NEXT-GEN PROMPT ATTACK DETECTION FOR LLMS USING INTEGRATED SMART SECURITY TECHNOLOGIES

MR.K.PREMKUMAR M.E(PH.D)
Professor & Head of Department, Computer Science and Engineering,
St. Joseph College of Engineering, Chennai-602117, Tamil Nadu,
Email Id – premkumarkrnd@gmail.com

Mr. HARRISON KESIN D,
Student, Computer Science and Engineering,
St. Joseph College of Engineering, Chennai-602117, Tamil Nadu,
Email Id – harrisondhk@gmail.com

Mr. MANOJ R ,
Student, Computer Science and Engineering,
St. Joseph College of Engineering, Chennai-602117, Tamil Nadu,
Email Id – manojrajendiran2910@gmail.com

Abstract -- Large Language Models (LLMs) have become a transformative technology in artificial intelligence, enabling advanced conversational systems, automated reasoning engines, content generation platforms, and enterprise-level decision support systems. Their ability to process natural language context and generate human-like responses has significantly enhanced automation across industries. However, this contextual flexibility introduces new security vulnerabilities in the form of prompt-based adversarial attacks.

Prompt injection, jailbreak manipulation, instruction override, and sensitive data extraction attempts exploit the probabilistic and instruction-following behavior of LLMs. Unlike traditional cybersecurity threats that target code-level weaknesses, prompt attacks operate at the semantic and contextual level, making them significantly harder to detect using conventional rule-based security systems.

This research proposes a Next-Generation Prompt Attack Detection Framework using Integrated Smart Security Technologies. The system combines structured preprocessing, transformer-based semantic embedding, and autoencoder-driven anomaly detection to identify malicious or abnormal prompt patterns in real time. By learning the behavioral distribution of normal prompts and computing reconstruction error thresholds, the framework detects suspicious inputs without requiring labeled attack datasets.

An integrated alert and logging mechanism provides real-time threat visualization through a web-based administrative dashboard. The proposed

architecture enhances reliability, scalability, adaptability, and operational safety of LLM deployments in enterprise and critical infrastructure environments.

I. INTRODUCTION

The evolution of transformer architectures has revolutionized artificial intelligence by enabling the development of Large Language Models capable of understanding long-range dependencies, contextual nuances, and complex reasoning tasks. These models are widely deployed in conversational assistants, automated report generation systems, intelligent tutoring platforms, healthcare triage tools, and enterprise AI copilots.

Despite their capabilities, LLMs present unique security challenges. Because they interpret instructions expressed in natural language, they can be manipulated through carefully crafted prompts that alter model behavior without modifying the underlying architecture. This vulnerability introduces a new category of threats known as prompt-based adversarial attacks.

Prompt injection attacks attempt to embed hidden instructions that override system-level policies. Jailbreak techniques encourage the model to ignore safety guidelines. Role manipulation attacks persuade the model to adopt unauthorized behavioral personas. Multi-stage conversational attacks gradually weaken alignment constraints across dialogue sessions.

Traditional security methods, such as keyword-based filtering and static rule enforcement, are insufficient in detecting such attacks because adversarial prompts can be paraphrased or contextually disguised. Moreover, many LLM deployments operate as external APIs, limiting the ability to modify internal model parameters.

Therefore, there is a critical need for an intelligent, adaptive, and scalable security framework capable of detecting abnormal semantic behavior before the prompt reaches the LLM processing stage. This paper introduces an integrated smart security architecture designed to protect LLM systems through contextual embedding analysis and unsupervised anomaly detection

II. BACKGROUND AND MOTIVATION

A. Growth of Generative AI Integration

The widespread adoption of generative AI has transformed digital ecosystems across industries. Organizations increasingly rely on LLMs for automating communication, summarizing documents, generating code, and assisting in analytical tasks. As reliance on these systems grows, so does the importance of ensuring their safe and secure operation.

B. Semantic Nature of Prompt-Based Attacks

Prompt-based attacks differ fundamentally from traditional cybersecurity threats. Instead of exploiting system vulnerabilities at the programming level, these attacks exploit the reasoning and contextual understanding capabilities of LLMs. Attackers craft inputs that subtly manipulate instruction hierarchy or attempt to override alignment constraints.

For example:

An attacker may instruct the model to "ignore previous instructions" or "act without restrictions." While these prompts may not contain explicit malicious keywords, their contextual meaning can significantly alter the model's response behavior.

