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-I (January – February) 2026

Frequency: Bimonthly

From Conversation to Command Execution: A Comparative Threat Modeling and Risk Analysis of OpenClaw and ChatGPT

Dr. Alex Mathew

Bethany College, West Virginia, USA

| **Received:** 17-02-2026 | **Accepted:** 25-02-2026 | **Published:** 28-02-2026

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

Abstract

The development of large language model (LLM)-based software has resulted in two trends: cloud-based conversational systems like ChatGPT and self-hosted autonomous agent systems like OpenClaw. Although they both use generative AI, their implementation authority, level of trust, and cybersecurity risk appear to vary significantly. In the given paper, the structure of the comparative analysis of three areas is conducted, that is, security architecture, paradigm of the threats modeling of the Structured Threat Reduction Inventory, and the qualitative risk measurement. Based on recent AI risk management (AI, 2024) and new studies on the weakness of LLMs (Dong et al., 2025; Gulyamov et al., 2026), we show that the autonomous agents change AI risk by transforming information-layer vulnerability to system-layer implementation threats that drastically change enterprise security posture needs. A comparison matrix framework and visualization of mitigation are given based on secure deployment decisions.

Keywords: *Autonomous AI agents, conversational AI, LLM security, STRIDE, prompt injection, zero trust, risk analysis, command execution security, OpenClaw, ChatGPT.*

1. Introduction

The current LLC systems have evolved into more than conversational assistance, into automation and execution of tasks. ChatGPT and similar cloud-based conversational AI are mainly prompt input and response output programs. This ability is expanded by autonomous agents such as OpenClaw, which can execute instructions, communicate with local systems, and maintain persistent operational memory. This difference creates radically new sources of security issues. Whilst conversational AI risks focus on data privacy and abuse, there are integrity and availability risks introduced by autonomous agents because they have system-level jurisdiction. AI-established command and control systems will have inherently different risk engineering solutions compared to passive analytical solutions.

2. Security Architecture Overview

2.1 OpenClaw Security Architecture

OpenClaw is an autonomous agent that is self-hosted in the host system environment, and that is continuously active and has direct access to host filesystems, shell command execution, and API credential storage. This architecture is a great source of increasing

DOI: 10.5281/zenodo.18811875

the attack surface as it combines reasoning and execution authority (Ahmed et al., 2025). The ecosystem of the agent presents supply chain factors, whereas its unlimited memory and automated schedule constitute continuous lines of operational risk.

2.2 ChatGPT Security Architecture

A controlled cloud system powers ChatGPT without access to the local system. A vendor manages the security controls, and they are centralized, and API access control allows integration of access (Verma et al., 2024; Luong et al., 2025). Although such architecture is primarily concerned with data governance and the security of the API, companies have to consider the implications of sending sensitive data to third-party infrastructure (Nagaraja & Bahsi, 2025).

3. Threat Modeling Analysis

3.1 STRIDE Comparative Mapping

Applying the STRIDE threat modeling framework reveals distinct risk profiles for each architecture:

STRIDE Category	OpenClaw Severity	ChatGPT Severity
Spoofing	Medium (NIST AI 100-1, 2024).	Medium (OWASP, 2023)
Tampering	High (Huang et al., 2025)	Medium (Team et al., 2023)
Repudiation	High (MITRE ATT&CK)	Medium (ENISA, 2025)
Information Disclosure	High (Vaidya et al., 2025)	Medium (Team et al., 2023)
Denial of Service	Medium (Mohan & Schön, 2026)	Low (ISO/IEC 27001)
Elevation of Privilege	Critical (Rawat et al., 2024)	Low (NIST AI 100-1, 2024)

3.2 Risk Assessment Matrices

OpenClaw Risk Matrix

Threat	Likelihood	Impact	Risk Level
Prompt Injection → Command Execution	High	Critical	□ Severe
Credential Leakage	Medium	High	□ High
Malicious Plugin Installation			
Local File Exfiltration	Medium	Critical	□ High
API Abuse	Medium	Medium	□ Moderate

ChatGPT Risk Matrix

Threat	Likelihood	Impact	Risk Level
Sensitive Data Exposure	Medium	High	□ Moderate
Prompt Injection	High	Medium	□ Moderate
API Key Theft	Medium	High	□ Moderate
Service Outage	Low	Medium	□ Low

4. Security Implications and Mitigations

4.1 OpenClaw Security Considerations

The use of shell command invocation and file manipulation capabilities presents system-level execution risk with OpenClaw. In this respect, the injection of prompt attacks is growing out of proportion, with the malicious inputs inducing command-destroying actions (Zhang et al., 2024). Email or web page external contents have the potential of having embedded commands that undermine the execution tier. The storage of credential keys on the machine makes exposure to lateral movement a high risk in the cases of integrating with such messaging platforms and cloud storage. The open-source presence of the ecosystem creates the vulnerable presence of supply chains in terms of potentially malicious dependencies, and the always-on agent platform raises the persistence risk due to the lack of adequate audit logging (Ma et al., 2023; Simpson et al. 2021).

