Proceedings of the 8th International
Scientific and Practical Conference
«Recent Advances in Global Science»
(August 26-28, 2026).
Vilnius, Lithuania

No
307

INFORMATION AND WEB TECHNOLOGIES

Engineering Approaches to Improving Security and Reliability in Generative Artificial Intelligence Systems

Babayev Mahammad Murad[1]

[1] Undergraduate Student in Computer Engineering
Baku Higher Oil School, Azerbaijan

Abstract. The rapid integration of generative artificial intelligence into software platforms, decision-support systems, information retrieval environments and autonomous applications has created a new class of engineering challenges. In addition to conventional software vulnerabilities, generative AI systems may produce factually incorrect information, follow maliciously injected instructions, reveal sensitive data, generate unsafe content or initiate unintended actions when connected to external tools. These characteristics make security and reliability inseparable requirements rather than independent design objectives. This article examines contemporary engineering approaches for improving the security and reliability of generative AI systems, with particular emphasis on large language model-based applications. The study adopts an engineering-oriented analytical review of recent research published mainly between 2024 and 2026. Prompt injection, jailbreaking, hallucination, excessive refusal, information leakage, unsafe tool use and component-level vulnerabilities are considered as major risk categories. A multilayer protection architecture is proposed that combines input validation, contextual separation, prompt-injection detection, model-level safety mechanisms, retrieval-based grounding, output verification, runtime action control, logging and continuous adversarial testing. Particular attention is paid to the trade-off between security and usefulness, since excessively restrictive guardrails can reduce the functional value of an otherwise safe system. A practical evaluation framework based on attack success rate, factual consistency, false refusal rate, utility retention, latency and auditability is also presented. The article argues that dependable generative AI cannot be achieved through a single safety filter. Instead, security and reliability should be treated as lifecycle-wide engineering properties supported by layered controls, continuous measurement and adaptive risk management.
Keywords: *generative artificial intelligence, large language models, AI security, reliability, prompt injection, jailbreak attack, hallucination, guardrails, trustworthy AI, runtime safety.*

Introduction. Generative artificial intelligence has moved from experimental language generation toward integration with software applications, enterprise information systems, search environments, programming tools and autonomous agents. This transition changes the nature of

INFORMATION AND WEB TECHNOLOGIES

the engineering problem. A stand-alone model that produces an inaccurate sentence may create an information-quality problem; the same model connected to databases, APIs, file systems or executable tools may transform an inaccurate interpretation into an operational security incident.

Traditional software engineering assumes that instructions and data can normally be separated by clearly defined syntax, interfaces and privilege boundaries. Large language models complicate this assumption because both instructions and external information are often represented in the same natural-language context. An apparently ordinary document, web page or retrieved database entry may therefore contain text that attempts to alter the behavior of the model. Recent work consequently treats prompt injection and jailbreaking not simply as undesirable model behavior, but as practical security problems affecting LLM-integrated systems [4].

Reliability presents a second challenge. Generative models can produce fluent and structurally convincing responses even when the underlying information is incomplete or incorrect. Hallucination is therefore particularly problematic in applications where users may interpret linguistic confidence as factual certainty. Production-oriented studies published in 2025 continue to identify hallucination detection and correction as essential requirements for dependable deployment [1].

Security mechanisms themselves can also introduce failure. A system designed to block every suspicious pattern may refuse legitimate requests, reducing usefulness and creating what is often described as over-defense. Recent prompt-injection research shows that protection must therefore be evaluated not only according to how many attacks are blocked, but also according to how accurately legitimate inputs are preserved [2-4]. These observations indicate that generative AI security cannot be reduced to a single moderation layer. The problem is better understood as an engineering assurance problem involving the model, application architecture, external data, tools, monitoring mechanisms and human oversight.The purpose of this article is therefore to examine recent approaches to generative AI security and reliability and to develop an integrated engineering framework for their practical application.

2. Research Approach. This study applies an engineering-oriented review of recent literature published mainly between 2024 and 2026. The analysis focuses on prompt injection,

jailbreak defense, hallucination mitigation, safety mechanisms, model-level protection and runtime control. The NIST Generative AI Risk Management Framework was also considered, particularly its recommendations related to governance, testing, content provenance and incident management [5]. The reviewed approaches were grouped into six areas: input protection, model-level security, grounding and verification, runtime and tool-use control, monitoring and incident response, and continuous adversarial evaluation.

3. Major Security and Reliability Risks. Generative AI systems face several key security and reliability risks. **Prompt injection** can manipulate a model by inserting malicious instructions, especially in retrieval-based and agentic systems [6]. **Jailbreak attacks** attempt to bypass built-in safety mechanisms, demonstrating that guardrails alone cannot provide complete protection. **Hallucinations** may lead to unsupported or inaccurate outputs, requiring grounding and response verification [7-10]. **Data leakage and component risks** may arise from external models, APIs, databases and third-party software, making access control and secure integration essential. At the same time, overly restrictive safety mechanisms may reject legitimate requests. Therefore, effective AI security should balance **protection, reliability and system utility** [9].

4. Engineering Approaches to Security and Reliability. The security and reliability of generative AI systems can be improved through a combination of input validation, trust classification and clear separation between instructions and external data. Layered guardrails, including input filters, prompt-injection detection, model-level safety mechanisms and output verification, provide stronger protection than a single control. Reliability can also be increased by grounding responses in trusted sources and verifying generated content before release. For AI agents capable of performing external actions, runtime authorization is essential, since generated decisions should not automatically be executed without additional control.

