International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.56472/ICCSAIML25-130
Eureka Vision Publication | ICCSAIML'25-Conference Proceeding

Original Article

Secure Messaging Protocols for Intelligent Chatbots: Enhancing User Trust and Data Privacy

Satya Karteek Gudipati
Principal Software Engineer, Texas, USA.

Abstract *- Interaction with chatbots has become a new normal in today's world. We interact with different types of bots without our knowledge. Given this context, how safe are the chatbots and can we trust them and provide our details? The industry is growing leaps and bounds in terms of AI/ML technologies and their seamless integration with chatbots. But these AI models lack contextual awareness as well as enhanced security. This paper comes up with a proposal to embed secure messaging protocol within the chatbot design, and we also touch the challenges related to privacy and trust of the end user. This paper lists down the common vulnerabilities, limitations of enhancing security and how we can bridge this gap through integration with real-time anomaly detection frameworks. This paper introduces a trust centric model that dynamically checks the risk assessment, bad-actor detection etc., The preliminary experimental data which is done under controlled test environments gave us a clear indicator that security breach attempts were mitigated significantly when compared with traditional chatbots. The concept of Trust-adjusted interactions safeguarded the chats and also improved the user confidence, reliability on system etc. The framework proposed thoroughly focuses on enhanced security, improving privacy, contextual awareness etc.*

Keywords *- Adaptive Trust Modeling, AI-Powered Messaging Systems, Chatbot Intelligence, Chatbot Security, Compliance and AI Ethics.*

1. Introduction

The dependency on chatbots is increasing day-by-day in both consumer and enterprise applications which led to research around conversational AI. We have powerful AI models evolving in terms of Natural Language Processing, on the other hand secure communication protocols for chatbots have become an untested territory . Majority of the chatbots in the enterprise world have integrations with OAuth 2.0, data is transferred over https, websockets etc. This generally gives the security of communication at the transport layer (layer 4), but companies do not stress on critical things like data privacy, data encryption of both data at rest and data in transit. Even the Technology giants like Facebook, X, Amazon, Google etc., face challenges every day to provide their chatbots with high-end security without compromising the performance, data privacy and compliance[3].

Most of the time, chatbots often route users and user data to centralized cloud services and this raises serious questions on the ownership of data, control and potential misuse of customer private data[1]. There are many scientific studies which highlight the importance of sustaining user trust in chatbots through transparency, integrity and compliance. It's always in the best interest when we include privacy by design and user-centric security during the design phase rather than layering them at later stages. However, most of the time enterprise giants give these principles a back burner due to time-to-market pressures. By following end-to-end encryption, forward secrecy and zero knowledge articles, apps like Signal, matrix and Telegram are gaining user trust and popularity. This is possible because these apps primarily focus on communication between two individuals. But adapting the same to chatbots poses some challenges, such as session persistence across user contexts, dynamic key management for bots, balancing latency with encryption complexity, ensuring explainability for non-technical users.

Secure messaging and chatbot intelligence look like two individual streams and when we combine both, there arises a significant void in the research data. When the chatbot frameworks combine secure messaging and chatbot intelligence, then we can achieve real-time conversational security, adaptive trust modeling and granular level data access control in chatbot applications. By focusing on a unified approach towards messaging security, user experience and AI system design, this paper tries to bridge the gaps in the chatbot frameworks. This paper targets to bring focus on unexplored territory within conversational AI: secure messaging protocols. This paper introduces a novel concept of integrating robust cryptographic technologies with adaptive trust modeling that safeguard user data during chatbot interactions. The goal here is to provide a strong privacy-first messaging architecture while evaluating performances through practical implementations and also addressing the vulnerabilities, we achieve a strong foundation to secure, transparent and trustworthy chatbot landscape.

2. Threat Landscape

Chatbots are omnipresent in the enterprise landscape, they exchange data with the backend systems and this has opened the door for malicious actors who are looking for opportunities to exploit vulnerabilities that present in the systems[4]. Traditional messaging systems have a context attached to their communications whereas chatbots interact anonymously across various platforms making them more susceptible to a larger spectrum of cybersecurity threats.

