

Securing Agentic AI: Emerging Security Threats, Vulnerabilities, and Defensive Mechanisms

Pratik Vishnu Bahirwal¹, Aarti Sanjay Gawai²

¹Student, Department of Information Technology and Computer Science, D.G. Ruparel College, Mumbai, Maharashtra, India

²Assistant professor, Department of Information Technology and Computer Science, D.G. Ruparel College, Mumbai, Maharashtra, India

¹Pratikbahirwal@gmail.com, ²aartigawai1611@gmail.com

ABSTRACT

The increasing development of artificial intelligence (AI) resulted in the appearance of agentic AI systems which could reason, plan, make decisions, communicate and interact with external tools and environments autonomously. In addition to allowing the agent to perform complicated operations with a minimum amount of human intervention, the mentioned features are a source of new threats and vulnerabilities related to security and privacy.

Agentic AI differs from traditional AI applications by gaining access to sensitive information, using external tools, having memory, and carrying out multi-step actions. This paper explores several kinds of threats associated with agentic AI (e.g., prompt injection, data poisoning, sensitive information disclosure, privilege escalation), as well as the vulnerabilities which allow those threats to happen. The defensive methods which protect agentic AI systems against different attacks are analyzed too. As multiple vulnerabilities might appear at different architectural levels (models, memory, tools, data, execution environment), the defense-in-depth strategy is necessary for ensuring the security of agentic systems.

Key Words: agentic AI, artificial intelligence, ai security, autonomous agents, prompt injection, vulnerability, cybersecurity, defensive mechanisms.

1. INTRODUCTION

AI has been evolving towards increasingly autonomous systems that are able to perform reasoning, planning, and interacting with the environment independently of a human being. Agentic AI stands for the paradigm in which AI systems are able to pursue the set objectives autonomously by processing information, making

decisions, invoking tools, and performing multi-step actions. The combination of LLMs with tools, memory, planning and application programming interface (API) has expanded the functionality and applicability of agentic AI systems which can be used in software engineering, healthcare, finance, education, and cybersecurity. However, an increase in the level of autonomy of agentic AI systems results in a wider attack surface compared with traditional AI applications. An agent may process untrusted inputs, receive information from an external source, have memory, invoke tools, access sensitive information, and perform actions on behalf of users/organizations. Therefore, vulnerabilities may emerge not only inside the AI model but also at data, memory, tool, communication and execution layers. Prompt injection, indirect prompt injection, data poisoning, sensitive information disclosure, insecure tool usage, excessive permissions, privilege escalation, and supply-chain attacks may result in manipulating or causing unintended actions of an agentic AI system. Prompt injection is especially critical in the case of agentic AI since it uses external sources of information (documents, web sites, emails, other apps) as part of its reasoning. Malicious commands embedded in such sources of information may potentially lead to the manipulation of an agentic AI agent to ignore the set objectives or perform unauthorized actions. Also, the vulnerability associated with excessive permissions and agency may be dangerous for security and privacy since the attacker may manipulate such an agent and lead to consequences worse than incorrect content generation (unauthorized access, data exposure, resource modification, harmful operations). Nevertheless, traditional mechanisms should be adjusted to the specifics of the autonomous AI system. Defenses may consist of least-privilege access control, authentication and authorization, sand-boxing, tool isolation, input and output validation, secure memory management, runtime monitoring, anomaly detection, human permission for impact operations, and defense-in-depth architecture. Security assessment should cover all phases of the agent lifecycle, including model development, deployment, integration with tools, runtime execution, and continuous monitoring. Hence, securing agentic AI calls for the application of the traditional cybersecurity techniques

together with those specifically addressing the issues related to autonomous AI's operations. This paper will study the security threats arising from the emergence of agentic AI, discuss vulnerabilities of various components of the agent architecture, and explore defensive mechanisms that help mitigate the threats.

2. EMERGING SECURITY THREATS

Agentic AI introduces security threats that go beyond the traditional attacks on AI models. As agents are capable of autonomously processing the information, interacting with the tools and other resources, and executing actions, the attackers can target various stages of the agent's operation. The major emerging security threats are described below.

2.1 Prompt Injection Attacks

The most serious security threat to any AI agent is prompt injection. In this attack, the adversary sends malicious instructions in order to change the AI's actions or its goals. These instructions can be either placed in the user's prompt directly or via some other sources, e.g., documents, websites, email messages, and information collected by the agent. If the agent takes malicious instructions for valid ones, it may execute undesirable actions or disclose confidential information.