C. Limitations of Existing Defense Mechanisms

Existing protective measures primarily rely on:

- Static blacklist filtering
- Hardcoded policy enforcement
- Manual moderation
- Simple pattern detection

These mechanisms fail when attackers use paraphrased instructions or indirect manipulation strategies. Furthermore, maintaining extensive rule sets becomes computationally inefficient and difficult to scale.

.

D. Motivation for Proposed Approach

The motivation for this research stems from the need to develop a security mechanism that:

- Learns normal prompt behavior automatically
- Detects unknown and zero-day prompt attacks
- Operates independently of proprietary LLM internals
- Supports real-time deployment environments

By integrating transformer-based semantic representation with anomaly detection modeling, the system aims to provide adaptive and robust LLM security.

III. NOVEL APPLICATIONS OF NEXT-GEN PROMPT ATTACK DETECTION FOR LLMS USING INTEGRATED SMART SECURITY TECHNOLOGIES

The proposed Next-Generation Prompt Attack Detection Framework introduces a novel and structured approach to securing Large Language Models (LLMs) against prompt-based adversarial threats. Unlike traditional filtering mechanisms that depend solely on static rule enforcement or keyword matching, this framework integrates semantic intelligence and behavioral modeling to detect malicious inputs at a deeper contextual level. The novelty of this work lies not merely in using machine learning techniques, but in the systematic integration of multiple intelligent layers to form a unified smart security architecture for LLM environments.

One of the key innovative aspects of the proposed system is the use of transformer-based contextual embeddings to analyze prompt semantics. Instead of evaluating prompts based on surface-level text patterns, the framework converts input text into high-dimensional embedding representations that capture contextual relationships, intent structure, and linguistic dependencies. This allows the system to detect malicious instructions even when they are paraphrased or disguised using indirect language. By focusing on semantic meaning rather than isolated keywords, the detection mechanism becomes significantly more robust against evolving adversarial strategies.

Another novel contribution of the framework is the implementation of an unsupervised autoencoder-based anomaly detection model for prompt validation. Traditional classification approaches require labeled datasets containing both malicious and non-malicious prompts. However, prompt-based attacks continuously evolve, making it difficult to maintain comprehensive labeled attack datasets. The proposed system overcomes this limitation by learning the behavioral distribution of normal prompts and identifying deviations using reconstruction error analysis. This enables detection of zero-day and previously unseen attack patterns without explicit supervision.

Furthermore, the architecture functions as an independent external security layer that does not require access to the internal architecture or attention weights of the LLM. This independence enhances practical applicability, especially for proprietary or API-based LLM systems where internal modifications are restricted. The security framework can be integrated seamlessly into existing AI pipelines as a pre-processing gateway, making it adaptable to both open-source and commercial deployments.

The integration of real-time alert generation and dashboard-based visualization further strengthens the practical significance of the proposed framework. Detected anomalies are logged with associated risk scores, timestamps, and classification categories. This not only improves monitoring efficiency but also enables organizations to analyze attack trends, evaluate system vulnerabilities, and refine defensive strategies over time. By combining detection, logging, and visualization in a single pipeline, the system transitions from passive monitoring to active threat intelligence.

Additionally, the modular design of the architecture ensures scalability and future extensibility. Each component—preprocessing, embedding generation, anomaly detection, classification, and alert management—can be independently upgraded or optimized. This modularity supports integration with reinforcement learning-based adaptive thresholds, hybrid detection systems, and cloud-based monitoring frameworks.

In summary, the novelty of the proposed framework lies in its intelligent fusion of semantic embedding analysis, unsupervised behavioral modeling, modular smart security layering, and real-time monitoring capabilities. The system moves beyond conventional rule-based filtering and establishes a proactive, adaptive, and scalable defense mechanism tailored specifically for securing modern LLM ecosystems.

IV. ROLE AND POTENTIAL OF NEXT-GEN PROMPT ATTACK DETECTION FOR LLMS USING INTEGRATED SMART SECURITY TECHNOLOGIES

Role:

A. Role in Proactive Security Enforcement – The proposed Prompt Attack Detection Framework plays a critical role as a proactive security layer in Large Language Model (LLM) environments. Unlike traditional moderation systems that analyze outputs after generation, this framework evaluates user prompts before they are processed by the language model. By intercepting malicious or suspicious inputs at the preprocessing stage, the system prevents harmful content generation, reduces policy violations, and minimizes the risk of adversarial exploitation. This proactive validation mechanism significantly enhances the safety and reliability of AI-driven applications.

