵測能力提升與誤報率變化對整體防禦效果之影響。

4.3 實驗結果

4.3.1 CRS參數調校結果

PARANOIA	BLOCKING PARANOIA	ANOMALY INBOUND	FPR(%)	FNR(%)
1	1	10	0.00	35.25
2	2	10	0.87	1.66
2	2	5	100.0	0.00

表一、CRS設置

由表一可觀察到，當 PARANOIA 與 BLOCKING_PARANOIA 設定為 1 時，雖然誤報率(FPR)為 0%，但漏報率(FNR)高達 35.25%，顯示大量攻擊流量未被成功攔截。當兩項參數提升至 2 並維持 ANOMALY_INBOUND 為 10 時，漏報率大幅下降至 1.66%，而誤報率僅增加至 0.87%，在安全性與可用性之間取得較佳平衡，因此選定此組參數作為後續實驗環境。若進一步降低 ANOMALY_INBOUND 至 5，雖可完全消除漏報，但誤報率上升至 100%，造成所有正常流量皆被阻擋，不適合實際部署。

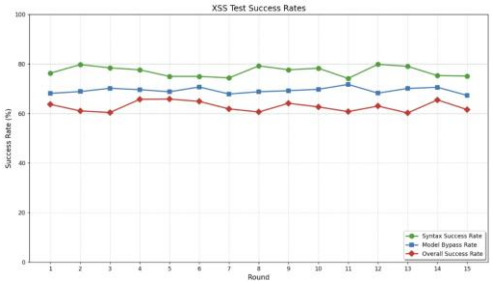
4.3.2 DDQN 攻擊效果評估

為評估 DDQN 對攻擊樣本之優化能力，本研究比較加入 DDQN 前後之攻擊結果。實驗結果顯示，未使用 DDQN 時，Overall Success Rate 平均約為 63%；加入 DDQN 後提升至 73%，Model Bypass Rate 亦由約 69%提升至 79%。

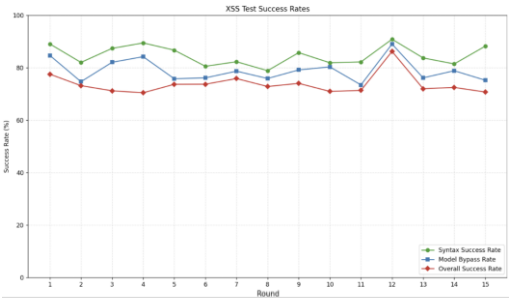
根據攻擊演化增益(Attack Evolution Gain, AEG)計算結果，DDQN 可額外提升約 15.9%

之繞過能力。結果顯示，DDQN 能有效學習較佳的變異策略，使產生之 Payload 具有更高的繞過成功率與攻擊效果。

由圖五可知，加入 DDQN 後，各項成功率均有明顯提升，其中 Overall Success Rate 由 63%提升至 73%，Model Bypass Rate 由 69%提升至 79%。實驗結果顯示，DDQN 能有效學習高價值變異策略，提升 Payload 之繞過能力，AEG 約為 15.9%。



圖四、每輪生成 payload bypass CRS 的成功率(no DDQN)



圖五、每輪生成 payload bypass CRS 的成功率(with DDQN)

4.3.3 改良 CRS 防禦效果評估

為驗證防禦模組之有效性，本研究比較原始 CRS 與加入攻擊特徵分析、LLM 規則生成及解碼增強機制後之改良版 CRS。

實驗結果顯示，原始 CRS 對 39,641 筆 XSS-Payload 僅成功攔截 3,275 筆，偵測率為 8.26%，仍有 36,366 筆攻擊樣本成功繞過防禦機制。加入改良規則後，所有 39,641 筆 XSS-Payload 皆被成功攔截，攻擊繞過數量降為 0，顯示新增規則能有效補足原始 CRS 之規則缺口。

此外，正常流量通過率仍維持 99.13%，誤報率僅為 0.87%，表示系統在提升偵測能力的同時，仍能維持良好的可用性。整體結果顯示，透過攻擊特徵分析與規則自動強化機制，可有效提升 WAF 對未知攻擊與變形攻擊之防禦能力。

防禦系統	XSS 攔截數	XSS 繞過數	偵測率 (%)	正常流量通過率(%)
原始 CRS	3,275	36,366	8.26	99.13
改良 CRS	39,641	0	100.00	99.13

表二、原始CRS與改良CRS防禦效果比較

5. 結語與展望

隨著網頁攻擊技術持續演進，傳統 WAF 在面對未知威脅與新型繞過手法時容易產生防禦缺口。本專題建構一套基於深度學習與大型語言模型之網路攻防演化框架，透過 LLM 生成攻擊樣本、DDQN 進行變異優化以及自動規則生成機制，建立攻擊與防禦共同演化之閉環流程，提升 WAF 對未知攻擊之防禦能力。

未來研究將朝向建構全自動化的網頁防禦沙盒演練架構。預計採用雙代理人（Dual-Agent）機制，將 XSS Payload 資料集依 7:3 比例切分。紅隊 Agent 持續生成變形對抗流量撞擊防火牆；藍隊 Agent 則針對漏洞即時重寫並優化正規表達式（Regular Expression）規則，透過反覆對抗訓練，直至生成低繞過、低誤殺之最佳防禦規則集，達成智慧化自動資安維運（SecOps）的目標。

6. 銘謝

感謝指導教授於專題研究期間提供寶貴的建議與指導，並給予充分的研究資源與發揮空間，使本專題得以順利進行。同時感謝實驗室學長姊分享研究經驗與技術協助，以及所有參與測試與提供建議的同學朋友。最後感謝專題組員一年來的相互合作與努力，共同克服研究與開發過程中的各項挑戰，使本專題得以順利完成。

7. 参考文献

[1] M. Amouei, M. Rezvani, and M. Fateh, "RAT: Reinforcement-Learning-Driven and Adaptive Testing for Vulnerability Discovery in Web Application Firewalls," IEEE Transactions on Dependable and Secure Computing, vol. 21, no. 4, pp. 2852–2868, Jul./Aug. 2024.

[2] Y. Yao, J. He, T. Li, Y. Wang, X. Lan, and Y. Li, "An Automatic XSS Attack Vector Generation Method Based on the Improved Dueling DDQN Algorithm," IEEE Transactions on Dependable and Secure Computing, vol. 21, no. 4, pp. 2852-2868, Jul./Aug. 2024.

[3] V. Babaey and A. Ravindran, "GenXSS: an AI-Driven Framework for Automated Detection of XSS Attacks in WAFs," arXiv preprint arXiv:2504.08176, Apr. 2025.

[4] C. Wu, J. Chen, S. Zhu, W. Feng, K. He, and R. Du, "WAFBooster: Automatic Boosting of WAF Security Against Mutated Malicious Payloads," IEEE Transactions on Dependable and Secure Computing, vol. 22, no. 2, pp. 1118–1131, Mar.–Apr. 2025.