International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(15s) 2026 pp. 1-9
https://doi.org/10.70917/ijcisim-2026-4343
Article

Analysis and Design of a Framework Handling Information Security in AI-Powered Service Bots

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

[1] Department of Computer Applications, Integral University, Dasauli, Bas-ha Kursi Road, Lucknow – 226026, Uttar Pradesh, India.
Email: ahmadmusa@student.iul.ac.in
[2] Department of Computer Applications, Integral University, Dasauli, Bas-ha Kursi Road, Lucknow – 226026, Uttar Pradesh, India.
Email: mdfaisal@iul.ac.in

Abstract: While AI-powered service bots have become an important building block in digital service delivery-e.g., in banking, healthcare, e-commerce, and public services-their pervasive handling of sensitive personal and transactional data opens up new information security risks. Recent incidents involving conversational AI systems have demonstrated many vulnerabilities related to data leakage, prompt injection, model inversion, and insecure integration with the host's legacy backends, challenging the traditional security controls designed for static web applications rather than continuously learning AI stacks. This paper analyzes contemporary information security threats specific to AI-powered service bots, and reviews state-of-the-art security mechanisms and standards for proposing a layered framework, which integrates secure model lifecycle management, privacy-preserving data pipelines, robust identity and access management, and continuous security monitoring. The framework is conceptually validated for its properties by mapping it against a taxonomy of chatbot information security dimensions and current best practices, thus proving its appropriateness for improving confidentiality, integrity, availability, and trust in AI-enabled services with the support for regulatory compliance and scalability in its deployment.
Keywords*:* AI-powered service bots, information security, framework, and security threats

1. Introduction

This trend is indicative of a general trend to the automation and always-on digital interaction that is also taking hold across customer support, citizen services, and enterprise workflows.[2]. Such systems are underpinned by NLP, large language models (LLMs), and integration with backend APIs to also fulfil tasks such as account inquiries, incident reporting, order management, and basic decision support [3]. While they yield huge gains in terms of efficiency and user experience, they also present expanded attack surfaces, including input channels, model behaviour, orchestration layers, and APIs, data lakes, and third-party services.[5].

Several recent studies and industry reports underpin that both benign and malicious bots now represent a significant part of Internet traffic and that AI-powered bots increasingly play the role of both targets and tools in cyber-attacks.[9]. For service bots interacting with end-users, these dynamics materialize as risks of sensitive data leakage, unauthorized actions via compromised sessions or APIs, adversarial manipulation of model outputs, and exploiting of misconfigurations in cloud-native architectures [10]. Traditional security frameworks based on perimeter protection and rule-based access control cannot suffice because AI components may be non-deterministic in their behaviour, can learn from live data, and are susceptible to becoming subtly adversarial in nature[11].

This paper addresses two main questions: within parentheses, following the example. Some components, such as multi-levelled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

Copyright: © 2026 by the authors

• What are the unique challenges related to information security that service bots powered by artificial intelligence must address as opposed to traditional Web and mobile apps?[4, 5]

• How might the framework be developed in order to mitigate the risks inherent in the lifecycle of AI Enabled Service Delivery from data gathering and model training through into retirement[12][4].

The rest of the paper is structured as follows: Section II discusses related work from chatbot information security and AI security frameworks. Section III identifies threats and vulnerabilities of AI service bots based on security taxonomies. Section IV describes a proposed multi-layered framework, including architectural details, processes, and security control procedures. Section V explores implementation aspects of standardization and compliance, while Section VI concludes with future research [3][5][6][1][2][4].

2. Related Work

A. Chatbot and conversational AI security

There has been systematic analysis of the chatbot's security issues concerning the risk of chatbots being subjected to a different set of threats such as the creation of malicious inputs, user profiling, context-aware attacks, and data breaches that may result from the improper handling and storage of data. Literature reviews demonstrate the fact that most of the chatbots used are prone to a lack of strong encryption techniques and user authentication and the need for specific chatbot behavior governance [3][4].

There are several other important technical considerations that are highlighted as best practices in the security of conversational AI, which include end-to-end encryption for all chats, penetration testing to ensure that all current and potential vulnerabilities are addressed, and the need for centralized bot management that can monitor all bot activities. Other important considerations highlighted by these resources include filtering and restricting rates and anomalies[6][2][3].