Mitigation Recommendations: Use OpenClaw on sandboxed or containerized systems that use the least privilege execution policies. Enforce zero-trust network segmentation, host management, and detailed audit assimilation. Manage credential storage in one of the secure secrets vaults and authenticate any external input to verify the incoming input in light of the injection attack (NIST AI 100-1, 2024).

4.2 ChatGPT Security Specializations.

The main information-layer risks that ChatGPT is exposed to are data privacy and compliance. Delicate prompts to the cloud resources can cause regulatory risk in the frameworks, including HIPAA or GDPR. When the system becomes embedded in automated processes, prompt injection attacks may optimize the

DOI: 10.5281/zenodo.18811875

results, whereas the availability of exposed API tokens allows the abuse of the service and impersonation. Vendor dependency brings the aspect of availability when a service is disrupted (Team et al., 2023).

Mitigation Recommendations: Do not transmit sensitive data to cloud LLM services. Leverage intensive API localized rotation and validation. Apply the enterprise data isolation levels where possible and watch API traffic trends towards any deviant usage (OWASP, 2023).

5. Discussion

The comparative analysis indicates that the realm of AI security engineering has changed. Informational risks are the major ones associated with conversational artificial intelligence and focus on confidentiality and usage. Risks that apply to autonomous AI agents include operational and system-level risks, which affect host integrity and availability (Dong et al., 2025). ChatGPT will broaden confidentiality issues in a scenario where OpenClaw will enhance integrity and availability issues because of an execution authority. Threat-driven development frameworks must be appropriately organized to address this extended attack surface despite the drastically different scope of secure engineering of autonomous agents, as explained by Ahmed et al. (2025).

The companies need to choose the architectures depending on the level of threat tolerance, maturity in governance, and operational requirements. ChatGPT is a secure profile when used personally or by non-technical users since the infrastructure manages it and only allows system access to a limited amount. In the case of controlled enterprise settings, where DevSecOps is highly developed, OpenClaw can be implemented safely through the use of proper hardening (Vaidya et al., 2025).

6. Conclusion

Changing the conversational AI to command performing agents can be considered one of the most significant developments in the engineering of cybersecurity risks. Whereas ChatGPT has a lower level of system risk, as sandboxed execution moves in the cloud, OpenClaw generates greater levels of integrity and privilege-based system risks, as a result of the execution authority. Self-sovereign LLM agents cause a significant shift in information-layer threats of AI to cyber-physical system-layer threats, and this paradigm shift in AI security engineering is a critical shift. The security architecture, monitoring regulations, and governance policies need to change as AI systems will move to be active operators and not passive responders.

