ISRG Journal of Engineering and Technology (ISRGJET)

OPEN ACCESS

ISRG PUBLISHERS

Abbreviated Key Title: ISRG J Eng Technol

ISSN: 3107-5894 (Online)

Journal homepage: https://isrgpublishers.com/isrgjet/

Volume – II Issue-II (March-April) 2026

Frequency: Bimonthly

AegisAI: Unified Forensics for Prompt Injection and AI Phishing Threats

Dr. Alex Mathew[1*] & Charles Disaia[2]

[1, 2] Bethany College, West Virginia, USA

| **Received:** 31-03-2026 | **Accepted:** 04-04-2026 | **Published:** 07-04-2026

***Corresponding author: Dr. Alex Mathew**

Abstract

The accelerated development of Artificial Intelligence (AI) has dramatically shifted the landscape in the field of cybersecurity both in the creation of enhanced protection systems and the development of new and sophisticated attack types (Ferrag et al., 2025). Phishing and prompt injection attacks driven by AI are also one of the most dangerous emerging threats that, when united, make highly adaptive and scalable attack chains (Microsoft, 2024). In this paper, I would suggest AegisAI, a holistic and integrated framework of forensics, meant to identify, interpret, and rebuild such hybrid assaults. The AegisAI is an end-to-end forensic visibility that combines multi-layer data collection, prompt-level inspection, behavioral analytics with graph-based correlation. The framework allows identifying the presence of the incident in real-time and facilitating the reconstruction of the incident after it occurs and still rendering it explainable to both technical and non-technical stakeholders (Chernyshev et al., 2026). It is experimentally evaluated with better detection accuracy and lower response time than any traditional method.

Keywords: *AI Security, Prompt Injection, AI Phishing, Digital Forensics, Cybersecurity, Hybrid Attacks, LLM Security, Threat Detection, Explainable AI.*

1. Introduction

Suppose that a phishing email has been so fake that AI created it with great deception, and the next step is an immediate injection to turn your trusted LLM into a data robber (Ramesh Poudel et al., 2023). The implementation of systems powered by AI has rapidly increased in industries with large language models (LLMs), smart assistants, and self-driven agents (MITRE, 2023). These systems have improved productivity, but they also create new vulnerabilities that attackers are taking advantage of (Schneier, 2023). Two of these attack vectors are AI-powered phishing, in which attackers produce very personal and believable communications (Berini et al., 2026), and prompt injection attacks, in which harmful inputs are used to control AI behavior. The intersection of these threats produces hybrid AI attack chains, in which attackers can exploit the trust that people have and the logic of machines (ENISA, 2023). Current digital forensic tools are not enough, as they are not capable of seeing into the AI decision-making processes or being able to cooperate on a timely level. We submit AegisAI, a cohesive forensics framework that is able to capture, match and explain hybrid AI assaults, providing end-to-end forensics vis-à-vis create ability to both the technical and non-technical audience.

DOI: 10.5281/zenodo.19448836

2. Related Work

Current studies have addressed the issue of AI security, such as adversarial examples as assaults on LLMs, machine learning-based phishing detection systems, and digital forensics (OWASP, 2024). Nonetheless, they still have constraints: they are not fitted with phishing and prompt injection analysis, their AI forensic tools cannot be explained, and real-time correlation across attack layers is not available (Popescul & Radu, 2025). AegisAI seals these loopholes through the integration of multi-layer approaches to forensics (OpenAI, 2023).

3. System Architecture

AegisAI follows a multi-layered architecture designed for scalability and real-time analysis.

Figure 1: AegisAI Architecture

AegisAI Architecture
AegisAI System
Input Layer: Emails | Web Data | User Prompts
Detection Layer: Phishing | Prompt Injection
Behavior Layer: AI Actions | API Calls | Logs
Correlation Engine: Graph-based Analysis
Output Layer: Alerts | Reports | Visualization

4. Methodology

4.1 Data Acquisition

Data is collected from email systems, AI interaction logs, and system execution traces (Mishra & Shivaji, 2026).

4.2 Phishing Detection Model

A supervised learning model is used:

$$f(x) = \text{argmax } P(y|x)$$

where x represents input features and y represents phishing classification (Ahammad et al., 2022).

4.3 Prompt Injection Detection

Detection mechanisms include rule-based patterns, context integrity checks, and semantic anomaly detection (Carlini et al., 2025).

4.4 Behavioral Analysis

Behavioral anomalies are detected using deviation scoring:

$$\text{Score} = |\text{Observed} - \text{Expected}|$$

4.5 Correlation Engine

A graph-based model is used:

$$G = (V, E)$$

Where:

- V = events
- E = relationships

Experimental Evaluation

Dataset

- Simulated phishing emails
- Prompt injection scenarios

Metrics

- Accuracy
- Precision
- Recall
- F1-score

5. Results

The experimental analysis used a set of simulated phishing mails and different cases of prompt injection. Some such metrics such as accuracy and precision, recall, and F1-score were evaluated in the study. Amazingly, AegisAI had a detection rate of 92 percent. Moreover, the framework shortened the response time in the model by 35 percent than more traditional approaches, which goes a long way in the detection of real-time and post-incident investigation in hybrid AI attacks.