B. AI and cybersecurity integration

There are examples of AI-related research into information security, which show that AI can enhance the security of threat detection, automate the response to security incidents, and make authentication more adaptive. At the same time, such research points to the increased use of AI by hackers to create more sophisticated attacks. A study of AI-related security issues maintains that the security of AI information systems needs to consider both information security principles and AI-related issues.

Research literature on organizational cybersecurity suggests that AI systems have lifecycle-wide implications, meaning security and governance considerations need to be integrated from data collection and model development to post-deployment and decommissioning phases of the lifecycle. Reference architectures for AI-decision support systems in cybersecurity applications demonstrate approaches for incorporating explanations and evidence-based reasoning to improve visibility and control over AI decisions. Such literature implications are useful to design a service bot from a security lifecycle approach for justifying or delimiting automated actions performed on behalf of users [13][1][5].

3. Secure architectures for conversational systems

Recent case studies suggest that a secure conversation AI system can be realized by leveraging cloud-native tools, automated PII identification, and access control based on policies. Specifically, for example, a conversation system based on managed cloud infrastructure uses automated data classification, data encryption at rest and in transit, and least-privilege access to secure the conversations that are cached in object repositories. Research prototypes of chatbots for cybersecurity purposes indicate that conversation assistants can assist users in various security operations and, in turn, remain secure using communication protection and backend security [14][7][2][4].

A 2024 systematic review suggests that there could be a taxonomy of information security in chatbots, covering technological, enterprise, as well as human aspects. This includes data privacy, authentication, logging, awareness, and controls in service-level agreements. Such a taxonomy helps as a theoretical basis to create a structured approach appropriate to service bots empowered by artificial intelligence [4].

4. Problem Definition

The differences brought by intelligent, service-oriented robots in terms of design in information security are in these three areas, which are significantly unique to traditional applications:

- These are capable of processing highly natural inputs involving potential personalized information that may even contain sensitive information like sexual behavior and expenses [2][4].
- Since their internal processes, determined by machine learning algorithms and context windows, may lack transparency and may in fact be unverifiable [12][5].
- Representing a crossroads of interacting systems that comprise channel front-ends, orchestration layers, model providers, logging infrastructures, and business services. Reference: [6][3].

These features pose many challenges for organizations:

- Confidentiality and privacy with regard to user communications potentially containing PII or other confidential data in the logs and the training data sets that can be reused for model enhancement [14][2].
- Preventing impairment of actions initiated by the bot to perform payments, update records, and provide access from abuse based on prompts, stolen tokens, and compromised APIs [3][6].
- Maintaining availability and robustness while exposing bots as public endpoints vulnerable to denial-of-service attacks or resource-exhausting conversations [10][9].
- Compliance with legislation such as GDPR, privacy legislation in sectors, and the increasing governance of artificial intelligence that require transparency, consent, and risk mitigation [1][14].

Current security infrastructures developed with respect to web APIs and microservices handle some of these challenges, yet they hardly ever address how large language models and conversations exhibit emergence, thus bridging the gap between AI safety research and security engineering practices in enterprises [5][12].

5. Threats and Vulnerabilities in AI Service Bots

1. Taxonomy overview

Building on the chatbot information security taxonomy, threats to AI-powered service bots can be grouped into four broad domains: data-centric threats, model-centric threats, integration-centric threats, and human-organizational threats [4].

TABLE I.

Domain	**Example threats**	**Core security objectives impacted**
Data-centric	Data leakage, insecure storage, weak encryption	Confidentiality, privacy
Model-centric	Prompt injection, adversarial examples, model inversion	Integrity, confidentiality
Integration-centric	API abuse, broken access control, weak secrets management	Integrity, availability
Human-organizational	Social engineering, misconfiguration, poor governance	All (C, I, A, accountability)

2. Data-centric threats

It is common for service bots to act as a conduit for valuable data, such as identifiable information, payment information, healthcare information, and documents shared in conversations, if it is shared in the course of a service Bot interaction. It is possible for such data to be stored in plaintext, shared with a model provider, and retain vulnerabilities with analytics platforms [14][2][3].

Cases like these, involving high-profile situations wherein workers copy and paste proprietary code snippets into public conversational AI systems, convey that sensitive information diffusion beyond an organization's limits isn't a difficult task when less clear guidelines aren't followed and appropriate technical measures aren't employed. Moreover, if conversation transcripts are repurposed for a model's training with ineffective anonymization techniques, there are chances that a model may retain and repeat sensitive information upon given prompts [5][2][4].