References

1. Ahmed, T., Hasan, S., Shorif, A., Hoque, A., Sajid, S., & Biplob, M. B. (2025). Secure Engineering of Autonomous AI Agents: A Threat-Driven Development Framework. https://www.preprints.org/frontend/manuscript/b173e9cb92a7f4fe8aa546eca869fa3e/download_pub
2. AI, N. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile. *NIST Trustworthy and Responsible AI Gaithersburg, MD, USA*. https://pathologyinnovationcc.org/s/R8b_1722118549750.pdf
3. Dong, Y., Mu, R., Zhang, Y., Sun, S., Zhang, T., Wu, C., ... & Huang, X. (2025). Safeguarding large language models: A survey. *Artificial intelligence review*, *58*(12), 382. https://link.springer.com/article/10.1007/s10462-025-11389-2
4. ENISA. (2025, August). *ENISA Threat Landscape 2025 | ENISA*. Europa.eu. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2025
5. Gulyamov, S., Gulyamov, S., Rodionov, A., Khursanov, R., Mekhmonov, K., Babaev, D., & Rakhimjonov, A. (2026). Prompt Injection Attacks in Large Language Models and AI Agent Systems: A Comprehensive Review of Vulnerabilities, Attack Vectors, and Defense Mechanisms. *Information*, *17*(1), 54. https://doi.org/10.3390/info17010054
6. Huang, L., Dave, D., Cody, T., Beling, P. A., & Jin, M. (2025, November). From Capabilities to Performance: Evaluating Key Functional Properties of LLM Architectures in Penetration Testing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 15890-15916). https://aclanthology.org/2025.emnlp-main.802/
7. ISO/IEC 27001:2022. (2024). Information Security Management Systems - A practical guide for SMEs. https://masm.isolutions.iso.org/publication/PUB100484.html
8. Khan, M. F. I. (2025). Risk management framework in the AI act. *International Journal of Science and Research Archive*, *14*(3), 466-471. https://www.researchgate.net/profile/Md-Fokrul-Islam-Khan/publication/389799328_Risk_Management_Framework_in_the_AI_Act/links/67d297edd759700065088291/Risk-Management-Framework-in-the-AI-Act.pdf
9. Luong, P. D., Bao, L. T. G., Tam, N. V. K., Khoa, D. H. N., Quyen, N. H., Pham, V. H., & Duy, P. T. (2025). xOffense: An AI-driven autonomous penetration testing framework with offensive knowledge-enhanced LLMs and multi agent systems. *arXiv preprint arXiv:2509.13021*. https://arxiv.org/abs/2509.13021
10. Ma, C., Yang, Z., Gao, M., Ci, H., Gao, J., Pan, X., & Yang, Y. (2023). Red teaming game: A game-theoretic framework for red teaming language models. https://openreview.net/forum?id=BrtOzgElD7
11. MITRE. (2025, November 11). *MITRE ATT&CK v18: What's in it — and why it matters | ReversingLabs*. ReversingLabs. http://reversinglabs.com/blog/mitre-attck-v18-whats-in-it--and-why-it-matters
12. Mohan, A., & Schön, T. (2026). Towards Robust Agents: A Survey of Adversarial Attacks and Defenses in Deep Reinforcement Learning. *IEEE Access*. https://ieeexplore.ieee.org/abstract/document/11363601/
13. Nagaraja, N., & Bahsi, H. (2025, February). Cyber Threat Modeling of an LLM-Based Healthcare System. In *ICISSP (1)* (pp. 325-336). https://www.scitepress.org/Papers/2025/132897/132897.pdf

DOI: 10.5281/zenodo.18811875

14. OWASP. (2023). *OWASP Top 10 for Large Language Model Applications | OWASP Foundation*. Owasp.org. https://owasp.org/www-project-top-10-for-large-language-model-applications/

15. Rawat, A., Schoepf, S., Zizzo, G., Cornacchia, G., Hameed, M. Z., Fraser, K., ... & Varshney, K. R. (2024). Attack Atlas: A Practitioner's Perspective on Challenges and Pitfalls in Red Teaming GenAI. *arXiv preprint arXiv:2409.15398*. https://arxiv.org/abs/2409.15398

16. Simpson, J., Oosthuizen, R., Sawah, S. E., & Abbass, H. (2021). Agile, antifragile, artificial-intelligence-enabled, command and control. *arXiv preprint arXiv:2109.06874*. https://arxiv.org/abs/2109.06874

17. Team, C. 4 A. (2023, September 19). *Best practices in cybersecurity for AI*. Tarlogic Security. https://www.tarlogic.com/blog/best-practices-cybersecurity-ai/

18. Vaidya, H., Nayak, K., & Thakur, A. Self-Hosted AI Coding Assistants for Secure and Real-Time Code Generation. *Architecture*, *14*, 15. https://www.researchgate.net/profile/Hemant-Vaidya/publication/396764236_Self-Hosted_AI_Coding_Assistants_for_Secure_and_Real-Time_Code_Generation/links/68f8a2ade7f5f867e6e17ca1/Self-Hosted-AI-Coding-Assistants-for-Secure-and-Real-Time-Code-Generation.pdf

19. Verma, A., Krishna, S., Gehrmann, S., Seshadri, M., Pradhan, A., Ault, T., ... & Phan, N. (2024). Operationalizing a threat model for red-teaming large language models (llms). *arXiv preprint arXiv:2407.14937*. https://arxiv.org/abs/2407.14937

20. Zhang, A. K., Perry, N., Dulepet, R., Ji, J., Menders, C., Lin, J. W., ... & Liang, P. (2024). Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. *arXiv preprint arXiv:2408.08926*. https://arxiv.org/abs/2408.08926

DOI: 10.5281/zenodo.18811875