B. Role in Alignment Protection and Behavioral Stability – Another essential role of the system is maintaining alignment integrity within LLM deployments. Modern LLMs are trained with safety and ethical alignment strategies to ensure controlled responses. However, prompt injection and jailbreak attacks attempt to override these constraints through carefully crafted instructions. The proposed framework detects such contextual

manipulation attempts by analyzing semantic anomalies and abnormal embedding behavior. This ensures that the model operates within its intended behavioral boundaries, preserving stability and consistency.

C. Role in Sensitive Data Protection – The system also serves as a safeguard for sensitive information processed by AI systems. In enterprise and institutional applications, LLMs may interact with confidential datasets, proprietary documents, or restricted system instructions. Malicious prompts may attempt to extract hidden data or induce unauthorized disclosures. By monitoring contextual deviations and suspicious semantic patterns, the framework reduces the probability of unintended information leakage and strengthens compliance with data protection regulations.

D. Role in Real-Time Monitoring and Threat Analysis – A further role of the framework is enabling real-time monitoring and anomaly tracking. The integrated logging and risk-scoring mechanism provides administrators with insights into suspicious prompt behavior. This facilitates continuous threat assessment, trend analysis, and detection of repeated attack patterns. The availability of structured anomaly reports enhances transparency and supports informed decision-making in AI governance.

E. Role in Strengthening AI Trust and Reliability – Ultimately, the proposed system plays a foundational role in building trustworthy AI ecosystems. By ensuring secure prompt interaction, preserving alignment mechanisms, and preventing adversarial manipulation, the framework enhances user confidence in LLM-based systems. Secure AI interaction is essential for widespread adoption, and the proposed architecture contributes significantly toward establishing robust and dependable generative AI infrastructures.

Potential:

The proposed framework demonstrates strong potential for future scalability and cross-domain integration. Its adaptability makes it suitable for diverse technological and industrial applications.

A. Enterprise-Level Deployment Potential – The proposed framework demonstrates strong potential for enterprise-level deployment across industries such as banking, healthcare, education, and e-commerce. As organizations increasingly integrate Large Language Models into customer support systems, document processing platforms, and decision-support tools, the need for reliable security mechanisms becomes critical. The framework can function as a preprocessing security gateway, ensuring that malicious or abnormal prompts are filtered before reaching the LLM. Its compatibility with existing infrastructures makes it highly practical for real-world commercial applications.

B. Compatibility with API-Based LLM Services – A major potential advantage of the proposed system is its independence from internal LLM architecture. Many commercial language models operate through API-based services where internal parameters are inaccessible. Since the framework operates externally using embedding analysis and anomaly detection, it can be integrated seamlessly with proprietary LLM platforms. This flexibility enhances its adoption potential in both open-source and commercial AI ecosystems.

C. Adaptability to Evolving Adversarial Threats – Prompt-based adversarial strategies are continuously evolving in structure and linguistic expression. The proposed system relies on semantic embedding analysis rather than static rule-based filtering, enabling it to detect zero-day and previously unseen attack patterns. With periodic retraining on updated prompt datasets, the framework can adapt dynamically to emerging threats, ensuring long-term sustainability and resilience in changing adversarial environments.

D. Expansion to Multi-Modal AI Systems – As AI technology progresses, future systems will increasingly handle multi-modal inputs including text, images, audio, and video. The embedding-based anomaly detection methodology can be extended to multi-modal representations, enabling cross-domain security monitoring. This scalability positions the proposed framework as a future-ready solution capable of evolving alongside advancements in generative AI technologies.

E. Cloud and Distributed Infrastructure Integration – Modern AI applications are commonly deployed in cloud-native and distributed computing environments. The modular architecture of the proposed framework allows it to be integrated into cloud pipelines, microservices, and API gateways. Its lightweight design supports real-time processing and scalability, making it suitable for large-scale deployments with high user interaction volumes.

F. Regulatory Compliance and AI Governance Support – With increasing global emphasis on AI governance and regulatory compliance, organizations must ensure transparency and accountability in AI operations. The logging, anomaly scoring, and monitoring capabilities of the framework provide traceable records of suspicious prompt activities. This supports audit processes, risk assessment, and compliance with emerging AI safety regulations, thereby strengthening responsible AI deployment practices.

G. Research and Innovation Opportunities – The framework also holds significant potential for academic and technological research advancement. It provides a foundation for exploring hybrid detection techniques, adaptive

threshold optimization using reinforcement learning, and federated anomaly detection models for distributed AI systems. Future enhancements can improve detection accuracy, reduce false positives, and strengthen overall AI security architectures.

