International Journal of Artificial Intelligence, Machine Learning and Data Analytics
www.ijaimda.org

| **Open Access** |

Investigating Information Security Challenges in AI-Powered Bots: A Literature Review

Ahmad Musa Bulama[1], Mohammad Faisal[2]

[1]Integral University Lucknow, ahmadmusa@student.iul.ac.in
ORCID: https://orcid.org/0009-0007-5543-8367
[2]Integral University Lucknow, mdfaisal@iul.ac.in
ORCID: https://orcid.org/0000-0003-2583-5293

Corresponding Author: Ahmad Musa Bulama, ahmadmusa@student.iul.ac.in

ARTICLE INFO

Article History:

Received 12/06/2026
Revised 06/07/2026
Accepted 15/07/2026

CITE AN ARTICLE:

Bulama A. M., Faisal M., (2026). ***Investigating Information Security Challenges in AI-Powered Bots: A Literature Review****. International Journal of Artificial Intelligence, Machine Learning and Data Analytics, 1(1), 51-55.*
https://www.ijaimda.org/index.php/ijaimda/article/view/4

ABSTRACT

AI-driven service bots have become game-changing technologies in various sectors, including cybersecurity, healthcare, and customer service. Their increasing adoption comes with complex information security issues, such as privacy risks, adversarial vulnerabilities, data protection gaps, and regulatory compliance issues. To investigate these issues and evaluate design approaches for safe and reliable AI bot systems, this literature review summarizes various recent studies conducted between the year 2021 and 2025. User privacy, adversarial threats, cybersecurity frameworks, AI-enhanced threat intelligence, and privacy-aware system design are among the thematic areas into which the review divides the studies. With suggestions for further research focusing on strong, scalable, and explicable security mechanisms in AI-driven bots, gaps in regulatory alignment, empirical testing, and interdisciplinary integration are noted.

Keywords **-** AI-Powered Chatbots, Privacy Concerns, Cybersecurity, Adversarial Attacks, Threat Intelligence, Regulatory Frameworks.

1. INTRODUCTION

In the ongoing process of global digital transformation, services that once required human to human interaction to solve problems or achieve some goal have started gradually being replaced by conversational agents, also referred to as "AI-powered bots". There are numerous advantages of using such agents, is that, unlike humans, they are available 24/7 and can instantly communicate with multiple users at the same time. Every year, cybercrime damages the global economy by over 400 billion dollars. The U.S. economy loses $ 8.19 million because of data breaches. Cybersecurity experts and researchers are alarmed by the rapid expansion of cybercrime (Alawadhi, S. A., et al, 2022). However, this rapid evolution has also created a complex and dynamic landscape of cyber threats and security challenges. As organizations and individuals continue to embrace the benefits of a digital world, the imperative to fortify cybersecurity measures becomes increasingly urgent.

AI-powered bots, more especially those using natural language processing (NLP) and large language models (LLMs), are increasingly integrated into services requiring sensitive user interactions. As these bots handle private data, analyse user behaviour, and operate autonomously, their security becomes paramount. Ensuring data security is a major challenge for developers, especially in AI-powered bots, where balancing privacy protection with response quality is crucial, given the need for conversation-driven development and data protection regulations. This review examines numerous contemporary academic work to assess the evolving information security landscape in AI-

powered bot systems, categorizing research into user privacy, adversarial robustness, cybersecurity applications, and system-level defence strategies.

2. USER PRIVACY IN AI CHATBOTS

The rapid advancement of Artificial Intelligence (AI) and Natural Language Processing (NLP) has paved the way for transformative innovations, with AI-powered bots at the forefront. These bots are designed to interact with users through text- or voice-based conversations, closely resembling human communication and revolutionizing how individuals engage with technology. Gumusel (2025) conducted a comprehensive review examining privacy concerns in conversational AI using a Social Informatics (SI) framework.

The study highlights how human-like interactions lead to user self-disclosure, underlining the importance of trust, social presence, and perceived intelligence. Despite theoretical advances, design frameworks and regulatory measures remain inconsistent. Throughout AI-powered bots, these bots have been designed to collect significant amounts of personal information without transparency and user consent. This practice has given rise to user privacy and security harms and risks (Dev, 2022). Similarly, Biswas et al. (2024) evaluated mental health chatbot apps, exposing severe privacy vulnerabilities including weak encryption, unclear data policies, and third-party sharing. These studies emphasize the urgent need for privacy-by-design principles and user empowerment features such as data deletion and control.

