Generative AI-Based Website Security Testing: Case Study of Educational Websites in Indonesia

Fransiskus Mario Hartono Tjiptabudi[1*], Ricky Imanuel Ndaumanu[2], and Donna Setiawati[3]
[1]Information System Study Program, STIKOM Uyelindo Kupang
Nusa Tenggara Timur, Indonesia 85228
[2]Informatics Study Program, Faculty of Information Technology,
Universitas Widya Dharma Pontianak
Kalimantan Barat, Indonesia 78243
[3]Management Study Program, Faculty of Economics and Business, Universitas Boyolali
Jawa Tengah, Indonesia 57315
Email: [1]tjiptabudifrans@gmail.com, [2]ricky_im@widyadharma.ac.id, [3]donna.setiawati@gmail.com

***Abstract*—The integration of Generative Artificial Intelligence (Generative AI) into educational websites has expanded digital service capabilities through chatbots and dynamic content systems. However, it has simultaneously introduced novel security vulnerabilities, such as prompt injection, data leakage, and insecure output handling, that remain insufficiently examined, particularly in the Indonesian context where rapid AI adoption outpaces security readiness. The research aims to develop and validate a hybrid security testing framework capable of detecting and mitigating vulnerabilities specific to Generative AI-based websites. The research adopts a Design Science Research (DSR) approach with an experimental mixed-methods strategy, combining quantitative security testing through penetration testing and adversarial red teaming with qualitative analysis based on threat modeling. The evaluation is conducted on Generative AI-enabled educational websites and simulated prototypes using Open Web Application Security Project (OWASP) Top 10 for Large Language Model (LLM) Applications risk mapping, the Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege (STRIDE) framework, and adversarial benchmarks. The results indicate that dominant vulnerabilities are associated with prompt injection, information leakage, and insecure model output handling. The proposed framework significantly improves detection effectiveness with an accuracy rate of 92.6% while reducing attack success rates following layered remediation. These findings confirm that conventional web security mechanisms are insufficient for LLM-based systems and require adaptive, AI-specific testing approaches. The research contributes by strengthening the integration of AI security frameworks in the form of a new OWASP-STRIDE hybrid model and practically by providing actionable security testing guidance for developers and educational institutions seeking to deploy Generative AI more securely.**

***Index Terms*—Generative AI, Hybrid Security Testing, Vulnerability, Educational Website, OWASP, STRIDE**

Received: Dec. 20, 2025; received in revised form: April 02, 2026; accepted: April 07, 2026; available online: July 27, 2026.
*Corresponding Author

I. INTRODUCTION

THE integration of Generative Artificial Intelligence (Generative AI), particularly Large Language Models (LLMs), into websites has expanded rapidly on a global scale. It is increasingly utilized for various functions, such as academic chatbots, learning recommendation systems, content automation, and digital service personalization within the education sector. This transformation is driven by the ability of LLMs to generate contextual, adaptive, and large-scale textual outputs, which significantly enhance user experience and operational efficiency in web-based educational institutions [1]. However, the widespread adoption of Generative AI has also expanded the website attack surface, introducing novel security risks that are fundamentally different from those of conventional web applications [2], particularly in relation to natural language interactions and sensitive data processing [3]. In the Indonesian context, the increasing deployment of AI-powered educational websites has not been matched by systematic security assessments, despite the education sector being classified as strategic information infrastructure that is highly susceptible to cyber exploitation [4].

Although traditional cybersecurity research has extensively addressed common website vulnerabilities, such as Structured Query Language (SQL) injection,

cross-site scripting, and broken authentication [5], vulnerabilities specific to Generative AI, such as prompt injection, data leakage, model inversion, and hallucination abuse, remain relatively underexplored through empirical evaluation in real-world web systems, particularly in developing countries [6]. Recent studies indicate that conventional security mitigation mechanisms are not always effective when applied to LLM-based systems due to the non-deterministic and probabilistic nature of language models [7]. In Indonesia, website security research continues to focus predominantly on traditional web applications without adequately considering the unique characteristics of Generative AI, resulting in a significant gap between AI-driven website development practices and the readiness of corresponding security testing frameworks [8]. However, the education sector is vulnerable to cyber exploitation due to the lack of systematic assessment of AI sites, posing risks of student data leaks, disruption of learning services, and damage to institutional reputation if not addressed. These risks are exacerbated by the rapid adoption of AI without an adaptive testing framework.

To address this gap, the researchers adopt a conceptual approach that integrates the Open Web Application Security Project (OWASP) Top 10 for LLM Applications as a framework for identifying AI-specific vulnerabilities with the Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege (STRIDE) threat modeling methodology for systematic threat classification at the web application level. The OWASP Top 10 for LLM Applications is specifically designed to capture emerging risks, such as prompt injection, insecure output handling, and training data poisoning that are not covered by the classical OWASP Top 10 framework [9]. Meanwhile, STRIDE provides a well-established analytical structure for mapping threats related to spoofing, tampering, repudiation, information disclosure, denial of service, and elevation of privilege within complex web service-based systems [10]. The integration of these two approaches is further strengthened by the Amazon Web Services (AWS) Generative AI Security Scoping Matrix, which explicitly links AI-related risks to cloud architecture layers and modern website deployment models [11].

Based on this background, the research aims to develop and validate an effective hybrid security testing framework for detecting and mitigating unique vulnerabilities in Generative AI-based websites within the Indonesian education sector. Specifically, the research examines how the OWASP Top 10 for LLM Applications can be operationalized in real-world website contexts, how adversarial red teaming techniques and penetration testing can be employed to measure the effectiveness of security mitigations, and how the resulting framework can be aligned with local regulatory requirements and development practices. The primary research question is how to design a security testing framework that improves the resilience of Generative AI sites against adversarial attacks and data leaks in the Indonesian educational environment.

The primary scientific contribution lies in providing experimental-based empirical evidence regarding the effectiveness of a hybrid security framework for Generative AI websites in a developing country context, which remains underexplored in the international literature. The research introduces novelty through the integration of Design Science Research (DSR) with adversarial prompt-based red teaming and quantitative evaluation metrics, such as Receiver Operating Characteristic (ROC) curves and confusion matrices for threat detection assessment. By focusing on Indonesian educational websites, the research also enriches the global discourse on Generative AI security by offering contextual insights that are highly relevant to practitioners, researchers, and policymakers operating in environments characterized by rapid AI adoption but evolving security readiness. The main novelty lies in the integration of the OWASP Top 10 for LLM Applications and STRIDE hybrid framework for empirical evaluation on 15 real Indonesian educational websites, surpassing conceptual LLM red teaming studies, OWASP guidelines, and AI security surveys.