V. INNOVATIVE INTEGRATION OF DEEP LEARNING AND NLP TECHNIQUES IN PROMPT SECURITY

A. NLP-Based Preprocessing and Linguistic Normalization – The first layer of the proposed framework integrates Natural Language Processing (NLP) techniques to preprocess and normalize incoming prompts before semantic analysis. Since prompts are expressed in natural language, they may contain noise, inconsistent formatting, redundant characters, or ambiguous linguistic structures. The system performs text cleaning, token normalization, and syntactic standardization to ensure consistent representation. This preprocessing stage enhances the quality of semantic embeddings generated in later stages. By applying NLP-based normalization, the framework reduces ambiguity and ensures that contextual meaning is preserved for accurate anomaly detection.

B. Contextual Embedding Generation Using Transformer Models – A major innovation in the framework is the use of transformer-based deep learning models to convert prompts into contextual embeddings. Unlike traditional bag-of-words representations, transformer architectures capture long-range dependencies, contextual relationships, and semantic intent within text. Each prompt is transformed into a high-dimensional vector representation that encodes its meaning in latent space. These embeddings serve as numerical representations of linguistic behavior. By leveraging deep learning–based contextual encoding, the system can detect subtle manipulations that may not be visible through surface-level keyword analysis.

C. Semantic Feature Extraction for Behavioral Modeling – After generating embeddings, the system extracts semantic features that characterize normal prompt behavior. This stage integrates NLP understanding with deep learning representation learning. The embedding vectors reflect linguistic intent, instruction hierarchy, tone, and contextual coherence. These features are used to model the distribution of legitimate prompt interactions. By learning semantic behavioral patterns, the system establishes a baseline for

detecting deviations. This combination of linguistic insight and neural representation enhances detection sensitivity against adversarial manipulation.

D. Deep Learning–Based Autoencoder for Anomaly Detection– The core detection mechanism of the framework relies on an autoencoder neural network trained on embeddings of legitimate prompts. The autoencoder compresses input vectors into a lower-dimensional latent representation and reconstructs them back to the original space. During training, the model learns the structural distribution of normal prompt embeddings. During inference, reconstruction error is computed using Mean Squared Error (MSE):

$$MSE = 1/n \sum(xi - x^{i})^2$$

Prompts that produce reconstruction errors exceeding a predefined threshold are classified as anomalous. This unsupervised deep learning approach enables detection of unknown or zero-day prompt attacks without requiring labeled malicious datasets.

E. Hybrid NLP–Deep Learning Interaction for Intent Analysis– The integration of NLP and deep learning allows the system to analyze prompt intent rather than isolated words. Adversarial prompts often attempt to override instructions indirectly through role manipulation or hidden directives. By examining contextual embedding shifts and linguistic dependency patterns, the framework detects suspicious intent transitions. This hybrid interaction between NLP linguistic processing and neural embedding analysis ensures deeper semantic understanding and improves robustness against sophisticated adversarial strategies.

F. Scalability Through Edge and cloud Integration – The architecture allows flexible deployment models:

- Edge deployment for low-latency intersection-level processing
- Cloud integration for centralized monitoring and analytics
- Hybrid models for distributed intelligence

This makes the system adaptable for both small-scale and city-wide implementation.

G. Adaptive Threshold Optimization and Continuous Learning– To enhance detection reliability, the framework incorporates adaptive threshold calibration mechanisms. Since prompt distributions may evolve over time, static thresholds can lead to false positives or false negatives. By periodically retraining the autoencoder with updated prompt datasets, the system dynamically adjusts its anomaly detection boundary. This continuous learning capability ensures that the integration of deep learning and NLP

remains responsive to changing adversarial environments. The adaptability strengthens long-term sustainability of the security architecture

H. Real-Time Detection and Smart Alert Integration Systems – The innovative integration extends beyond detection into real-time monitoring and visualization. Once an anomaly is detected, the system assigns a risk score based on reconstruction error magnitude and semantic deviation intensity. These results are logged and displayed in an administrative dashboard. NLP-based metadata analysis categorizes the type of suspicious behavior (e.g., instruction override, policy bypass attempt, or data extraction attempt). This intelligent reporting mechanism transforms the framework into an active security intelligence system rather than a passive filter.