The information gathered during these interactions could be prone to unauthorized access, data breaches, and misuse, potentially leading to adverse consequences for users. Addressing these concerns within the context of SI not only safeguards individual rights but also lays the foundation for a more resilient and trustworthy technology ecosystem.

Table 1: Key Privacy Vulnerabilities Identified in Chatbot Systems

Study	Key Vulnerabilities	Recommendations
Gumusel (2025)	Self-disclosure, lack of user trust	Multidisciplinary design, privacy frameworks
Biswas et al. (2024)	Weak encryption, poor data policy transparency	GDPR/HIPAA compliance, user-controlled data

3. ADVERSARIAL THREATS TO NLP AND AI SYSTEMS

Research has shown that machine learning is vulnerable to adversarial examples, both theoretically and practically. Until now, such attacks have primarily targeted visual models, exploiting the gap between human and machine perception. Although text-based models have also been attacked with adversarial examples, there has been a struggle to preserve semantic meaning and indistinguishability in such attacks. Adversarial attacks pose serious threats to AI bots. Boucher et al. (2021) revealed imperceptible NLP attacks using Unicode perturbations, which compromise models without human detection. These vulnerabilities affect systems from major providers like Google and Microsoft. Tidjon and Khomh (2022) expanded this scope, examining 89 ML attack scenarios and vulnerabilities in 854 ML repositories. Their work identifies lifecycle-wide threats, including data poisoning, adversarial training, and dependency exploits. Both studies call for proactive encoding controls, multi-layered defenses, and threat modeling frameworks.

The diagram below shows key security issues faced by AI bots in three main areas: privacy, cybersecurity, and human factors. It shows how users overshare personal data, hackers exploit technical flaws, and teams struggle with skill gaps. It also points out that AI training data isn't always reliable. For each risk, it maps out fixes from stronger tech like encryption to smarter policies and user awareness. The clear, color-coded layout helps make sense of how these threats connect and where to focus protection efforts.

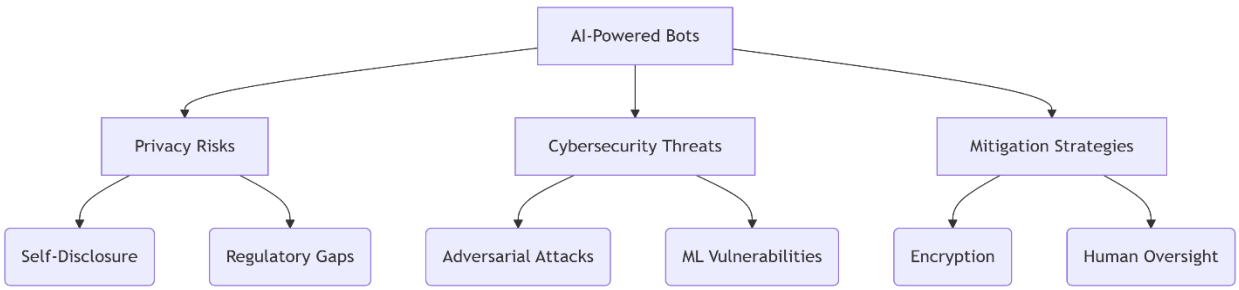

Figure 1. A structured Information Security Challenges in AI-Powered Bots diagram based on the literature review

4. CYBERSECURITY APPLICATIONS AND AI INTEGRATION

Cybersecurity protects computing systems, networks, and data against unauthorized access, use, or damage. It protects society and the economy from online threats that could expose cybersecurity and Applications, the confidentiality, integrity, and availability of various sectors, including business, the military, healthcare, and education. Admass et al. (2024) discuss AI's integration into cybersecurity across sectors like smart grids, vehicular networks, and eHealth. The study emphasizes real-time detection and intelligent automation while highlighting legal, technical, and organizational gaps. Similarly, Liew et al. (2021) propose a novel methodology for analyzing the safety and security of cyber-physical systems through the use of custom metrics. To enable a more comprehensive investigation, the authors introduce an approach that integrates Systems-Theoretic Process Analysis (STPA), a top-down hazard analysis tool with customized matrices. This integration aims to address the limitations of STPA. STPA is employed to identify hazardous control scenarios that could lead to unintended losses. The likelihood of these scenarios being exploited by malicious actors is then assessed using the custom matrices. Akhtar & Rawol (2024) complement this by presenting a mixed-methods study showing AI's effectiveness in reducing false positives and enhancing authentication. However, they caution against algorithmic bias and scalability issues, advocating for ethical frameworks and transparency in AI systems. In 2023, from the figure below; Statista reported that 36 percent of businesses in the United States and the United Kingdom (UK) reported that ransomware attacks were the most significant cyber threat they faced. Furthermore, 22 percent of respondents indicated that phishing attacks pose a major threat to their operations. AI powered Bot attacks were highlighted by 11 percent of respondents, while nine percent of companies said that data exfiltration attacks were also significant.

