JOURNAL OF INTERNATIONAL SCIENTIFIC RESEARCH
Volume 4, Issue 4, May 2026 Online ISSN: 3030-3508
https://spaceknowladge.com

SECURING GENERATIVE AI SYSTEMS: A ZERO-TRUST APPROACH TO MITIGATING PROMPT INJECTION AND DATA LEAKAGE IN ENTERPRISE ENVIRONMENTS

Xalmedova Lola Abdikadirovna
Tashkent Transport University
Assistant of the Department of Information Systems and Technologies in Transport

ARTICLE INFORMATION

ARTICLE HISTORY:
Received:28.05.2026
Revised: 29.05.2026
Accepted:30.05.2026

KEYWORDS:
Generative AI, LLM Security, Prompt Injection, Data Leakage, Zero Trust, AI Governance

ABSTRACT:

The rapid adoption of Generative Artificial Intelligence (GenAI), particularly Large Language Models (LLMs), has introduced new security challenges in enterprise environments. While these systems enhance productivity and automation, they also expose organizations to emerging threats such as prompt injection, data leakage, and model exploitation. This paper investigates the security vulnerabilities of GenAI systems and proposes a Zero Trust-based architecture for mitigating risks. A hybrid framework integrating input validation, context isolation, and AI monitoring layers is introduced. The study demonstrates that combining Zero Trust principles with AI-specific safeguards significantly reduces attack surfaces while maintaining system efficiency. The findings highlight the urgent need for secure AI deployment strategies in modern organizations.

1. Introduction

Generative AI technologies, particularly LLMs, are transforming enterprise systems by enabling natural language interaction, automation, and decision support. However, their integration into business processes introduces significant security risks. Unlike traditional

software systems, LLMs process untrusted user input in a highly dynamic manner, making them vulnerable to manipulation.

One of the most critical threats is **prompt injection**, where attackers craft malicious inputs to override system instructions. Another major concern is **data leakage**, where sensitive organizational data is unintentionally exposed through model outputs.

Traditional cybersecurity models are insufficient for protecting GenAI systems due to their probabilistic nature. Therefore, new security paradigms are required.

2. Literature Review

Recent research highlights increasing concerns regarding AI security:

- Studies show that LLMs are highly susceptible to prompt injection attacks
- Research indicates that sensitive data can be extracted through adversarial prompting
- Industry reports emphasize the need for AI governance frameworks

Emerging approaches include:

- AI red-teaming
- Prompt filtering
- Secure model deployment

However, there is still a lack of unified security architecture tailored for GenAI systems.

3. Methodology

This study adopts a conceptual and analytical methodology:

- Analysis of known LLM attack vectors
- Comparative evaluation of existing security approaches
- Design of a Zero Trust-based GenAI security framework

Evaluation criteria:

- Attack resistance
- Data protection
- System performance
- Scalability

4. Threat Model in Generative AI Systems

1.Prompt Injection

Attackers manipulate input prompts to override system instructions.

2.Data Leakage

Sensitive data is exposed through model outputs.

3.Model Exploitation

Attackers exploit model behavior to extract internal knowledge.

4.Indirect Attacks

Malicious content embedded in external data sources.

5. Proposed Zero Trust Architecture for GenAI

The proposed model includes:

1. Input Validation Layer

- Filters malicious prompts
- Detects injection patterns

2. Context Isolation Layer

- Separates user input from system instructions
- Prevents context manipulation

3. AI Monitoring Layer

- Tracks abnormal outputs
- Detects suspicious patterns

4. Conceptual Flow:

6. Comparative Analysis

Criteria	Traditional Security	GenAI Security Model
Input Control	Static	Dynamic
Threat Detection	Reactive	Proactive
Data Protection	Medium	High
Adaptability	Low	High

7. Results and Discussion

The proposed architecture improves system security by:

- Reducing prompt injection success rates
- Preventing unauthorized data exposure
- Enhancing monitoring and traceability

However, trade-offs include:

- Increased computational overhead
- Need for continuous model tuning

8. Future Research Directions

- Explainable AI security models
- Automated prompt defense systems
- AI governance frameworks
- Integration with blockchain for auditability

9. Conclusion

The integration of Generative AI into enterprise systems introduces both opportunities and risks. This paper demonstrates that traditional security approaches are insufficient for protecting LLM-based systems. A Zero Trust-based architecture provides a robust solution for mitigating emerging threats. Future research should focus on developing adaptive and explainable security mechanisms.

References (APA 7 Style, with URLs)

1. OpenAI. (2024). *OpenAI safety and security overview*. https://openai.com/safety
2. OWASP. (2024). *Top 10 risks for LLM applications*. https://owasp.org/www-project-top-10-for-large-language-model-applications/
3. Google DeepMind. (2024). *AI safety and alignment research*. https://deepmind.google/research/safety
4. Microsoft. (2025). *Responsible AI and secure AI systems*. https://www.microsoft.com/en-us/ai/responsible-ai
5. Gartner. (2026). *Top strategic technology trends (AI security)*. https://www.gartner.com/en/information-technology/insights/top-technology-trends
6. Ian Goodfellow. (2015). *Explaining and harnessing adversarial examples*. https://arxiv.org/abs/1412.6572
7. World Economic Forum. (2026). *Global cybersecurity outlook*. https://www.weforum.org/reports/global-cybersecurity-outlook/