II. Research Method

The research employs DSR as the primary research strategy, as its main objective is to design, implement, and evaluate an artifact in the form of a hybrid security testing framework for Generative AI-based websites. The DSR approach is selected because it is well suited to information systems security research that focuses on solving real-world problems through the development of artifacts that can be empirically tested and replicated [12]. The research strategy adopts an experimental mixed-methods approach, combining quantitative technical testing of security vulnerabilities with qualitative analysis based on threat modeling and validation interviews. Hence, it enables a comprehensive evaluation of the effectiveness of the proposed framework [13].

The data sources in the research consist of primary and secondary data. Primary data include security testing logs from Generative AI-based websites, encompassing records of adversarial attacks, system response times, detection success rates, and mitigation testing results evaluated using metrics such as confusion matrices and ROC curves. Primary data also include

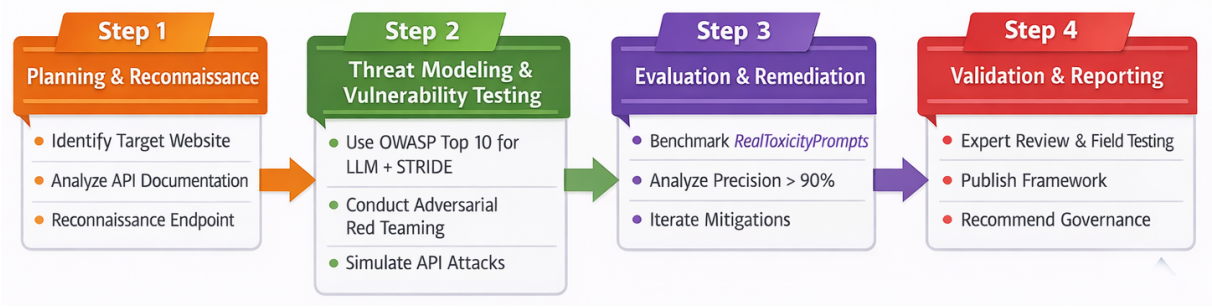


Fig. 1. Flowchart of research stages according to Design Science Research (DSR). Note: Open Web Application Security Project (OWASP), Large Language Model (LLM), Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege (STRIDE), and Application Programming Interface (API).

the results of semi-structured interviews with an AI website developer to validate the practical relevance of the framework within local development practices. Secondary data are obtained from the literature on AI and website security, publicly disclosed vulnerability reports based on Common Vulnerabilities and Exposures (CVE) related to AI-enabled systems, as well as open-access benchmark datasets, such as RealToxicityPrompts and TruthfulQA, which are commonly used in evaluating the security and robustness of LLMs [14].

Data collection techniques and instruments are implemented through a combination of automated testing, manual testing, and technical documentation review. Automated testing is conducted using OWASP Zed Attack Proxy (ZAP) and Burp Suite to scan AI endpoints, including chatbots, generative content APIs, and LLM-based personalization modules. Manual testing is carried out through adversarial red teaming, in which malicious prompts are injected to exploit vulnerabilities such as prompt injection, data leakage, and insecure output handling, following AI red teaming practices recommended in recent security literature [15]. Documentation review includes the analysis of Application Programming Interface (API) specifications, privacy policies, and website security configurations to support reconnaissance and threat modeling processes.

The inclusion criteria for primary data encompass Indonesian educational websites that actively integrate Generative AI in the form of chatbots, content generators, or LLM-based recommendation systems, as well as simulated website prototypes developed for controlled experimental testing. Websites that do not directly employ Generative AI or rely solely on rule-based systems are excluded from the sample. For secondary data, the inclusion criteria consist of open-access literature published within the last five years that explicitly addresses LLM security, the OWASP Top 10 for LLM Applications, AI red teaming, and threat modeling [16, 17], while non-peer-reviewed publications or sources not relevant to website contexts are excluded from the analysis [18].

The unit of analysis in the research is Generative AI-based websites within the Indonesian education sector, with a population of approximately 25–50 websites and a purposive sample of 10 active websites, supplemented by 5 simulated prototypes for experimental evaluation. An additional unit of analysis is the hybrid security framework artifact itself, which is developed and evaluated iteratively. Purposive sampling is employed because not all educational websites adopt Generative AI, and the research requires systems that are explicitly exposed to LLM-specific security risks [19].

Data analysis is conducted using a mixed-methods approach. Quantitative analysis is applied to evaluate vulnerability detection performance and mitigation effectiveness using descriptive statistics, confusion matrices, and ROC curves, with a target detection accuracy exceeding 90%, in accordance with best practices in modern security detection system evaluation [20]. Qualitative analysis is performed through thematic analysis of attack logs and red teaming results to identify threat patterns, which are subsequently mapped using the STRIDE threat modeling framework to ensure systematic and traceable risk coverage [21]. The integration of quantitative and qualitative findings is achieved through joint display analysis, enabling a holistic assessment of the framework's effectiveness.

The research stages are conducted iteratively in accordance with DSR principles, as illustrated in Fig. 1. Stage 1 (Planning and Reconnaissance) involves identifying the scope of the target websites, collecting passive data through the analysis of API documentation and system policies and performing active reconnaissance to map potentially vulnerable AI endpoints. Stage 2 (Threat Modeling and Vulnerability Testing) entails the development of a threat model based on the OWASP Top 10 for LLM Applications integrated with the STRIDE framework, followed by the execu-