2.2 Indirect Prompt Injection

Indirect prompt injection is the attack where the instructions to perform malicious actions are placed not in the prompt but in some other source of information. For instance, the hacker may embed instructions in a website, document, or database record that will be processed by the agent at some point in time. Since the agent does not distinguish between the trustworthy instructions and the other data, it may influence the future decisions and operations performed by the agent.

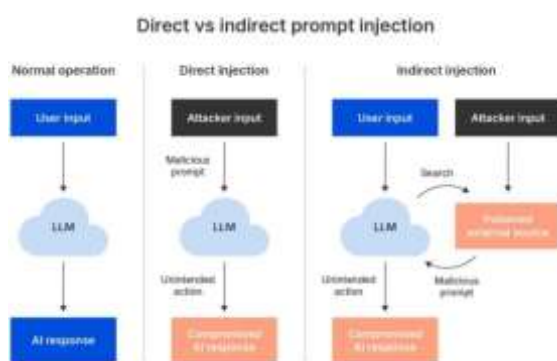


Fig-1 : Direct vs Indirect Prompt injection

2.3 Excessive Agency and Unintentional Actions

Often, agentic AI systems need access to tools and services in order to accomplish the assigned task. When the permission given to an agent is too broad, then this gives room for the agent to perform actions which go beyond what was expected from the agent initially. An agent which can be manipulated or hacked can use such permissions to change files, access confidential information, make communications, and perform transactions. This increases the severity of the consequences in case of a security breach associated with AI.

2.4 Data Poisoning and Manipulation

Data poisoning entails the inclusion of malicious or fake information in the dataset, knowledge base, memory system, or any other source of information available to the agent. The agent uses the information retrieved from the source to plan and decide how to act. In addition, if the information source is persistent memory, then the malicious information becomes a greater threat since it will continue to be used in further interactions.

2.5 Sensitive Information Disclosure

An agentic system may have access to confidential information of the users of the organization, documents, credentials, or any other proprietary knowledge. Handling the information in the wrong way leads to information leakage.

2.6 Tool and API Abuses

AI agents routinely interface with various APIs, databases, software applications, and other tools. The attacker might abuse vulnerabilities of tool interfaces or make the agent invoke a tool using malicious parameters. In consequence, a poorly secured tool can act as an avenue through which the attacker would influence the agent or access secure resources.

2.7 Supply-Chain Attacks

As is mentioned above, agentic AI systems are likely to utilize third-party models, libraries, plugins, APIs, datasets, and tools. Compromised third party in this case can inject malicious capabilities or vulnerabilities into the entire AI system. As is evident from the nature of agent ecosystems, supply chain security becomes one of the aspects to consider when developing and deploying an agentic AI system.

3. VULNERABILITIES IN AGENTIC AI SYSTEMS

The threats discussed above can exploit vulnerabilities that occur in different layers of an agentic AI system. Such vulnerabilities may arise from the AI model itself,

memory mechanisms, tool interfaces, access control mechanism, data sources, or execution environments.

3.1 Model-Level Vulnerabilities

Vulnerabilities of the AI model may include susceptibility to adversarial input, adversarial instructions, hallucination, and inadequate separation of system instructions from external information. As was mentioned earlier, since an agent relies on model output to make decisions, such vulnerabilities may affect other parts of the system.

3.2 Memory and Context Flaws

Agents may utilize short-term or persistent memory to preserve the context through interactions. Vulnerable memory allows attackers to inject malicious data that would later affect decision-making. Sensitive data kept in memory could be compromised if proper access controls, encryption, and management policies are not enforced.

3.3 Tool-Use Vulnerabilities

The ability to call third-party tools constitutes a major security barrier. Lack of proper validation of tool inputs, inadequate authorization, and uncontrolled access to tools can result in unintended execution. The agent itself can choose an inappropriate tool or pass incorrect parameters due to misleading instructions and wrong reasoning.

3.4 Identity and Authorization Flaws

Identity and authorization mechanisms that are not sufficiently enforced allow agents to gain access to resources without proper constraints. Wide permissions granted to an agent increase the risks of accidents as well as successful attacks. Thus, authorization mechanisms should check whether a particular operation is allowed for the agent, user, resource, and current situation.