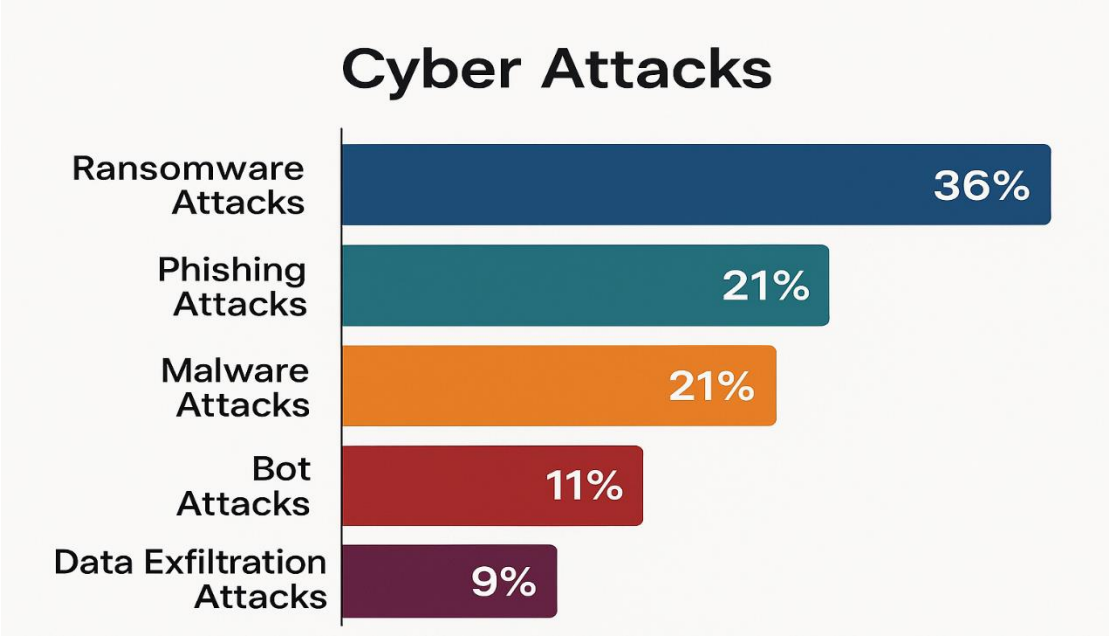

Figure 2. Bot attacks faced by businesses in the US and UK in 2023

5. THREAT INTELLIGENCE AND LLM-DRIVEN SECURITY

Alturkistani & Chuprat (2024) reviewed how AI, especially LLMs, transforms cyber threat intelligence (CTI). Their findings stress the limitations of static models and promote adaptive, real-time, and privacy-preserving techniques like federated learning. Nacheva & Azeroual (2024) focused on ML-based intrusion detection using synthetic datasets, revealing strengths and limitations of classifiers like SVM and kNN. The need for real-world validation and robust adversarial defenses is a shared theme. Machine learning algorithms allow AI systems to process large volume of data and identify potential security threats as early as possible, aiding safety experts in spotting and mitigating risks more successfully. These algorithms also perform predictive analysis, allowing security experts to anticipate and avoid security breaches before they occur. This shift in AI practices has led to a surge in popularity in areas like natural language processing (Alawadhi, S. A. et al., 2022).

6. PRIVACY-AWARE SYSTEM DESIGN

Achieving data privacy is a major challenge for software developers, especially AI-powered bots, where harmonizing privacy protection with response quality is key, given the need for conversation-driven development and data protection regulations. Silva & Canedo (2025) explored privacy in chatbot development, especially under Conversation-Driven Development (CDD). Their study offers eight different privacy requirements, from secure storage to decentralized architectures. While developers favour foundational techniques like encryption, challenges in adopting complex privacy mechanisms remain. Kumar et al. (2024) proposed a secure AI-based document chatbot using AES-256 and OAuth, emphasizing modular architecture and usability. However, empirical validation and threat modelling are underexplored.

7. RESEARCH GAPS

Across the reviewed literature, several key gaps have emerged:

- **Regulatory Inconsistencies**: Discrepancies between privacy laws and implementation hinder trust.
- **Empirical Validation**: Synthetic data and lab settings dominate; real-world testing is rare.
- **Explainability and Ethics**: Black-box models limit transparency, raising concerns about accountability.
- **Scalability and Interoperability**: Many proposed solutions lack deployment benchmarks or cross-system compatibility.
- **Adversarial Defense**: Robust, universal defenses against imperceptible attacks remain elusive.