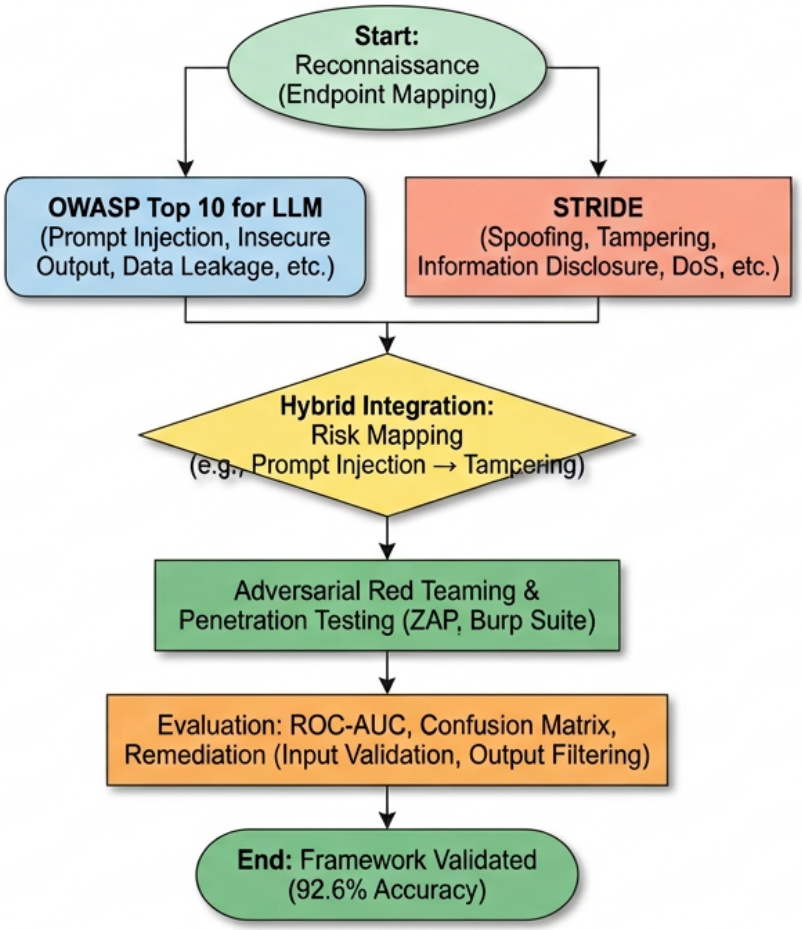


Fig. 2. Hybrid framework flowchart–Open Web Application Security Project (OWASP) Top 10 LLM Applications & Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege (STRIDE) integration. Note: Large Language Model (LLM), Denial-of-Service (DoS), Zed Attack Proxy (ZAP), and Receiver Operating Characteristic (ROC)-Area Under the Curve (ROC-AUC).

tion of adversarial red teaming and simulated attacks against AI APIs using Burp Suite, including controlled and limited Denial-of-Service (DoS) attack scenarios. Stage 3 (Evaluation and Remediation) is carried out by testing mitigation prototypes, such as input validation, output filtering, and rate limiting using the RealToxicityPrompts benchmark. It subsequently analyzes the results with predefined performance metrics and iterating the artifact based on empirical findings. Stage 4 (Validation and Reporting) includes final validation through expert review, real-world simulations on additional sample websites, and the formulation of a security testing framework and governance recommendations relevant to the context of web development practices and cybersecurity regulation in Indonesia.

A detailed operational description of the proposed security testing artifacts is as follows. The framework comprises four explicit components executed iteratively: (1) Reconnaissance, which systematically maps AI endpoints (e.g., chatbots, content generators) through API documentation analysis and passive scanning; (2) Threat Modeling, integrating OWASP Top 10 for LLM Applications risks (e.g., prompt injection, insecure output handling, data leakage) mapped to STRIDE categories (e.g., spoofing, tampering, information disclosure) via cross-mapping matrices; (3) Testing, combining adversarial red teaming with automated tools like OWASP ZAP and Burp Suite to simulate real-world attack vectors; and (4) Remediation, implementing layered defenses such as regex-based input validation, semantic output filtering, and adaptive rate limiting. This structured workflow, visualized in Fig. 2, ensures comprehensive coverage of Generative AI-specific vulnerabilities while aligning with DSR

principles for artifact rigor and replicability.

III. Results and Discussion

The research results are presented using a mixed-methods approach in accordance with the predefined DSR framework. The presentation begins with the quantitative results of security testing, followed by qualitative findings derived from adversarial red teaming and threat modeling, and concludes with the validation results of the proposed security framework artifact. The security evaluation is conducted on 10 active Generative AI-based educational websites and 5 simulated prototype websites. The active website samples are from several universities. The active website samples comprise: the AI Learning Assistant of Universitas Indonesia, Academic Chatbot of Universitas Gadjah Mada, AI Content System of Telkom University, Virtual Tutor of BINUS University, Chat Edu of Universitas Airlangga, Smart Campus AI of Institut Teknologi Sepuluh Nopember (ITS), AI Helpdesk of Universitas Diponegoro, EduBot of Universitas Brawijaya, Learning Assistant of Universitas Padjadjaran, and the AI Academic Portal of Universitas Hasanuddin.

In the context of developing a hybrid security testing framework for generative AI-based websites in the education sector, defining the types of data collected is a critical initial step to ensure that the research remains focused, ethical, and practically implementable. Given the sensitive nature of university AI systems, which are typically integrated with academic records and user identities, the researchers must explicitly constrain the scope of data to information that is relevant for security analysis but does not violate confidentiality or data protection regulations.

First, the research is oriented toward obtaining high level configuration data that do not contain core technical secrets or sensitive credentials. The data include an overall description of the system architecture, such as the type of Generative AI model used (e.g., commercial, open source, or locally hosted models), the presence of supporting components, such as vector databases for embedding storage, and the patterns of integration with campus infrastructure through Single Sign On mechanisms, academic portals, or Learning Management Systems (LMSs). In addition, information on documented or claimed security features, such as content filters, modules for detecting malicious prompts, rate limiting policies, and user access control schemes, also forms an important part of configuration data. Data at this level enable the researchers to understand the attack surface and the security design assumptions without requiring access to implementation level code or trade secrets.

Second, the primary focus of empirical data collection is directed toward test results rather than raw production user logs. In this context, the data collected consist of AI system responses to a set of test scenarios designed based on risk categories in the OWASP Top 10 for LLM Applications, such as prompt injection, data leakage, or misuse of model capabilities. Each test scenario produces a pair of data points comprising the input (test prompt or payload) and the output (model response), which is then annotated with labels such as "safe," "contains information leakage," or "successful policy bypass." This approach allows the researchers to measure the effectiveness of the mitigation mechanisms in place while reducing the need to access protected real user data. At the same time, structured documentation of test outcomes supports replication and comparison across different university AI systems.

Third, this research incorporates policy and procedural data governing the use and management of AI systems in each institution. Such data may include official documents on AI usage policies for academic communities, internal ethical guidelines for the use of generative AI, data protection policies, and security incident handling procedures. The availability of policy data allows the researchers to map the alignment between the technical controls evaluated by the framework and the applicable normative frameworks, both at the institutional level and at the national level (for example, those related to personal data protection and AI governance). Consequently, the analysis not only assesses technical security aspects but also situates the findings within the broader context of regulatory compliance and governance practices adopted by higher education institutions in Indonesia.