3.5 Vulnerabilities Relating to Data and Information Retrieval

Retrieval agents can receive data from external sources that could be untrustworthy, out-of-date, manipulated or even harmful. The inability of the retrieval methods to differentiate between trustworthy and non-trustworthy sources exposes the reasoning component of the process to an indirect attack. Moreover, poor verification of the retrieved data leads to erroneous and dangerous conclusions.

3.6 Execution Environment Vulnerabilities

Computational environments have to be created in order for AI agents to perform their actions by running various tools, scripts, applications or workflows. An inadequately secured computational environment allows a

compromised agent to access operating system resources, files, network systems or other applications. Consequently, the impact of compromising an agent becomes higher.

3.7 Gaps in Monitoring and Auditability

Agentic artificial intelligence systems are usually multi-step and autonomous. Thus, it is hard to find out why a certain action was taken. Poor logging, monitoring and auditing lead to inability to find a possible attack and investigate it. Moreover, the lack of transparency prevents administrators from detecting abnormal agent behavior.

4. DEFENSIVE MECHANISMS FOR SECURING AGENTIC AI

Agentic AI protection entails using a layered approach that is built on both AI-focused techniques and conventional cybersecurity measures. The defensive approaches would have to be designed to offer security not only to the AI model but also the data, memory, tools, communication methods, execution environment, and all the while ensuring human supervision.

4.1 Least-Privilege Access Control

The idea of least privilege should also be considered in terms of permissioning AI agents. Each agent should only have just enough permissions that it needs to complete a given task. Permissions would be limited to the user, task, resource, and the overall context of the activity. This will minimize the effect of any compromised or manipulated agents.

4.2 Strong Authentication and Authorization

Strong authentication and authorization policies should also be put in place not only for the users but the AI agents as well. All operations that are carried out must be linked to a given authenticated identity as per the authorization policy.

4.3 Prompt and Input Validation

Input validation can assist in detecting malicious instructions that may affect the decision-making process of the agent. It is important to differentiate between trusted instructions and untrusted external inputs. Further validation checks may be performed on the user prompts, extracted information, tool parameters, and commands issued by the agent.

4.4 Sandbox and Tool Isolation

A sandboxing solution offers an isolated environment where possibly dangerous actions of the agent can be performed. Isolation of tools and limitation of the network, file system, and process access permissions are preferred whenever possible. Tool isolation can mitigate the risks caused by malicious instructions or unexpected actions of the agent.

4.5 Secure Memory Management

Information stores need to provide access controls, encryption if required, data validation, and retention management. The incoming data that needs to be stored in the persistent memory has to be checked for validity. Unnecessary retention of sensitive information is not recommended, as well as identification, modification, or removal of malicious or stale memory records.

4.6 Runtime Monitoring and Anomaly Detection

Continuous monitoring can detect any deviations from expected behavior, unauthorized use of tools, unusual access to data, and any other anomalous activity. Behavioral analysis and anomaly-detection systems can trigger alarms if any deviations from the norm are detected by the agent. In addition, comprehensive logging facilitates post-incident forensics.

4.7 Human-in-the-Loop Controls

Human intervention is especially critical in cases when the actions taken by the agent might lead to irreversible results. Operations like making payments, deleting crucial data, changing security configurations, or communicating with external parties might need human intervention beforehand. "Human-in-the-loop" methods add an extra layer of control when the automatic decision may have severe consequences.

4.8 Output Validation and Action Verification

The output of the agent and the actions initiated should be verified prior to their execution. Instead of allowing an agent to automatically execute all the commands that it generated, there should be a layer for enforcement of security policies to see whether the operation is safe.

4.9 Defense-in-Depth Architecture

No matter how advanced any particular security measure may be, it is not enough to ensure adequate protection of an autonomous AI system from all possible threats. The "defense-in-depth" security architecture involves authentication, authorization, input validation, sandboxing, monitoring, secure memory, tool isolation, and human intervention. Different security layers decrease the probability that the breakdown of one of them will cause the complete failure of the whole system.

4.10 Continuous Security Testing and Evaluation

Agentic AI systems should be subjected to continuous security evaluation. Such tests may include red-team exercises, adversarial testing, vulnerability assessment, attack simulation, and other types of security audits. This is especially critical because the behavior of an agentic AI system may vary depending on changes in its models, tools, data sources, configuration, etc.