5. Risk-Control Matrix. Generative AI risks require different safeguards; therefore, effective protection depends on combining specialized controls rather than relying on a single defense mechanism.

INFORMATION AND WEB TECHNOLOGIES

Table 1.

Compact risk-control framework for generative AI systems

Risk	Typical failure	Main control	Metric
Prompt injection	Malicious instructions alter model behavior	Trust classification and prompt filtering	Attack Success Rate
Jailbreak	Safety restrictions are bypassed	Safety classifiers and model-level protection	Jailbreak Success Rate
Hallucination	Unsupported or incorrect output	RAG and evidence verification	Factual Consistency Rate
Data leakage	Sensitive information is exposed	Access control and output scanning	Leakage Rate
Unsafe / over-restricted behavior	Harmful actions or unnecessary refusals	Runtime authorization and balanced guardrails	Unsafe Action / False Refusal Rate

6. Evaluation of Security and Reliability. The security and reliability of generative AI systems can be assessed through several complementary indicators. Attack Success Rate (ASR) measures resistance to malicious attacks, False Refusal Rate (FRR) reflects unnecessary rejection of legitimate requests, Factual Consistency Rate (FCR) evaluates the accuracy of generated information, and Utility Retention (UR) shows whether safety mechanisms preserve normal system performance. Operational factors such as latency, cost, false detections, human intervention and auditability should also be considered. For integrated assessment, this study proposes the Generative AI Assurance Index (GAI-AI):

$$GAI\text{-}AI = w_1(1 - ASR) + w_2FCR + w_3UR + w_4AC$$

where AC represents audit coverage. The index combines security, reliability, usability and auditability in a single evaluation framework, while allowing the weights to be adjusted according to application requirements.

7. Discussion. Recent research shows a shift from model-centered safety toward system-level assurance in generative AI. Security depends not only on the model itself, but also on its interaction with external data, tools, users and application components. Since jailbreak and prompt-injection techniques continue to evolve, static protection is insufficient and should be supported by continuous testing, monitoring and control updates. The proposed L-GAISRA

architecture addresses this challenge through layered protection that balances security with system usability.

Conclusion. Generative AI systems require security and reliability to be treated as core engineering requirements. Key risks such as prompt injection, jailbreak attacks, hallucination, data leakage and unsafe tool use cannot be addressed by a single defense mechanism. Therefore, a layered approach combining input validation, model-level safeguards, output verification, access control and continuous monitoring is necessary. The proposed L-GAISRA architecture supports this approach by integrating multiple protection layers while maintaining system usability.

References:

[1] Bin, W., Jiazheng, Q., Yu, X., Hansen, H., Hao, Y., Gao, A., Wan, Z., Li, H., & Tsang, I. (2026). Safety Sidecar: Reflection-Driven Runtime Control for Safer Agents. Findings of the Association for Computational Linguistics: ACL 2026, pp. 30842-30856. Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.findings-acl.1542

[2] Chang, Z., Li, M., Huang, Y., Jiang, Z., Jia, X., Xiong, Q., Wang, J., Li, Z., & Wang, Q. (2026). Know Thy Enemy: Securing LLMs Against Prompt Injection via Diverse Data Synthesis and Instruction-Level Chain-of-Thought Learning. Findings of the Association for Computational Linguistics: ACL 2026, pp. 3540-3561. Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.findings-acl.173

[3] Wang, X., Jian, S., Li, S., Li, X., Li, Z., Ji, B., Wang, B., & Yu, J. (2026). JPU: Bridging Jailbreak Defense and Unlearning via On-Policy Path Rectification. Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics, pp. 7661-7674. Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.acl-long.348

[4] Chen, Y., Li, H., Zheng, Z., Wu, D., Song, Y., & Hooi, B. (2025). Defense Against Prompt Injection Attack by Leveraging Attack Techniques. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, pp. 18331-18347. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.897

[5] Li, H., Liu, X., Zhang, N., & Xiao, C. (2025). PIGuard: Prompt Injection Guardrail via Mitigating Overdefense for Free. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, pp. 30420-30437. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.1468

[6] Hackett, W., Birch, L., Trawicki, S., Suri, N., & Garraghan, P. (2025). Bypassing LLM Guardrails: An Empirical Analysis of Evasion Attacks against Prompt Injection and Jailbreak Detection Systems. Proceedings of the First Workshop on LLM Security (LLMSEC), pp. 101-114. Association for Computational Linguistics.

[7] Wang, S., Wang, X., Mei, J., Xie, Y., Chen, S.-Q., & Xiong, W. (2025). Developing a Reliable, Fast, General-Purpose Hallucination Detection and Mitigation Service. Proceedings of NAACL 2025: Industry Track, pp. 971-978. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-industry.72

[8] Xu, Z., Liu, Y., Deng, G., Li, Y., & Picek, S. (2024). A Comprehensive Study of Jailbreak Attack versus Defense for Large Language Models. Findings of the Association for Computational Linguistics: ACL 2024, pp. 7432-7449. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.443

[9] Xie, Y., Fang, M., Pi, R., & Gong, N. (2024). GradSafe: Detecting Jailbreak Prompts for LLMs via Safety-Critical Gradient Analysis. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pp. 507-518. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.30

INFORMATION AND WEB TECHNOLOGIES

[10] Zhao, W., Li, Z., Li, Y., Zhang, Y., & Sun, J. (2024). Defending Large Language Models Against Jailbreak Attacks via Layer-Specific Editing. Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 5094-5109. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.293