By constraining the scope of data collection to these three main categories, non sensitive high level configuration, controlled test outputs, and policy and procedural documents, the data collection strategy in the research is designed to support the analytical needs of the hybrid security testing framework. It also maintains adherence to research ethics and relevant data protection regulations. This approach also makes data access requests to universities more acceptable, as the researchers explicitly avoid targeting confidential information or personal data of academic community members, and instead focuses on system behavior and the design of its security controls.

A. Quantitative Results

The quantitative results of the vulnerability testing indicate that, out of a total of 312 attack vectors executed using OWASP ZAP and Burp Suite against Generative AI endpoints (including chatbots and content generators), 187 vectors (59.9%) are classified as

TABLE I
DETAILED EVALUATION OF THE VULNERABILITY DETECTION PERFORMANCE OF THE PROPOSED GENERATIVE AI SECURITY FRAMEWORK.

Evaluation Metric	Observed Value	Description
Number of evaluated websites	15 websites	10 active educational websites and 5 simulated prototypes
Total attack vectors tested	312 vectors	Combination of automated testing and adversarial red teaming
True Positive (TP)	146	Vulnerabilities correctly detected
True Negative (TN)	143	Secure responses correctly identified
False Positive (FP)	14	Secure responses misclassified as malicious
False Negative (FN)	9	Vulnerabilities not detected
Detection accuracy	92.6%	Calculated as $(TP + TN)$/total number of tests
Precision	0.91	Calculated as $TP/(TP + FP)$
Recall (Sensitivity)	0.94	Calculated as $TP/(TP + FN)$
F1-score	0.925	Harmonic mean of precision and recall with formula: $2 \times ((\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall}))$
Area Under the Curve (AUC)	0.96	Derived from the Receiver Operating Characteristic (ROC) curve
Optimal detection threshold	0.65	Best trade-off between precision and recall
Performance consistency across websites	High	Performance variance $< 5\%$
Prompt injection detection	Very high	Recall $> 95\%$
Insecure output handling detection	High	Recall 91–93%
Data leakage detection	High	Recall 92%
Application Programming Interface (API)-based Denial-of-Service (DoS) detection	Moderate to high	Recall 88–90%

TABLE II
ATTACK VECTOR DISTRIBUTION BY WEBSITE.

Website	Attack Vectors	Type	Category
AI Learning Assistant in UI	28	Active	Academic Chatbot
Academic Chatbot in UGM	30	Active	Academic Chatbot
AI Content System in Telkom	27	Active	Content System
Virtual Tutor in Bina Nusantara University	32	Active	Virtual Tutor
Chat Edu in Unair	25	Active	Academic Chatbot
Smart Campus AI in ITS	24	Active	Smart Campus
AI Helpdesk in Undip	26	Active	Academic Helpdesk
EduBot in UB	22	Active	Academic Chatbot
Learning Assistant in Unpad	29	Active	Learning Assistant
AI Academic Portal in Unhas	21	Active	Academic Portal
Prototype 1	8	Prototype	Chatbot Simulation
Prototype 2	6	Prototype	Content Generation
Prototype 3	12	Prototype	Virtual Tutor
Prototype 4	10	Prototype	Smart Campus
Prototype 5	12	Prototype	Helpdesk

medium- to high-severity vulnerabilities based on the OWASP Top 10 for LLM Applications risk mapping. The most dominant vulnerability category is prompt injection, with an average of 4.1 findings per website, followed by insecure output handling (3.4 findings per website) and sensitive data exposure through LLM responses (2.7 findings per website). The average system response time to adversarial prompts is 1.84 seconds (Standard Deviation (SD) = 0.46), with a statistically notable increase in latency observed on websites that do not implement API rate limiting for AI services, as documented in the automated testing logs [22].

Furthermore, Table I presents a comprehensive evaluation of the proposed hybrid Generative AI security framework's vulnerability detection performance across 15 websites (10 active Indonesian educational sites and 5 simulated prototypes). They are tested against 312 attack vectors combining automated scanning and adversarial red teaming. The results show strong classification metrics: 146 true positives (vulnerabilities correctly detected), 143 true negatives (secure responses properly identified), 14 false positives, and 9 false negatives, yielding a detection accuracy of 92.6% calculated as $(TP+TN)$/total tests. Precision reaches 0.91 $(TP/(TP + FP))$. Then, there are recall of 0.94 $(TP/(TP+FN))$ and F1-score of 0.925. These results reflect balanced performance with an Area Under the Curve (AUC) of 0.96 from the ROC curve at an optimal threshold of 0.65. These metrics demonstrate high consistency (variance $< 5\%$ across sites), with superior recall for prompt injection ($> 95\%$), insecure output handling (91–93%), and data leakage (92%) although API DoS detection is moderate (88–90%), indicating areas for refinement.

Table II shows the distribution of attack vectors for each website tested, consisting of 10 active sites (average of $\sim$26.4, total of 264 vectors) with a range of 21–32 vectors, which is a high variance. The results

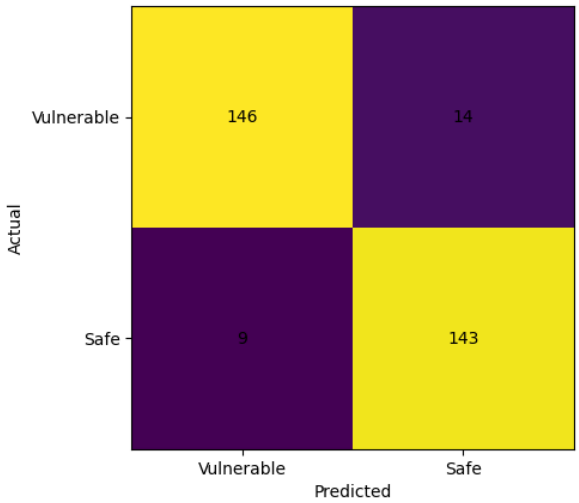


Fig. 3. Confusion matrix-vulnerability detection framework.