5. CONCLUSION

Agentic AI is a next-level development in the field of artificial intelligence because of the ability of these systems to reason, plan, use tools, gather information, memorize facts and carry out actions. Nevertheless, such functionality brings considerable challenges for security and privacy of agentic AI systems. Agentic AI systems can operate with an environment and use various tools and APIs. They can access sensitive information and execute complex actions consisting of several steps. Thus, any vulnerability in the system may lead to the leakage of data and compromise the overall security of an agentic AI system. In this research, the most important threats related to agentic AI were analyzed. Direct and indirect prompt injection, data poisoning, sensitive information disclosure, excessive agency, privilege escalation, insecure tool usage, and supply-chain attack were selected as examples. At the same time, different types of vulnerabilities were analyzed including those associated with AI model, memory and context, tools and APIs, authentication and authorization, data and retrieval systems, execution environment, and monitoring. Recently, NIST highlighted the significance of identity, authorization, auditing, and controls for AI agents' interaction with data, tools, and applications. As the results of analysis show, the problem of agentic AI security cannot be solved using one security measure. Thus, such measures as least privilege access, proper authentication and authorization, input/output validation, sandboxing and isolation of tools, memory management, runtime monitoring, anomalies detection, actions verification, and human monitoring are critical. The point is that the permissions and capabilities of an agent play an important role in ensuring agentic AI security since the environment and tools can increase the potential impact of an action. NIST emphasizes the importance of least privilege, agent identity, authorization, auditing, and prompt injection mitigation for agentic AI security. As such, it is essential that the security of agentic AI is evaluated at all stages of the lifecycle – from design, development, implementation to monitoring and ongoing security testing. With agentic AI becoming increasingly competent and autonomous, future research should concentrate on developing standardized tests of security, strong systems of identity and authorization of agents, architecture of memory and tools used, detection of

indirect attacks, and approaches to ensuring meaningful human control. In summary, agentic AI systems have significant promise when it comes to supporting highly complex and autonomous tasks; however, their use requires proper security and privacy measures. Proper technical controls, access limitations, monitoring, proper architecture of the systems in question, and human supervision are likely to decrease the danger posed by the use of autonomous AI systems.

ACKNOWLEDGEMENT

Finally, the authors would like to extend their sincere thanks to all those who helped us with our research either directly or indirectly. We would like to thank our faculty members and academic mentors for their guidance, useful advice, and constant encouragement throughout the process. We would also like to express our sincere appreciation to the researchers, institutions, and organizations that have provided valuable resources concerning artificial intelligence, cybersecurity, privacy, and security of agentic AI through their public publications and other freely available materials. Last but not least, we are thankful to our university, colleagues, friends, and family members for their support and encouragement in our efforts