2.1 Common Attack Vectors

- Man-in-the-Middle (MITM) Attacks: In MITM attacks, malicious actors will exploit the vulnerabilities and try to intercept the handshake between the user and the chatbots. This will open a security threat and will eventually lead to sensitive data leaks. The TLS protocol will only prevent basic interception, but it cannot prevent the data theft once it's decrypted at the server or application level.
- Data Leakage and Unauthorized Access: Chatbots have to store conversational data at the client side and most of the time this data is stored in plain text without any encryption. This approach will eventually become a liability to the enterprise when a potential data leak triggers an audit and ends up into a compliance violation[2].
- Impersonation and Spoofing: It is fairly easy for attackers to impersonate a chatbot and steal the conversational data that sits in the user's device. This will lead to potential theft of the user's confidential data. Methods like spoofing can also expose vulnerability of weak authentication mechanisms in a session-based chatbot application.
- Adversarial Prompting and Injection: Seasoned and professional attackers will use methods such as prompted injection or adversarial inputs which manipulate the behavior of NLP models within a chatbot. This way the attacker can penetrate and alter the responses or redirect data. They can even bypass internal filters and safeguards.
- Session Hijacking: This is becoming popular and more problematic in today's world. Weak tokens and identifiers will open the flood gates where attackers may gain access to the entire conversation and also live sessions by replaying mechanisms through the weak session keys.

2.2 Case Studies and Real-World Incidents

There are many real world 'accidents' that occurred due to vulnerabilities in chatbots:

- In 2020, a health care chatbot is responsible for a patient's private data leak due to a bad configuration which resulted in granting access control to a cloud storage bucket.
- A data migration error resulted in exposing transaction history via a finance chatbot which is integrated with a third-party CRM.
- Attackers are able to trick bots and get access to internal API documentation. This happened for a customer service chatbot with prompt injection vulnerabilities.
- These incidents are only the tip of an iceberg. Incidents like these show us the importance of security within the entire chatbot application.

2.3 Risk Modeling and Impact Analysis

When the primary focus of chatbot architecture is mainly on conversational context, the importance to security takes a back seat. But the risk model reveals:

- Multi platform deployments (web, mobile app) always have a larger surface exposure to vulnerabilities.
- Medium-to-high impact severity based on the type of data handled (PII, financial, medical).
- Sectors like finance and healthcare have cascading effects on user trust, regulatory compliance and brand reputation.

3. Proposed Secure Messaging Framework

The previous sections mainly focussed on bringing the security and privacy challenges into discussion. In the sections that follow, this paper proposes a messaging framework that is modular, adaptable and secure enough to mitigate real-time threats and security issues while maintaining positive user experience and also achieving conversational efficiency and performance[5][6].

3.1 Architectural Overview:

The framework consists of the following core components. These components are loosely coupled through a PUB-SUB model resulting in a low latency while isolating sensitive tasks.

- Client Interface Layer: The chatbot front end (web, mobile, or voice assistant) interacting with the user.
- Secure Communication Core: Handles message encryption, decryption, and signature verification.
- Trust Management Engine: Assigns and adjusts trust levels based on user behavior, context, and historical patterns.
- Session Management Module: Maintains secure and tamper-resistant conversational sessions.

- Anomaly Detection Subsystem: Monitors for suspicious patterns in communication or system behavior in real time.

3.2 Encryption and Data Protection:

The proposed framework employs end-to-end encryption (E2EE) between the user and the chatbot processing engine. Each message is encrypted using a hybrid scheme like Asymmetric encryption (e.g., RSA-4096 or ECC) for key exchange, Symmetric encryption (e.g., AES-256) for actual message content. To ensure forward secrecy, ephemeral session keys are generated per session and rotated periodically. Message headers have threshold limits and metadata is obfuscated to reduce leakage.

3.3 Zero-Knowledge Authentication:

A zero-knowledge proof (ZKP) system is integrated for identity verification, allowing the chatbot to confirm a user's credentials or access rights without ever storing or exposing those credentials. This comes handy while dealing with sensitive data.

3.3.1 Use cases:

- Authenticating a user without password entry.
- Verifying user consent without storing biometric markers.
- Access control for multi-user chatbot interfaces.

3.4 Context-Aware Trust Scoring:

The Trust Management Engine calculates a dynamic trust score for each interaction using a weighted model based on:

- Historical chat behavior (e.g., frequency, sentiment stability)
- Device fingerprint and network trustworthiness
- Conversation topic sensitivity (e.g., casual chat vs. account recovery)
- Anomalous interaction markers (e.g., sudden language change or unnatural input rhythm)

3.4.1 Based on the trust score, the chatbot can:

- Request additional authentication
- Disable certain functions
- Escalate to a human operator
- Limit or redact sensitive responses

3.5 Secure Session Handling:

To prevent session hijacking, replay attacks, and unauthorized access in shared-device scenarios each chatbot-user interaction is encapsulated in a secure session with:

- Expiry mechanisms
- HMAC-based integrity checks
- Secure cookie/token-based identification (never stored in plaintext)
- In-memory session logs with automated purging

3.6 Anomaly Detection and Real-Time Defense:

Using lightweight machine learning classifiers and pattern-matching rules, the Anomaly Detection Subsystem flags:

- Repetitive automated queries
- Credential stuffing attempts
- Language that may indicate coercion or fraud

It integrates with the trust engine to escalate or lock down responses in real time, enabling conversational resilience.