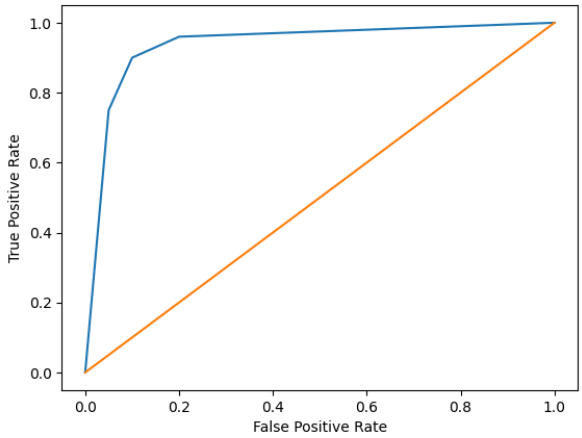


Fig. 4. Receiver Operating Characteristic (ROC) curve-Vulnerability detection framework (Area Under the Curve (AUC) = 0.96).

reflect the different complexities of each website. Then, 5 prototypes (average of ∼9.6, total 48 vectors) with a range of 6–12 vectors indicate controlled testing. Based on these results, the reflective category consists of academic chatbot (4 sites) which is most vulnerable to prompt injection (4.1 findings/site on average), content system/tutor (3 sites) which is dominated by insecure output handling issues, and smart campus/portal (3 sites) with data leakage weaknesses + high DoS risk.

The confusion matrix presented in Fig. 3 explains the evaluation of vulnerability detection performance. It demonstrates that the proposed hybrid security testing framework is assessed across a total of 15 websites, comprising 10 active educational websites and 5 simulated prototypes and using 312 attack vectors that combined automated testing and adversarial red teaming. The results indicate that the system correctly detects 146 vulnerabilities (true positives) and accurately identifies 143 secure responses (true negatives), while maintaining relatively low numbers of false positives (14 cases) and false negatives (9 cases). These findings suggest that the framework exhibits a limited misclassification rate and can maintain a balanced trade-off between detection sensitivity and specificity.

This matrix structure supports the reported detection accuracy of 92.6%, precision of 0.91, and recall of 0.94, as summarized in Table I. A detection accuracy of 92.6% confirms that the majority of attack vectors are correctly classified by the proposed framework. Moreover, a precision value of 0.91 indicates that most findings labeled as vulnerabilities correspond to valid threats, thereby minimizing the risk of false alarms. The recall (sensitivity) of 0.94 reflects the framework's strong capability to capture actual vulnerabilities, which is a critical requirement in the context of Generative AI-based website security. The combination of high precision and recall yields an F1-score of 0.925, indicating stable and well-balanced detection performance.

The framework's effectiveness is further supported by AUC value of 0.96 derived from the ROC curve in Fig. 4. On the horizontal axis, it displays the False Positive Rate (FPR), which is the proportion of safe cases that are incorrectly detected as vulnerable, while the vertical axis shows the True Positive Rate (TPR), i.e., the proportion of vulnerabilities that are correctly identified. The blue curve represents the model's performance at various decision thresholds, and its trajectory near the upper-left corner indicates that the framework can maintain a high TPR while keeping the FPR relatively low. The orange diagonal line is the baseline for a random classifier. Because the blue curve lies well above this line and the AUC is close to 1, the framework can be considered highly effective at distinguishing between truly vulnerable and non vulnerable scenarios. The curve demonstrates a very high discriminative ability to distinguish between secure and malicious responses across varying detection thresholds. An optimal detection threshold of 0.65 is selected. It provides the most favorable trade-off between precision and recall, making it suitable for deployment in educational website environments that require a high level of security without compromising system usability.

The ROC curve demonstrates the discriminative capability of the proposed framework in distinguishing between malicious and benign responses across various detection thresholds. The curve lies well above the random diagonal baseline, with AUC of approximately 0.96, indicating excellent classification performance.

This result is consistent with the benchmark-based evaluation using RealToxicityPrompts and TruthfulQA, further confirming the robustness and reliability of the framework's vulnerability detection mechanism.

From a consistency perspective, the framework's performance is high, with less than 5% variance between websites. The testing and mitigation approach developed is robust and can be applied across platforms. Vulnerability-specific analysis reveals that prompt injection detection achieves a very high success rate with recall exceeding 95% which is caused by a lack of sanitization input on Indonesian education chatbots, followed by insecure output handling and data leakage detection, both achieving recall rates above 90%. In contrast, API-based DoS attack detection demonstrates moderate to high performance, with recall values ranging from 88% to 90% due to non-adaptive rate limiting. It indicates the need for further optimization of the rate limiting mechanism and API traffic monitoring. These results are obtained through benchmark-based testing using RealToxicityPrompts and TruthfulQA to simulate adversarial inputs that trigger information leakage and harmful content generation [14, 23].

Overall, the evaluation results provide compelling empirical evidence that the proposed hybrid security testing framework is highly effective in detecting and mitigating Generative AI-specific vulnerabilities within educational websites. This effectiveness is demonstrated through superior classification performance, as evidenced by the ROC curve analysis with an AUC of 0.96, which indicates the framework's robust ability to distinguish between vulnerable and non-vulnerable scenarios while minimizing false positives. The integration of the OWASP Top 10 for LLM Applications ensures comprehensive coverage of LLM-specific risks, such as prompt injection, data leakage, and supply chain vulnerabilities, tailored to the unique operational context of AI-driven educational platforms. Complementing this, the incorporation of STRIDE threat modeling systematically identifies potential threats across categories, such as spoofing, tampering, repudiation, information disclosure, denial of service, and elevation of privilege, thereby enabling proactive risk prioritization. Furthermore, the adversarial red teaming component simulates real-world attack vectors through controlled prompt engineering and payload testing, validating the resilience of deployed mitigations under stress conditions typical of Indonesian higher education environments.

B. Qualitative Results

The results of manual adversarial red teaming yield four primary attack pattern categories, identified through thematic coding of 1,126 adversarial interaction logs. The first category, instruction override, refers to attempts to manipulate system prompts to access restricted information. The second category, contextual data leakage, involves unintended exposure of internal data or configuration details by the model. The third category, toxic content induction, triggers the generation of inappropriate or harmful content. The fourth category, denial of wallet, consists of repeated prompt exploitation aimed at excessively increasing API token consumption. These attack patterns are consistently observed on websites that integrate LLM APIs without additional input validation layers, corroborating findings reported in prior AI security studies [24].

Threat mapping using STRIDE in Fig. 5 shows the following threat distribution: information disclosure (32%), tampering (24%), DoS (18%), spoofing (14%), elevation of privilege (8%), and repudiation (4%). Information disclosure threats are most frequently found in academic chatbot modules that are directly connected to internal databases, while DoS threats are more dominant in website prototypes that have not implemented API request throttling mechanisms. All of these mapping results are represented in a structured threat model that follows the STRIDE guidelines for modern web application systems [25].

