國立臺北大學資訊工程學系114學年度專題

基於深度學習與大型語言模型之網路攻防框架

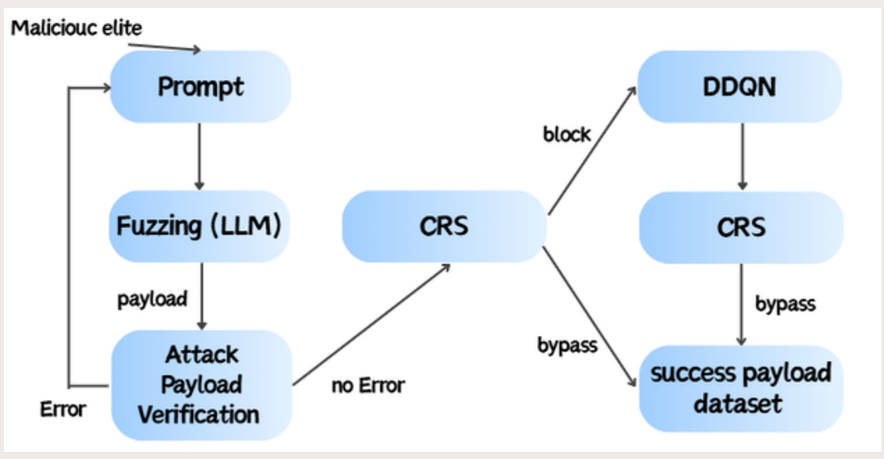
Deep Learning and LLM-driven Cyber Attack and Defense Framework

成員：郭庭均、陳華侑、陳胤圻、粘芸瑄、吳翟

動機

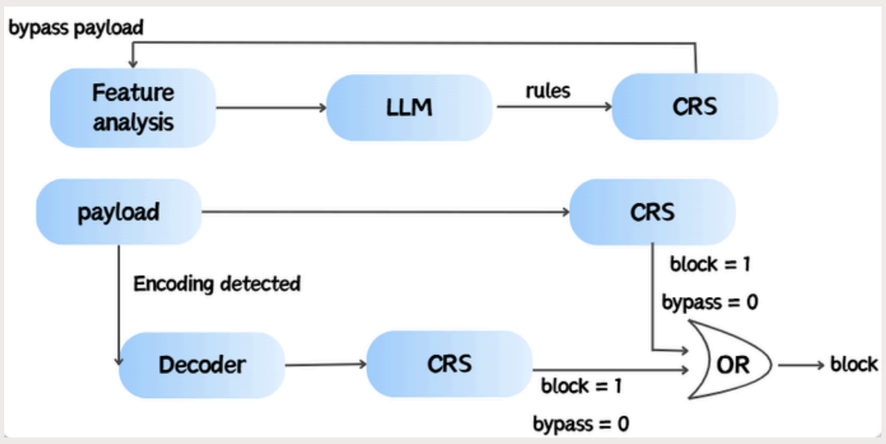
隨著科技發展，網頁攻擊手法日新月異，傳統WAF因防禦滯後、缺乏閉環機制，難以應對Zero-Day威脅。本研究旨在建立一套基於LLM 與深度學習的「主動式演化防禦框架」。透過 LLM 自動生成高對抗性樣本，結合強化學習與規則自動化，構建持續驗證的攻防閉環。此架構不僅能自動發掘防禦缺口、克服語義失效問題，更實現了資安防護從「被動修補」向「主動演化」的轉型，大幅提升防禦韌性並降低維護成本。

攻擊流程介紹



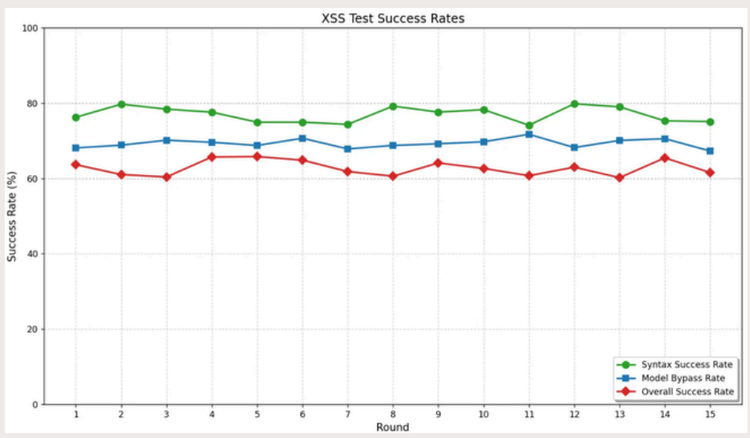
整體攻擊流程：首先蒐集具有高繞過潛力之攻擊特徵，並將其作為輸入提供大型語言模型生成候選攻擊樣本。生成後之攻擊樣本將透過瀏覽器動態驗證機制檢測其實際執行能力，確認攻擊有效性後再送入防火牆環境進行測試。若成功繞過防火牆則納入攻擊資料集；若遭到攔截，則送入DDQN 模組進行進一步變異與優化，直到成功繞過或達到最大迭代次數為止。

防禦流程介紹



整體防禦流程：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。

實驗結果



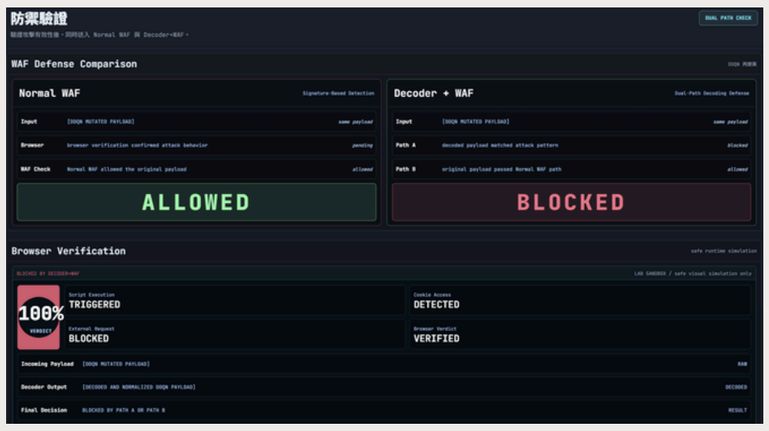
每輪生成 payload bypass CRS 的成功率(no DDQN)

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

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

使用者界面

本系統透過Web互動式介面整合LLM Payload生成、DDQN變異強化、Browser Verification驗證及WAF防禦檢測，並以視覺化方式呈現攻擊生成到防禦驗證的完整攻防流程與結果。



結論

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。