I. Scalability and Future Research Expansion – The integration of deep learning and NLP within the proposed framework establishes a scalable architecture suitable for future expansion. The embedding-based detection mechanism can be extended to multi-modal AI systems incorporating text, image, and audio inputs. Additionally, future research may integrate reinforcement learning for dynamic policy enforcement or hybrid supervised–unsupervised models for improved accuracy. The modular design ensures that each component—NLP preprocessing, embedding generation, anomaly detection, and alert management—can be independently upgraded. This flexibility positions the framework as a sustainable and future-ready solution for securing generative AI ecosystems.

Output Moderation and Content Filtering Improvements – Significant progress has also been made in output-level content moderation systems. Modern LLM platforms employ toxicity detection, bias identification, and harmful content filtering using supervised classification models. While output moderation helps reduce visible harmful responses, it remains reactive in nature. It cannot prevent the internal reasoning pathway triggered by malicious prompts. The proposed system enhances security by acting at the input stage, complementing output moderation strategies and reducing risk before content generation occurs.

Semantic Embedding and Behavioral Modeling Techniques – Recent research trends emphasize semantic embedding analysis for detecting adversarial behavior. Transformer-based embedding models have enabled contextual representation of prompts in high-dimensional vector spaces. By modeling normal behavioral distributions of embeddings, anomaly detection systems can identify deviations indicating malicious intent. This approach aligns directly with the methodology of the proposed framework,

which integrates deep learning–based embedding analysis with unsupervised anomaly detection to identify prompt-based attacks.

Zero-Day Attack Detection and Unsupervised Learning Approaches – Traditional supervised security models depend on labeled datasets of known attack patterns. However, LLM-based adversarial strategies evolve rapidly, creating zero-day threats that are not present in training datasets. Recent advancements therefore focus on unsupervised and semi-supervised learning techniques for anomaly detection. Autoencoders, one-class classifiers, and clustering-based approaches have shown promising results in detecting previously unseen malicious inputs. The proposed system builds upon these advancements by employing an autoencoder-based reconstruction error mechanism for adaptive prompt anomaly detection.

VII.CHALLENGES

False Positive and False Negative Risk– One of the major challenges in the proposed Prompt Attack Detection Framework is balancing detection sensitivity and accuracy. Since the system relies on semantic embeddings and anomaly thresholds, highly creative but legitimate prompts may sometimes appear unusual in vector space representation. This can result in false positives, where valid prompts are incorrectly flagged as malicious. Conversely, subtle adversarial prompts that closely mimic normal behavior may bypass detection, leading to false negatives. Designing an optimal reconstruction error threshold and fine-tuning the autoencoder model is therefore critical to achieving reliable performance.

Evolving Adversarial Prompt Strategies – Prompt-based attacks are continuously evolving, with attackers employing paraphrasing, indirect instruction embedding, contextual manipulation, and multi-step conversational strategies to bypass security systems. This creates a dynamic threat environment where prompt structures change over time. As a result, the distribution of normal and malicious embeddings may shift, causing performance degradation if the model is not periodically retrained. Addressing concept drift and ensuring adaptive learning capabilities remains a significant technical challenge for long-term deployment.

Computational Complexity and Real-Time Processing – The integration of transformer-based embedding models and deep learning–based anomaly detection introduces computational overhead. Generating contextual embeddings and calculating reconstruction errors in real time requires optimized resource management, especially in high-traffic environments. In enterprise-scale applications with thousands of concurrent users,

maintaining low latency while preserving detection accuracy becomes challenging. Efficient model optimization, lightweight embedding strategies, and scalable deployment architectures are essential to mitigate this issue.

Limited Standardized Datasets for Evaluation – Unlike traditional cybersecurity domains, prompt injection and jailbreak attack datasets are relatively new and lack standardized benchmarks. This makes model evaluation, comparative analysis, and performance validation more complex. Creating a high-quality dataset that accurately represents normal prompt behavior while incorporating diverse adversarial examples requires careful curation and continuous updating. The absence of universally accepted datasets poses a challenge in demonstrating robustness across different LLM environments.

VIII.CONCLUSION

The proposed Prompt Attack Detection Framework presents a robust and intelligent security solution for safeguarding Large Language Models against prompt-based adversarial attacks. By innovatively integrating Natural Language Processing techniques with deep learning–based semantic embeddings and autoencoder-driven anomaly detection, the system enables proactive identification of malicious or abnormal prompts before they influence model behavior. Unlike traditional rule-based filtering mechanisms, the framework focuses on contextual understanding and behavioral modeling, allowing it to detect zero-day and evolving attack strategies effectively. Its modular architecture, real-time monitoring capability, and compatibility with API-based LLM deployments make it scalable and practical for enterprise, academic, and research environments. Overall, this work contributes toward building secure, reliable, and trustworthy generative AI systems by establishing a proactive, adaptive, and future-ready prompt security framework.