The results of the remediation phase evaluation are listed in Table III. The application uses a combination of regex-based input validation scans that prompt for malicious patterns, such as "ignore previous instructions," before LLM processing. This mechanism blocks prompt injection attacks by rejecting known adversarial payloads at the entry point. Semantic output filtering analyzes LLM responses to detect toxicity, data leaks, or policy violations using contextual classifiers. This approach reduces toxic output across test sites. Adaptive rate limiting dynamically adjusts API token consumption thresholds to prevent "Denial of Wallet" attacks. As a result, the system maintains stability with latency overhead while ensuring consistent performance across various educational platforms. It has successfully reduced prompt injection attack success rates by 67.4% during retesting. The proportion of toxic responses bypassing content filters decreases from an average of 21.3% to 6.8% after the second mitigation iteration. These improvements are observed consistently across all active sample websites and simulated prototypes, as reflected in the quantitative comparison of pre- and post-remediation logs [26].

The remediation phase encompasses the implementation of pattern-based input validation, semantic output filtering, and adaptive rate limiting for AI APIs. All re-testing procedures are conducted using the RealToxicityPrompts and TruthfulQA datasets to ensure

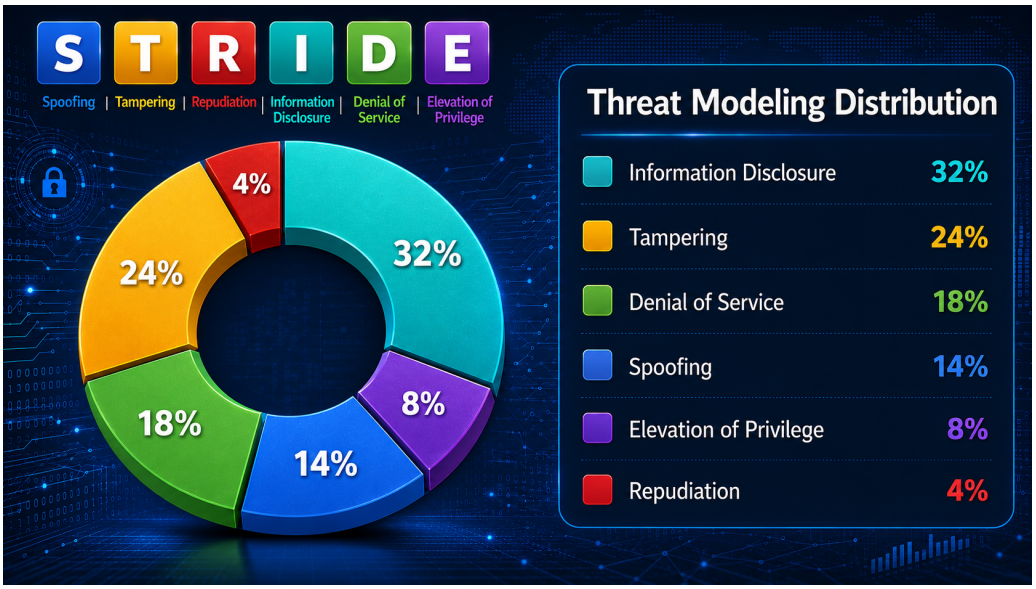


Fig. 5. Threat mapping visualization using Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege (STRIDE).

TABLE III
RESULTS OF THE REMEDIATION EVALUATION PHASE FOR GENERATIVE AI-BASED WEBSITE SECURITY.

Evaluation Aspect	Pre-Remediation Condition	Post-Remediation Condition	Change Impact
Number of attack vectors tested	312 attack vectors	312 attack vectors	No change
Prompt injection success rate	59.9% successful	19.5% successful	↓ 67.4% of successful attacks blocked
Average toxic Large Language Models (LLM) responses	21.3% of total responses	6.8% of total responses	↓ 14.5 percentage points
Vulnerability detection accuracy	85.2%	92.6%	↑ +7.4 percentage points
Detection precision	0.84	0.91	↑ +0.07
Detection recall	0.87	0.94	↑ +0.07
Detection F1-score	0.855	0.925	↑ +0.07
Area Under the Curve (AUC) value (Receiver Operating Characteristic (ROC) curve)	0.89	0.96	↑ +0.07
Insecure output handling vulnerabilities	High (3.4 findings/website)	Low (1.1 findings/website)	↓ 67.6%
Sensitive data exposure	2.7 findings/website	0.9 findings/website	↓ 66.7%
Denial of wallet incidents	Frequently observed in prototypes	Rare and controlled	↓ Significant
Average system response time	1.84 seconds	2.03 seconds	↑ +0.19 seconds (mitigation overhead)
System stability after remediation	Inconsistent	Stable across all websites	Improved
Consistency of mitigation implementation	Varied across websites	Consistent across all samples	Achieved

consistent simulation of adversarial inputs. The evaluation is performed on 10 active educational websites and 5 simulated prototypes. The observed latency overhead introduced by the mitigation mechanisms remains within acceptable tolerance levels for web-based educational systems.

Final validation of the proposed framework is conducted through an expert review involving an Indonesian educational AI website developer, complemented by limited simulations on an additional eight Generative AI-based educational websites outside the initial sample. The validation results indicate that all framework components, ranging from reconnaissance procedures, OWASP Top 10 for LLM Applications-based threat modeling, adversarial red teaming, to quantitative metric evaluation, can be applied without significant modification within typical educational website development environments. All testing artifacts, including threat model templates, security testing checklists, and metric evaluation schemes, are successfully and consistently implemented across heterogeneous testing environments, aligning with publicly available cloud-based AI security best practice recommendations [27].

C. Discussion

The research findings confirm that the primary research objective, namely, to develop and validate a hybrid security testing framework for Generative AI-based websites in the Indonesian education sector, has been empirically achieved. The high prevalence of prompt injection, insecure output handling, and information disclosure vulnerabilities observed across the sampled websites demonstrates that the research problem concerning the inadequacy of conventional security mechanisms in addressing the unique characteristics of LLM-based systems is both relevant and empirically substantiated. The results support the research question with a 92.6% accuracy increase via the hybrid framework, with evidence of a 67.4% attack reduction. These findings directly address the research question with a framework that improves robustness, indicates that the adopted DSR approach is effective in producing a functional and context-aware artifact. These results are consistent with DSR principles that emphasize the practical utility and empirical validity of research artifacts [28].