3. Model-centric threats

Model-centric threats rely on the behavior of artificial intelligence models, which enable service bots. Prompt injection can attack these by aiming to bypass system commands or jailbreaking safety nets, which could result in the release of internal data, undesired behavior, or harmful output from the bot. The attack could output undesired behavior or model decisions, especially for triage or access decisions done by service bots [11, 12, 6, 2, 5].

Model inversion and membership inference attacks are attempts at reversing models in order to gain access to the training data and determine whether specific examples are part of the training dataset, which also poses risks even where direct access is prevented. Such threats are more particularly relevant in training models and fine-tuning models for sensitive call logs without access controls [4][5].

4. Integration-centric threats

Service bot applications are not frequently operated solo; rather, Service bot are orchestrated using user directory services, customer relationship management platforms, banking cores, or ticket services using APIs, web hooks, or message queues. Issues in-authenticated APIs, deficiencies in authorizations, or substandard secrets management practices (such as token hard coding) can slyly allow attackers to misuse service bots as a proxy for backend applications access [7][6][3].

When bots are left publicly accessible via channels, they can be vulnerable to automated abuse such as credential stuffing attacks, resource exhaustion from bot-Script interactions, or the injection of illicit content meant for poisoning analytics data. Unchecked error handling or responses from external resources can spread erroneous or sensitive information back to the user [9][10][6][2].

5. Human and organizational factors

Human factors are still significantly exposed in AI service bot security. End users could over-trust these service bots, which would provide them with more information than they would provide in a traditional form, whereas employees could misuse internal service bots for convenience by bypassing secure communication channels. System administrators or developers could also misunderstand security controls because of complex cloud-native AI technology, allowing unnecessary ports to stay open, incorrect access scopes, or even disable logging [15][14][2][3][4].

There are also some challenges for organizations in the area of governance, such as establishing a clear accountability framework for AI-bot actions, using explainability to enhance automated decision-making, and aligning AI-risk management to ISMS. Lack of clear policies, training, and control can result in a diminished effectiveness of highly technical architecture in terms of risky utilization practices [15][1][5][4].

6. PROPOSED FRAMEWORK FOR SECURE AI-POWERED SERVICE BOTS

1. Design principles

The proposed framework applies to all AI service bots and is based on four design principles: Security by Design, Privacy by Default, Defense in Depth, and Continuous Assurance. By Security by Design, it is meant that threat analysis, secure code development, and security testing should not be treated as an afterthought in the development life cycle of service robots. Privacy by Default means that there should be least data collection, appropriate context of data retention, and good consent practices, particularly in the context of personal data/PII.

Defense-in-depth is achieved by stacking controls on channels, models, orchestration, APIs, and infrastructure so that a single point of failure will not directly compromise. Continuous assurance is necessary due to dynamic threat environments, realizing AI is constantly operating within a constantly changing environment, thus adapting their controls either automatically or with human-in-the-loop methods [10][6][3][4].

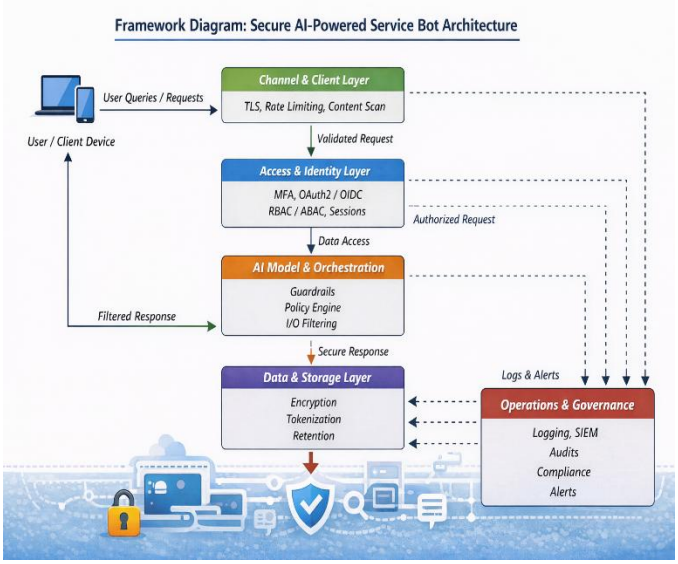

Figure 1:

The figure presents a layered Data Flow Diagram (DFD) that illustrates how user data is securely processed within an AI-powered service bot, from initial interaction to governance and monitoring. Each layer represents a distinct security boundary that enforces specific controls, ensuring defense-in-depth.