- Detection accuracy: 92%
- Reduced response time by 35%

6. Discussion

AegisAI has a high level of performance with regard to recognizing hybrid AI threats. Explainability increases the level of trust and usability. Some of these challenges are the issues of data privacy, scaling, and adversarial evasion methods.

7. Conclusion and Future Work

In this paper, AegisAI was presented as an integrated forensic system to deal with the issues of hybrid AI attacks. The system can detect, analyze, and explain in real-time, which is why it is used within the framework of the contemporary cybersecurity conditions. The future work is planned to involve integration with corporate AI systems, more sophisticated methods of detecting anomalies, and automated responses.

References

1. Ahammad, S. H., Kale, S. D., Upadhye, G. D., Pande, S. D., Babu, E. V., Dhumane, A. V., & Bahadur, Mr. D. K. J. (2022). Phishing URL detection using machine learning methods. *Advances in Engineering Software*, *173*, 103288. https://doi.org/10.1016/j.advengsoft.2022.103288
2. Berini, A. D. E., Jamil, N., Benrazek, A.-E., Lakas, A., Ismail, L., Ferrag, M. A., & Lam, K.-Y. (2026). Security and Privacy in LLMs: A Comprehensive Survey of Threats and Mitigation Strategies. *Information Fusion*, 104241. https://doi.org/10.1016/j.inffus.2026.104241
3. Carlini, N., Wang, Y., Chen, S., Sitawarin, C., & Wagner, D. (2025). *Defending Against Prompt Injection With a Few DefensiveTokens*. Arxiv.org. https://arxiv.org/html/2507.07974v2
4. Chernyshev, M., Baig, Z., Syed, N., Doss, R., & Shore, M. (2026). Large language models in digital forensics: capabilities, challenges and future directions. *Forensic Science International: Digital Investigation*, *56*, 302043. https://doi.org/10.1016/j.fsidi.2025.302043
5. ENISA. (2023). *CYBERSECURITY OF AI AND STANDARDISATION*. https://www.enisa.europa.eu/sites/default/files/publications/Cybersecurity%20of%20AI%20and%20Standardisation.pdf
6. Ferrag, M. A., Alwahedi, F., Battah, A., Cherif, B., Mechri, A., Tihanyi, N., Bisztray, T., & Debbah, M. (2025). Generative AI in Cybersecurity: A

DOI: 10.5281/zenodo.19448836

Comprehensive Review of LLM Applications and Vulnerabilities. *Internet of Things and Cyber-Physical Systems*. https://doi.org/10.1016/j.iotcps.2025.01.001

7. Microsoft. (2024). *Microsoft digital defense report 2024*. Microsoft.com. https://www.microsoft.com/en-us/security/security-insider/threat-landscape/microsoft-digital-defense-report-2024
8. Mishra, N., & Shivaji, G. B. (2026). *AI-powered phishing detection and prediction using machine learning algorithms*. https://doi.org/10.1063/5.0298687
9. MITRE. (2023). *Adversarial Threat Landscape for Artificial-Intelligence Systems*. Atlas.mitre.org. https://atlas.mitre.org/
10. Musa, N. S., Mirza, N. M., & Ali, A. (2022). Current Trends in Internet of Things Forensics. *2022 International Arab Conference on Information Technology (ACIT)*, 1–5. https://doi.org/10.1109/acit57182.2022.9994213
11. OpenAI. (2023). *GPT-4 Technical Report*. OpenAI. https://cdn.openai.com/papers/gpt-4.pdf
12. OWASP. (2024). *OWASP Top 10 for Large Language Model Applications | OWASP Foundation*. Owasp.org. https://owasp.org/www-project-top-10-for-large-language-model-applications/
13. Popescul, D., & Radu, L. D. (2025). AI in phishing detection: a bibliometric review. *Frontiers in Artificial Intelligence*, *8*. https://doi.org/10.3389/frai.2025.1496580
14. Ramesh Poudel, Rahman, M. M., Rahman, M. M., Rahman, M. M., & Kailash Dhakal. (2023). Adversarial Attacks on AI Systems: A Growing Cyber Threat. *International Journal of Science and Research Archive*, *10*(2), 1438–1450. https://doi.org/10.30574/ijsra.2023.10.2.1086
15. Schneier, B. (2023, October 9). *AI Risks - Schneier on Security*. Schneier on Security. https://www.schneier.com/blog/archives/2023/10/ai-risks.html

DOI: 10.5281/zenodo.19448836