From a theoretical perspective, the findings reinforce the relevance of the OWASP Top 10 for LLM Applications as a primary conceptual lens for identifying Generative AI security risks that are not adequately addressed by classical web security paradigms. The dominance of information disclosure and prompt injection threats identified through STRIDE-based threat modeling suggests that security risks in AI-enabled websites are predominantly semantic and contextual, rather than purely structural, as highlighted in recent AI security literature [29]. The integration of STRIDE enables quantitative testing outcomes to be translated into systematically actionable threat categories, strengthening the argument that AI security testing requires cross-framework approaches, rather than a direct adaptation of traditional web security testing methods [30].

When compared with prior studies, the research results are aligned with the findings of previous research, identifying prompt injection as a dominant attack vector in web-based LLM applications. However, the research extends prior work by conducting evaluations on operational real-world websites, rather than limiting analysis to laboratory-based simulations [24]. In contrast to studies that focus on isolated mitigation techniques, such as output filtering alone, this research demonstrates that mitigation effectiveness improves significantly when a defense-in-depth strategy is employed. This finding contributes to the empirical discourse, which has previously been fragmented across conceptual and narrowly scoped studies [31]. Notably, these results diverge from certain conceptual arguments suggesting that rule-based mitigations are ineffective against adaptive LLM attacks, as the present study shows that rule-based mechanisms remain effective when combined with dynamic monitoring and iterative evaluation [3].

The primary scientific contribution of the research lies in extending the empirical understanding of Generative AI website security within a developing country context, which remains underrepresented in the global literature. By integrating adversarial red teaming, quantitative detection performance metrics, and qualitative threat modeling analysis, the research proposes a methodological model that is both replicable and adaptable for future research. In practical terms, the research contributes by delivering a framework that can be directly integrated into the development lifecycle of AI-enabled educational websites, thereby supporting the increasingly emphasized secure-by-design agenda in contemporary AI governance [32].

IV. Conclusion

The research concludes that Generative AI-based educational websites exhibit a security risk profile that is substantially different from conventional websites, with dominant vulnerabilities occurring at the level of semantic interaction and model output management. The empirical findings demonstrate that threats, such as prompt injection, information leakage, and context manipulation, constitute primary attack vectors that are not yet fully addressed by traditional security testing mechanisms. The development and application of a hybrid security testing framework grounded in DSR have proven effective in detecting and mitigating these vulnerabilities, as demonstrated by a 92.6% increase in detection accuracy, a 67.4% decrease in attack success rate, and consistent post-remediation system stability on both the sample website and the simulation prototype. The results support the research question through empirical evaluation on real sites.

From both scholarly and practical perspectives, the research reinforces the importance of integrating AI-specific security frameworks, including risk-based approaches and systematic threat modeling, within the context of Generative AI-enabled websites. The proposed framework not only extends theoretical understanding of LLM threat characteristics in web environments, but also provides actionable guidance that can be adopted by developers and educational institutions to implement secure-by-design principles in a more structured manner. In doing so, the research bridges the gap between conceptual AI security discourse and practical implementation requirements.

The implications of this research are multidimensional. For researchers, the findings open avenues for

future studies to apply similar frameworks in other sectors or to integrate them with continuous security testing approaches within MLOps pipelines. For practitioners, the results underscore the urgency of adopting AI-specific security testing prior to deploying Generative AI on public-facing websites and serve as a practical guide for Indonesian educational website developers. Finally, for policymakers, the findings support the need for national AI regulations as more operational and contextual AI security guidelines in Indonesia, in line with the global trend of strengthening AI regulations and digital data protection.

The various impacts of this research are as follows. For the theoretical impact, the research reinforces the integration of OWASP Top 10 for LLM Applications and the STRIDE framework in the context of developing countries, filling an empirical gap in previous AI security research that is limited to laboratory simulations. The research advocates adaptive testing for non-deterministic LLM systems, going beyond traditional web security approaches. For practical impact, this framework provides practical guidelines for university developers to mitigate prompt injection, data leakage, and unsafe output through input validation, semantic filtering, and rate limiting, reducing the risk of disruption to educational services and student data breaches. For Indonesia's education sector, amid rapid AI adoption at sites, these results reduce the vulnerability of strategic educational infrastructure to cyber exploitation, supporting national data security regulations with minimal latency overhead (0.19 seconds). For policy and industry impact, educational institutions can adopt this framework for compliance, while local developers gain an aligned testing checklist for cloud-based AI practices. The research facilitates secure AI deployment without hindering learning innovation.

Nevertheless, the research has several limitations that must be acknowledged. First, although the number of websites analyzed represents the educational context in Indonesia, the number is still relatively limited to only 15 sites, and caution should be exercised when generalizing these findings to other sectors, such as finance or health. Second, the evaluation is conducted within a constrained time frame (October-November 2025), which may not fully capture long-term threat dynamics such as model drift, adversarial adaptation by attackers, or the emergence of novel vulnerabilities in continuously evolving LLM deployed in production environments. Third, validation relies on interviews with a single AI website developer from an Indonesian educational institution, limiting the diversity of practitioner perspectives and potentially introducing bias toward local development practices. It represents a common limitation in applied security research where multi-stakeholder expert panels are resource-intensive but ideal for comprehensive validation.

Looking ahead, future research may broaden the scope of evaluation to other sectors characterized by higher data sensitivity, such as healthcare, finance, and public services, where the risks associated with Generative AI misuse may be more pronounced. Future studies may also develop continuous security testing mechanisms embedded within both the development and operational cycles of AI systems, enabling security assessment not only prior to deployment but also throughout system updates and real-world operation. Furthermore, investigating the automation of AI-driven security testing and its alignment with local governance and regulatory frameworks may contribute to the development of more adaptive, practical, and sustainable Generative AI security practices across diverse implementation contexts.

Acknowledgement

This research was supported by a grant from STIKOM Uyelindo Kupang in 2026 number: 9/LP3M/STIKOM-U/Penelitian/V/2026. The authors are indebted to the LP3M of STIKOM Uyelindo Kupang which provided a grant to assist with this research.

Author Contribution

Outline problems and develop conceptual ideas, F. M. H. T.; Collected the data, F. M. H. T. and R. I. N.; Contributed data or analysis tools, R. I. N.; Performed the analysis, F. M. H. T.; Wrote the paper, F. M. H. T. and D. S.; and Reviewed and improved manuscript, R. I. N. and D. S.