2. *Layered architecture overview*

The framework is organized into five layers: channel and client, access and identity, AI model and orchestration, data and storage, and operations and governance [3][4].

TABLE II.

Layer	**Focus**	**Core controls (examples)**
Channel & client	User interfaces and endpoints	TLS, device checks, rate limiting, content scanning
Access & identity	Authentication and authorization	MFA, OAuth2/OpenID Connect, RBAC/ABAC, session security
AI model & orchestration	Model interaction and business logic	Guardrails, input/output filters, policy engines
Data & storage	Data lifecycle management	Pseudonymization, encryption, tokenization, retention
Operations & governance	Monitoring, compliance, and human factors	Logging, SIEM, audits, policies, training

Each layer is described in more detail below, along with representative controls and processes.

3. *Channel and Client Layer*

The channel and client layer encompasses web chat widgets, mobile apps, messaging services, and voice interfaces, where users interact with service bots. Protecting this level first involves transport-level security, utilizing up-to-date transport-layer security protocols, pinning certificates when necessary, and implementing down grade Protection and man-in-the-middle protections [15][2][3].

Rate limiting and behavior analysis on edge assist with dealing with potential misuse of automation scripts on bots exposed on public-facing sites or messaging services. Scanning for malicious content based on obvious malware, malicious URLs, scripting attempts can also be done before the input reaches the AI algorithm to avoid injection into backend logs [9][10][2][6].

Further protect clients by using secure token storage on mobile devices, avoiding the embedding of secrets within client-side code, and utilizing secure cookies and headers to guard against cross-site scripting and request forgery when web channels are present. For voice interfaces, the problem of both spoofing and replay must be addressed, and voice biometrics can be utilized to enhance assurance when verifying identities [12][15][3].

4. Access and Identity Layer

Access and identity regulate who can interact with a bot and the kind of actions they can tell the bot to perform. External users should be authenticated via standardized mechanisms such as OAuth2 or OpenID Connect with MFA applied on sensitive operations like payments, access to data, or account changes. Internal users-e.g., employees accessing enterprise bots-should be using single sign-on with strong identity verification and device- or network-based conditional access[1][15][3][4].

In any case, role-based access control or attribute-based access control shall be implemented at the backend service level and not just in the conversational layer, so, even if there is a manipulation of prompts, a bot wouldn't be able to perform any actions outside the scope of authenticated user's privileges. Token management, including short-lived access tokens, refresh token rotation, and secure storage reduces the impact of token theft or replay [15][6][3].

Given that conversational systems usually span many turns and even channels, session management becomes a critical task; techniques include binding sessions to devices or contexts, enforcing idle timeouts, and requiring step-up authentication as the conversation moves from low-risk queries to high-risk transactions. Robust identity auditing and traceability support incident investigation and regulatory requirements for accountability in AI-mediated interactions [1][2][5][4].

5. AI Model and Orchestration Layer

The model and orchestration layer is the heart of the bot and is concerned with the processing of the user's input and responses as well as the execution of the business logic to be applied to it or act upon it in some manner.

For a model to be designed securely, system calls, policy engines, and guardrail elements are required to limit its model performance to abide by organizational policies regardless of whether it's given any malicious input or not.

Filtering and normalizing the input and output messages are critical initial steps where the messages could be checked for malicious activity, length, and types of content that are unauthorized prior to injecting them into the model. Filtering the outputs can be useful for ensuring that the outputs do not reveal internal details of the configuration or messages and that dangerous instructions are blocked while the content of the outputs is safe [14][2][6][4].

The orchestration layer is responsible for facilitating calls between the model and backend systems, enforcing explicit policies for allowable API calls, with what parameters, and as what user identity. Methods like whitelists for tool calling, schema validation, and policy-as-code help to prevent the model from making arbitrary or unvalidated calls based on potentially misleading prompts. There may be approval flows with human-in-the-loop approval for critical changes, involved for risky operations, when approval is necessary before making a critical change request via the bot [7][11][5][1][3].

In terms of lifecycle management, machine learning model training and tuning must adhere to secure MLOps best practices such as training data access control, versioning, replicability, and testing for adversarial behavior and fairness. Deployment pipelines must then implement integrity checks, code signing, and environment containment in order to ensure that subversive components can't change the code in the background [12][5].