8. CONCLUSION

This review underscores the increasing importance of addressing information security challenges in AI-powered bots. As these systems become deeply embedded in domains ranging from customer service to cybersecurity, ensuring their trustworthiness and resilience is very important. The study revealed various issues, including privacy vulnerabilities, adversarial manipulation, incomplete regulatory adherence, and technical design limitations.

While there are notable advancements that have been made in frameworks for privacy protection, adversarial robustness, and the application of AI in cyber threat intelligence, persistent gaps continue to hinder real-world implementation. These include the lack of solid empirical validation using live systems and insufficient transparency of AI models, and underdeveloped, scalable solutions. Moreover, variations between privacy laws and their technological enforcement create more friction in the deployment of secure AI systems.

future research must adopt a multidisciplinary approach that combines insights from computer science, social science, ethics, and law. Key priorities should include the development of understandable and adaptive AI systems, broader adoption of federated and privacy-preserving learning techniques, and real-world pilot testing in diverse environments. Regulatory bodies, industry practitioners, and academic researchers must work in concert to create robust, standardized frameworks that facilitate innovation while safeguarding users.

Eventually, the secure design of AI-powered bots requires a user-centric bot that gives priority to privacy, accountability, and resilience by default. By bridging theoretical advancements with practical implementation, the AI community can build systems that perform effectively and earn and sustain user trust in an increasingly connected digital world.

REFERENCES

[1] Admass, W. S., Munaye, Y. Y., & Diro, A. A. (2024). Cyber security: State of the art, challenges and future directions. *Cyber Security and Applications*, *2*, 100031.

[2] Akhtar, Z. B., & Rawol, A. T. (2024). Enhancing cybersecurity through AI-powered security mechanisms. *IT Journal Research and Development*, *9*(1), 50-67.

[3] Alturkistani, H., & Chuprat, S. (2024). Artificial Intelligence and Large Language Models in Advancing Cyber Threat Intelligence: A Systematic Literature Review.

[4] Alawadhi, S. A., Zowayed, A., Abdulla, H., Khder, M. A., & Ali, B. J. (2022, June). Impact of artificial intelligence on information security in business. In *2022 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)* (pp. 437-442). IEEE.

[5] Biswas, A. A., Zulfiker, M. S., Rahman, M. M., Jani, M. R., & Anwar, M. M. (2025, March). Data Privacy and Security Analysis for Mental Health Chatbot Applications. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 1245-1249). IEEE.

[6] Binhammad, M., Alqaydi, S., Othman, A., & Abuljadayel, L. H. (2024). The role of AI in cyber security: Safeguarding digital identity. *Journal of Information Security*, *15*(2), 245-278.

[7] Boucher, N., Shumailov, I., Anderson, R., & Papernot, N. (2022, May). Bad characters: Imperceptible nlp attacks. In *2022 IEEE Symposium on Security and Privacy (SP)* (pp. 1987-2004). IEEE.

[8] Dev, J. (2022). Privacy-preserving conversational interfaces. Indiana University.

[9] Gumusel, E. (2025). A literature review of user privacy concerns in conversational chatbots: A social informatics approach: An Annual Review of Information Science and Technology (ARIST) paper. *Journal of the Association for Information Science and Technology*, *76*(1), 121-154.

[10] Kumar, D. A., Sharif, S. M., & Aravind, T. (2024, October). Development and Implementation of a Secure Document Analysis and Chatbot System for Context-Aware User Interactionsr. In *2024 Global Conference on Communications and Information Technologies (GCCIT)* (pp. 1-6). IEEE.

[11] Liew, L. S., Sabaliauskaite, G., Kandasamy, N. K., & Wong, C. Y. W. (2021, December). A novel system-theoretic matrix-based approach to analysing safety and security of cyber-physical systems. In *Telecom* (Vol. 2, No. 4, pp. 536-553). MDPI.

[12] Nacheva, R., & Azeroual, O. (2024, November). Security of AI-Powered Systems: Threat Intelligence on the Edge. In *2024 8th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)* (pp. 1-6). IEEE.

[13] Silva, G. R. S., & Canedo, E. D. (2025). Privacy in Chatbot Conversation-Driven Development: A Comprehensive Review and Requirements Proposal. *ACM Transactions on Software Engineering and Methodology*.

[14] Tidjon, L. N., & Khomh, F. (2022). Threat assessment in machine learning based systems. *arXiv preprint arXiv:2207.00091*.