IX. REFERENCE

1. **A. Jamdade and B. Liu, "A pilot study on secure code generation with ChatGPT for web applications," in Proceedings of the ACM Symposium on Software Engineering (ACM SE), (New York, NY, USA), Association for Computing Machinery, 2024.**

2. **T. Liu, Y. Li, Z. Deng, G. Meng, and K. Chen, "Llm4shell: Discovering and exploiting rce vulnerabilities in real-world llm-integrated frameworks and apps." Briefing at Black Hat Asia 2024, 2024.**

3. **M. Lin, H. Zhang, J. Lao, R. Li, Y. Zhou, C. Yang, Y. Cao, and M. Tang, "Are your llm-based text-to-sql models secure? exploring sql injection via backdoor attacks," arXiv preprint arXiv:2503.05445, 2025.**

4. **S. Chen, J. Piet, C. Sitawarin, and D. Wagner, "Struq: Defending against prompt injection with structured queries," 2025. arXiv preprint arXiv:2402.06363 (accepted to USENIX '25).**

5. **R. Pedro, M. E. Coimbra, D. Castro, P. Carreira, and N. Santos, "Promptto-sql injections in llm-integrated web applications: Risks and defenses," in Proceedings of the 47th IEEE/ACM International Conference on Software Engineering (ICSE 2025), (Ottawa, Canada), pp. 1768–1780, IEEE Computer Society, 2025.**

6. **Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al., "Constitutional ai: Harmlessness from ai feedback," 2022. arXiv preprint arXiv:2212.08073, Accessed: 28 August 2025.**

7. **N. S. Dasari, A. Badii, A. Moin, and A. Ashlam, "Enhancing sql injection detection and prevention using generative models," arXiv preprint arXiv:2502.04786, 2025. Accessed: 28 August 2025.**

8. **G. Shen, A. Goeva, O. Kuchaiev, et al., "Nemo-aligner: A toolkit for efficient alignment of large language models," arXiv preprint arXiv:2405.01481, 2024. Accessed: 28 August 2025.**

9. **H. Lu, J. Liu, Z. Xu, H. Zhan, Z. Xu, Y. Wang, H. Chen, X. Wu, Y. Zhang, T. Zhang, et al., "Alignment and safety in large language models: Safety mechanisms, training paradigms, and emerging challenges," arXiv preprint arXiv:2507.19672, 2025.**

10. **S. Liu, W. Fang, Z. Hu, J. Zhang, Y. Zhou, K. Zhang, R. Tu, T.-E. Lin, F. Huang, M. Song, Y. Li, and D. Tao, "A survey of direct preference optimization," arXiv preprint arXiv:2503.11701, 2025.**

11. **T. Wu, L. Mei, R. Yuan, L. Li, W. Xue, and Y. Guo, "You know what i'm saying: Jailbreak attack via implicit reference, " arXiv preprint arXiv:2410.03857, 2024.**

12. **B. Arasteh, B. Aghaei, B. Farzad, K. Arasteh, F. Kiani, M. Torkamanian-Afshar, et al., "Detecting sql injection attacks by binary gray wolf optimizer and machine learning algorithms," Neural Computing and Applications, vol. 36, pp. 6771–6792, 2024.**

13. **D. Leung, O. Tsai, K. Hashemi, B. Tayebi, and M. A. Tayebi, "Xploitsql: Advancing adversarial sql injection attack generation with language models and reinforcement learning," in Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM), (Boise, Idaho, USA), 2024.**

14. **J. Zulu, B. Han, I. Alsmadi, and G. Liang, "Enhancing machine learning based sql injection detection using contextualized word embedding," in Proceedings of the 2024 ACM Southeast Conference, pp. 211 –216, 2024.**

15. **S. Sanguino, "Enhancing security in industrial application development," Applied Sciences, vol. 14, no. 9, pp. 3780–3794, 2024.**

16. **OWASP CRS Project, "Owasp modsecurity core rule set (crs)," tech. rep., OWASP Foundation, 2024. Implementation reference for baseline WAF configuration; not a peer-reviewed publication. Accessed: 28 August 2025.**

17. **OWASP Foundation, "Owasp top 10 for large language model applications,"**

tech. rep., OWASP Foundation, 2024. Official Technical Report. Accessed: 2025-05-30.