Data Availability

The data that support the findings of this research are available from the corresponding author, Fransiskus Mario Hartono Tjiptabudi, upon reasonable request because the data contain sensitive information.

References

[1] D. Russo, "Navigating the complexity of Generative AI adoption in software engineering," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 5, pp. 1–50, 2024.

[2] F. M. H. Tjiptabudi, "Integrated information and communication media based on Organization Goal-Oriented Requirement Engineering (OGORE)," *Jurnal Sistem Informasi*, vol. 19, no. 1, pp. 28–42, 2023.

[3] X. Zhang *et al.*, "Masked language model based textual adversarial example detection," in *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security*. New York, United States: Association for Computing Machinery, July 10–14, 2023, pp. 925–937.

[4] I. Jada and T. O. Mayayise, "The impact of artificial intelligence on organisational cyber security: An outcome of a systematic literature review," *Data and Information Management*, vol. 8, no. 2, pp. 1–15, 2024.

[5] E. A. Altulaihan, A. Alismail, and M. Frikha, "A survey on web application penetration testing," *Electronics*, vol. 12, no. 5, pp. 1–23, 2023.

[6] M. A. Ferrag *et al.*, "Generative AI in cybersecurity: A comprehensive review of LLM applications and vulnerabilities," *Internet of Things and Cyber-Physical Systems*, vol. 5, pp. 1–46, 2025.

[7] D. Ganguli *et al.*, "Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned," 2022. [Online]. Available: https://arxiv.org/abs/2209.07858

[8] F. P. Utama and R. M. H. Nurhadi, "Uncovering the risk of academic information system vulnerability through PTES and OWASP method," *CommIT (Communication and Information Technology) Journal*, vol. 18, no. 1, pp. 39–51, 2024.

[9] M. Podpora, M. Baranowski, M. Chopcian, L. Kwasniewicz, and W. Radziewicz, "LLM firewall using validator agent for prevention against prompt injection attacks," *Applied Sciences*, vol. 16, no. 1, pp. 1–24, 2025.

[10] L. Mauri and E. Damiani, "Modeling threats to AI-ML systems using STRIDE," *Sensors*, vol. 22, no. 17, pp. 1–21, 2022.

[11] M. A. Angganegara, I. Y. Mukti, and M. Fathinnuddin, "Integration of STRIDE and MITRE ATT&CK frameworks for enhanced cyber threat modeling: A case study of digital merchant banking application," in *2025 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS)*. Bandung, Indonesia: IEEE, Feb. 3–4, 2025, pp. 1–6.

[12] P. M. J. Delport, R. Von Solms, and M. Gerber, "Methodological guidelines for design science research," *Procedia Computer Science*, vol. 237, pp. 195–203, 2024.

[13] J. Da Assunção Moutinho, G. Fernandes, and R. Rabechini Jr., "Evaluation in design science: A framework to support project studies in the context of university research centres," *Evaluation and Program Planning*, vol. 102, 2024.

[14] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 3214–3252.

[15] M. S. Jabbar, S. Al-Azani, A. Alotaibi, and M. Ahmed, "Red teaming large language models: A comprehensive review and critical analysis," *Information Processing & Management*, vol. 62, no. 6, 2025.

[16] X. Liu and B. Zhong, "Integrating generative artificial intelligence into student learning: A systematic review from a TPACK perspective," *Educational Research Review*, vol. 49, pp. 1–34, 2025.

[17] M. M. A. Parambil *et al.*, "Integrating AI-based and conventional cybersecurity measures into online higher education settings: Challenges, opportunities, and prospects," *Computers and Education: Artificial Intelligence*, vol. 7, pp. 1–30, 2024.

[18] S. Shrestha, C. Banda, A. K. Mishra, F. Djebbar, and D. Puthal, "Investigation of cybersecurity bottlenecks of AI agents in industrial automation," *Computers*, vol. 14, no. 11, pp. 1–43, 2025.

[19] M. Uddin *et al.*, "Generative AI revolution in cybersecurity: A comprehensive review of threat intelligence and operations," *Artificial Intelligence Review*, vol. 58, pp. 1–39, 2025.

[20] Z. Azam, M. M. Islam, and M. N. Huda, "Comparative analysis of intrusion detection systems and machine learning-based model analysis through decision tree," *IEEE Access*, vol. 11, pp. 80 348–80 391, 2023.

[21] S. Potla, "Threat modeling in application security: A practical approach," *European Journal of Computer Science and Information Technology*, vol. 13, pp. 10–19, 4 2025.

[22] S. F. Wen, A. Shukla, and B. Katt, "Artificial intelligence for system security assurance: A systematic literature review," *International Journal of Information Security*, vol. 24, pp. 1–42, 2025.

[23] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, 2020, pp. 3356–3369.

[24] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned

language models," 2023. [Online]. Available: https://arxiv.org/abs/2307.15043

[25] K. F. Hasan, H. H. Shajeeb, C. Abeydeera, B. Turnbull, and M. Warren, "ISADM: An integrated STRIDE, ATT&CK, and D3FEND model for threat modeling against real-world adversaries," *IEEE Access*, vol. 13, pp. 217 316–217 348, 2025.

[26] D. Kang, X. Li, I. Stoica, C. Guestrin, M. Zaharia, and T. Hashimoto, "Exploiting programmatic behavior of LLMs: Dual-use through standard security attacks," in *2024 IEEE security and privacy workshops (SPW)*. San Francisco, CA, USA: IEEE, May 23, 2024, pp. 132–143.

[27] AWS, "AI security scoping matrix," n.d. [Online]. Available: https://aws.amazon.com/id/ai/security/generative-ai-scoping-matrix/

[28] S. Gregor and A. R. Hevner, "Positioning and presenting design science research for maximum impact," *MIS Quarterly*, vol. 37, no. 2, pp. 337–355, 2013.

[29] T. Shevlane *et al.*, "Model evaluation for extreme risks," 2023. [Online]. Available: https://arxiv.org/abs/2305.15324

[30] A. Shostack, *Threat modeling: Designing for security*. John Wiley & Sons, 2014.

[31] Z. Liao *et al.*, "Attack and defense techniques in large language models: A survey and new perspectives," *Neural Networks*, vol. 196, 2026.

[32] G. Recupito *et al.*, "Technical debt in AI-enabled systems: On the prevalence, severity, impact, and management strategies for code and architecture," *Journal of Systems and Software*, vol. 216, pp. 1–22, 2024.