6. Data and Storage Layer

The data and storage aspect includes the scope of what the layer will cover, including the capture of conversation transcripts, logs, the corpus for training, and analytics. A privacy by design model adopts a privacy-conscious system that commences the process with the aspect of minimization. This includes the collection of only

the needed information for the specific requested service and the strong discouragement or prevention of the collection of extremely sensitive information into general-purpose bots.

PII identification systems like cloud-based classifiers would be able to identify the sensitive fields in conversation data and then redact, mask, or tokenize this data prior to storage or processing. Data at rest protection by more robust algorithms in conjunction with encrypting all data flows in transit will protect against storage or network infrastructure breaches [2][15][3][14].

A data retention policy must consider legal and business requirements, ensuring that conversations can be retained only for a needed period and then deleted or anonymized securely. When models are trained or fine-tuned using conversations, strong de-identification methods, secure access mechanisms for data scientists, and tracking data lineage can help minimize risks associated with re-identification or memorization of data. Logging settings must be carefully considered, ensuring a balance between making data useful for forensics and a lack of full payload logging as a means of ensuring higher privacy values [5][6][1][4][14].

7. Operations and Governance Layer

The operational and governance layer ensures that technical controls are suitably monitored and maintained, and aligned with organizational objectives and regulatory expectations. Centralized logging and SIEM platforms should ingest events from all the following layers: channels, access control, models, APIs, and infrastructure to support correlation and anomaly detection.

Security operations teams can also use AI-driven analytics to spot unusual conversational patterns, spikes in failed authentication attempts, or anomalous API call sequences indicative of prompt injection, credential abuse, or automated attacks. Regular penetration testing and red-teaming exercises against the bot include simulated jailbreaks and data exfiltration to help validate guardrails and filters [11][6][3].

From a governance perspective, service bot security needs to be integrated into the broader ISMS of organizations, including risk assessments, policy frameworks, and compliance audits. Accountability is also facilitated through clear allocations of roles and responsibilities; for instance, product owners, security architects, model owners, and data protection officers. Training and awareness programs for both end users and staff are necessary, which would promote safe use of bots, explain limitations, and reduce the possibility of an accidental leak of data or social engineering [1][15][4][5][2].

7. DISCUSSION

1. Alignment with standards and regulations

Furthermore, this framework is consistent with established best practice for information security, for example, ISO/IEC 27001, but also takes into consideration new regulatory obligations for AI. For instance, the principles of confidentiality, integrity, and availability can be ensured at each level by employing encryption, access controls, auditing, and security controls, and compliance with obligations for privacy, for example, purpose and storage limitations, could be included within the data life cycle [4][5][1][14].

As a result, as regulators are increasingly concerned with high-risk AI systems such as those impacting the use of essential services, organizations will be required to show that they are managing risk well in their AI systems[11]. This has been enabled through the framework's focus on explainability in decision logic, well-defined limits on the role of the bot, as well as human-in-the-loop when critical actions are required. However, there may be different interpretations in a particular regulated environment [13][12].

2. Benefits and limitations

The benefits of adopting this structured approach include: segregation of responsibilities on various technical levels, more systematic protection against threats, and improved communication between security, data, and product teams. It is also easy to adopt this approach incrementally because an organization could start improving specific levels, such as PIIs within logs or guardrails, while planning to adopt this approach at a bigger transformation scale in the future [6][3][14][18].

However, some limitations still need attention. The models might continue to behave erratically even under strong guard rails, and detecting prompt injection attacks as well as data exfiltration attacks through natural language is still an area of research[10, 11, 5, 14].Small firms could also find monitoring, in addition to in-house knowledge,

resource-intensive, further increasing reliance on external platforms that could, in turn, have a strong risk profile that requires attention [15].

8. Conclusion and Future Work

Service bots enhanced by AI are now an essential part of modern servicing but represent a complex information-security environment because of their specific combination of conversational interfaces and powerful models highly integrated into business processes. The paper assessed various threats in relation to information contained in data, models, integrations, and humans and has outlined a multi-layered approach that could incorporate secure channel management, high-quality identity management, model safety guidelines, privacy-focused data management, and sound operating and governance processes[3][1][4].

Future directions will be to empirically evaluate the framework using case studies and risk reduction metrics, as well as to explore new techniques such as the formal verification of orchestration policies, the application of differential privacy to the training data of conversational AI, and the identification of adversarial interactions in the natural language. The involvement of researchers in AI, security experts, and representatives from the relevant agencies and consumer advocacy organizations will be very important to ensure that service bots are not only smart and optimal, but also safe and trust-worthy[12][5][16][17].