REFERENCE

1. Srinivasan, J., Regi, S. A., Anbarasan, A. K., Suresh, A., Vetrivel, T., & Venu, S., "Detection and analysis of prompt injection in Indian multilingual large language models," *Scientific Reports*, vol. 16, 16208, 2026, doi: 10.1038/s41598-026-43883-0. — Vellore Institute of Technology, India.
2. Balakrishnan, P., & Leema, A. A., "Vulnerabilities and Defenses: A Monograph on Comprehensive Analysis of Security Attacks on Large Language Models," *Indian Journal of Information Sources and Services*, 2025, doi: 10.51983/ijiss-2025.IJISS.15.2.54. — Vellore Institute of Technology, India.
3. Mathew, E. S., "Enhancing Security in Large Language Models: A Comprehensive Review of Prompt Injection Attacks and Defenses," *Journal on Artificial Intelligence*, vol. 7, pp. 347–363, 2025, doi: 10.32604/jai.2025.069841. — Motilal Nehru National Institute of Technology Allahabad, India.
4. Thakur, V., & Garg, U., "Defending against prompt injection attacks in LLMs," *AIP Conference Proceedings*, vol. 3297, 050021, 2025, doi: 10.1063/5.0286966. — Delhi Technological University, New Delhi, India.
5. Jain, B., Pawar, P., Gada, D., Patwa, T., Kanani, P., Patil, D., & Kurup, L., "Comprehensive Analysis of Machine Learning and Deep Learning Models on Prompt Injection Classification Using Natural Language Processing Techniques," *International Research Journal of Multidisciplinary Technovation*, vol. 7, no. 2, pp. 24–37, 2025, doi: 10.54392/irjmt2523. — Dwarkadas J. Sanghvi College of Engineering, Mumbai, India.
6. Teckani, H., Pandya, D., & Makwana, H., "Secure Artificial Intelligence (SAI): A Dual-layer Defence Model Against Prompt Injection and Prompt Poisoning Attacks," *International Journal for Multidisciplinary Research*, vol. 7, no. 1, 2025, doi: 10.36948/ijfmr.2025.v07i01.35371. — Silver Oak University, Ahmedabad, India.
7. Saini, L., Motwani, Y., Singh, S., & Patil, S., "Exploring the Landscape of AI Security: A Taxonomy and Review of Jailbreaking and Prompt Injection," *NCTAAI 4.0 Conference Proceedings*, 2025. — D. Y. Patil International University, Pune, India.
8. Honawade, S. T., Sachin, I. C., Mangala, L., Bharath, S. V. S., & Kumara, S., "Securing LLMs Against Prompt Injection: A Hybrid Rule-Based and Machine Learning Detection Framework," *Journal of Emerging Technologies and Innovative Research (JETIR)*, vol. 12, no. 11, 2025. — Dayananda Sagar College of Engineering, Bengaluru, India.
9. Dhande, S. N., & Dhande, D. S., "Prompt Security and Bias Mitigation Techniques in Large Language Models," *International Journal of Engineering Research & Technology (IJERT)*, vol. 14, no. 2, 2026, doi: 10.17577/IJERTCONV14IS020157. — Ashoka Center for Business and Computer Studies, Nashik, India.
10. Kanthale, R., Nagar, H., Shah, A., & Abhang, R., "Automated Software Vulnerability Detection using Large Language Models: A Comparative Study from Cloud to Edge," *Open Access International Journal of Science and Engineering*, vol. 9, no. 1s, pp. 99–106, 2026, doi: 10.65521/oaijs.v9i1s.3609. — Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, India.
11. https://promon.io/security-news/prompt-injection-attacks-emerging-critical-risk-mobile-app-security?utm_source=chatgpt.com
12. Booth, H., Fisher, W., Galluzzo, R., & Roberts, J. (2026). *Accelerating the Adoption of Software and Artificial Intelligence Agent Identity and Authorization: Concept Paper*. National Institute of Standards and Technology (NIST), National Cybersecurity Center of Excellence.
13. A. Chhabra, S. Datta, S. K. Nahin, and P. Mohapatra, "Agentic AI security: Threats, defenses, evaluation, and open challenges," *IEEE Access*, 2026, doi: 10.1109/ACCESS.2026.3675554.
14. S. Arora and J. Hastings, "Securing agentic AI systems—A multilayer security framework," in *Proc. 2025 IEEE 5th Int. Conf. on Robotics*,

- Automation and Artificial Intelligence (RAAI)*, 2025, doi: 10.1109/RAAI67517.2025.11423374.
15. V. S. Narajala and O. Narayan, "Securing agentic AI: A comprehensive threat model and mitigation framework for generative AI agents," *arXiv preprint arXiv:2504.19956*, 2025, doi: 10.48550/arXiv.2504.19956.
16. "Securing agentic AI: A policy-driven governance architecture for cloud security workflows," in *Proc. 2025 7th International Academic Exchange Conference on Science and Technology Innovation (IAECST)*, 2025, doi: 10.1109/IAECST68792.2025.11415118.
17. "Securing tool-using AI agents against injection and authority misuse," *Computation*, vol. 14, no. 5, Art. no. 98, 2026, doi: 10.3390/computation14050098.
18. D. Pant and S. Lohani, "Securing Agentic AI: A Comprehensive Threat Analysis of Model Context Protocol Systems with Layered Defense Strategies," *Preprints*, 2025, doi:10.20944/preprints202510.2087.v1.
19. V. Gupta Unfiltered, "How to Secure AI Agents & MCP Servers with Zero Trust: Full Hands-On Tutorial," *YouTube*, Sep. 5, 2026.
20. C. T. M. Praveen Kumar, S. H. G., S. R. Kulkarni, Y. S. Kotian, and R. B. S., "Agentic AI—Email Classification Using XGBoost, Ollama (gemma3:1b) with Interpretable Dashboard," *International Journal of Scientific Research in Engineering and Management (IJSREM)*, vol. 10, no. 3, Mar. 2026, doi: 10.55041/IJSREM57339.

BIOGRAPHIES



Mr. Pratik Vishnu Bahirwal
Msc. Computer Science,
Student at D.G. Ruparel college, Mumbai,
Maharashtra, India



Ms. Aarti Sanjay Gawai
Asst. Prof., Dept. of IT/CS,
D. G. Ruparel College, Mumbai,
Maharashtra, India