Acknowledgment

The authors acknowledge the support of the Department of Computer Application, Integral University, Lucknow, India. This manuscript has been assigned Institution ManuscriptCommunication Number IU/R&D//2026-MCN0004618.

References

1. I. Jada and T. O. Mayayise, "The impact of artificial intelligence on organisational cyber security: An outcome of a systematic literature review," Data and Information Management, vol. 8, no. 2, p. 100063, 2024.
2. A. Kumar, D. A. Khan, and R. Amin, "Ensuring secure conversational commerce: Anomaly detection in chatbot interactions," Expert Systems with Applications, vol. 295, p. 128769, 2026.
3. S. Narula, M. Ghasemigol, J. Carnerero-Cano, A. Minnich, E. Lupu, and D. Takabi, "Exploring AI security: A systematic mapping study," IEEE Access, 2025.
4. J. Yang, Y. L. Chen, L. Y. Por, and C. S. Ku, "A systematic literature review of information security in chatbots," Applied Sciences, vol. 13, no. 11, p. 6355, 2023.
5. V. Bhardwaj, B. K. Dhaliwal, S. K. Sarangi, T. M. Thiyagu, A. Patidar, and D. Pithawa, "Conversational AI: Introduction to chatbot security risks, their probable solutions, and the best practices to follow," in Conversational Artificial Intelligence, pp. 435–457, 2024.
6. A. M. Buttar, F. Shahzad, and U. Jamil, "Conversational AI: Security features, applications, and future scope at cloud platform," in Conversational Artificial Intelligence, pp. 31–58, 2024.
7. V. Bhardwaj, S. S. Khan, G. Singh, S. Patil, D. Kuril, and S. Nahar, "Risks for conversational AI security," in Conversational Artificial Intelligence, pp. 557–587, 2024.
8. R. M. Potluri, A. K. Jumasseitova, and L. S. Potluri, "The role of artificial intelligence (AI) in combating digital marketing fraud and bot attacks," in AI-Enabled Threat Intelligence and Cyber Risk Assessment, CRC Press, pp. 17–28, 2025.
9. B. Guembe, A. Azeta, S. Misra, V. C. Osamor, L. Fernandez-Sanz, and V. Pospelova, "The emerging threat of AI-driven cyber attacks: A review," Applied Artificial Intelligence, vol. 36, no. 1, p. 2037254, 2022.
10. M. M. Nair, A. Deshmukh, and A. K. Tyagi, "Artificial intelligence for cyber security: Current trends and future challenges," in Automated Secure Computing for Next-Generation Systems, pp. 83–114, 2024.
11. S. Neupane, S. Mitra, I. A. Fernandez, S. Saha, S. Mittal, J. Chen, et al., "Security considerations in AI-robotics: A survey of current methods, challenges, and opportunities," IEEE Access, vol. 12, pp. 22072–22097, 2024.
12. C. Ebert and M. Beck, "Artificial intelligence for cybersecurity," IEEE Software, vol. 40, no. 6, pp. 27–34, 2023.
13. H. W. Lee, T. H. Han, and T. J. Lee, "Reference-based AI decision support for cybersecurity," IEEE Access, vol. 11, pp. 143324–143339, 2023.
14. M. Bhanushali, H. Parekh, R. Mistry, and Y. Mane, "TAKA cybersecurity chatbot," in Proc. 2023 Int. Conf. on Advanced Computing Technologies and Applications (ICACTA), pp. 1–6, Oct. 2023
15. N. Javed, T. Ahmed, and M. Faisal, "An efficacy comparison of supervised machine learning classifiers for cyberbullying detection and prediction," *International Journal of Bullying Prevention*, pp. 1–20, 2024.
16. A. M. Bulama and M. Faisal, "Human-AI collaboration in education: Enhancing learning environments," in Proc. Int. Conf. Trends and Development in Science and Engineering, Sep. 29, 2025, doi: 10.17577/IJERTCONV13IS06039**.**
17. [17] R. A. Khan and M. F. Farooqui, "Validation of performance impact in S-AMMA using AHP-A multi-criteria decision-making approach," IJECS, vol. 8, no. 1, pp. 68–74, 2026

18. A. M. Bulama, M. Faisal, H. M. Bello, and M. M. Fadama, "Securing AI Chatbots: A Framework for Trust and Standardization," International Journal of Drug Delivery Technology, vol. 16, no. 49s, pp. 908–914, 2026, doi: 10.25258/ijddt.16.49s.102.