4. Results and Discussion

A mult-environment prototype chatbot is developed to validate the effectiveness of the proposed secure messaging framework. The section below illustrates the technical setup, implementation details and evaluation results.

4.1 Implementation Setup Technology Stack

- Frontend: React.js chatbot UI (Web & Mobile)
- Backend: Python (FastAPI) + Node.js (for session handlers)

- Encryption: Asymmetric: Elliptic Curve Cryptography (ECC-P521)
- Symmetric: AES-256-GCM
- ZKP Protocols: zk-SNARK via libsnark
- Session Management: Redis (in-memory) + JWT with HMAC-SHA256
- Trust Engine: Scikit-learn + rule-based logic
- Anomaly Detection: LSTM-based classifier trained on synthetic attack data

4.2 Experimental Scenarios

- Scenario A: Customer Support Chatbot
- Handles billing queries, account status, and live support escalation
- Security test: man-in-the-middle simulation and session hijack
- Scenario B: Healthcare Symptom Checker
- Collects symptom data and provides triage advice
- Security test: prompt injection, impersonation, data leakage risk

4.3 Metrics for Evaluation

Table 1. Key Performance Metrics for Secure Communication Systems

Metric	Description
Encryption Overhead	Time added per message for secure transmission
Detection Accuracy	Precision/recall of anomaly detection module
Trust Model Responsiveness	Time taken to adjust trust level dynamically
Message Round Trip Time	Overall latency with and without security
User Trust Rating	Survey-based feedback on perceived safety

4.4 Results: Performance vs. Baseline

Table 2. Security Threat Mitigation Performance and User Trust Impact

Test Scenario	Detection Accuracy	Overhead (ms)	Trust Score Adjustment	User Trust Rating (1-5)
MitM Simulation	98.2%	+42 ms	Fast (≤ 300ms)	4.7
Session Hijack Attempt	100% blocked	+36 ms	Immediate	4.8
Prompt Injection	95.6%	+49 ms	Conditional	4.6
Impersonation via Spoofed ID	97.3%	+31 ms	Gradual	4.5

Compared to a conventional chatbot using only HTTPS and static tokens:

- Security breach attempts dropped by over 95%
- Trust-adjusted interaction decisions prevented 2 out of 5 privacy breaches in testing
- Users perceived the secure system as more “serious” and “reliable”

5. Discussion

By implementing the proposed secure messaging framework and through a thorough evaluation, we try to achieve integration of privacy-preserving protocols into chatbots. This is the need of the hour for the chatbots which are evolving in the intelligence part but falling behind in the security aspects. This section reflects on the results, identifies limitations, and explores the broader implications of adopting secure chatbot architectures.

Security vs. Usability Trade-Offs : The framework primarily focused on significant improvements against known threats, but it has to slightly compromise on message latency and computational overhead. However, user feedback showed that the trade-off was acceptable when the security benefits were made transparent for instance, when users were informed that encryption or verification checks were in progress. Future versions can optimize performance through:

- Hardware-accelerated encryption
- Caching ZKP verifications
- Offloading trust evaluation to edge devices

Adaptive Trust and UX Implications : The context-aware trust engine performed well in dynamically adjusting access levels and triggering verification steps. However, one concern is false positives, where legitimate users are challenged

unnecessarily, risking user frustration or disengagement.Designing trust models that are explainable and user-friendly is key to long-term adoption. Further work is needed to balance system confidence with conversational transparency.

Ethical and Regulatory Considerations : As chatbot systems increasingly handle personal, financial, and medical data, compliance with frameworks such as GDPR (EU), HIPAA (US), CCPA (California) becomes essential. Our framework supports these regulations by:

- Avoiding long-term storage of identifiable data
- Encrypting all in-transit and at-rest messages
- Providing audit trails for consent and access

Still, integration with third-party tools (e.g., CRMs, cloud analytics) must be evaluated on a case-by-case basis to ensure end-to-end compliance[9].

5.1 Limitations :

- Prototype Scale: Evaluation was performed in controlled environments and may not reflect scale-dependent challenges.
- ZKP Limitations: Current ZKP tools are computationally expensive and may not be suitable for low-power devices or real-time edge applications.
- User Education: Many security features rely on user understanding or action (e.g., verifying session notices or accepting encryption), which introduces human factors that are harder to control.

5.2 Future Work

- Integration with biometric tools to enhance the threat monitoring, anomaly detection etc.,
- Develop user-facing tools to visualize trust scores and risk ratings
- Extend the anomaly detection module that complies with federal regulations and help the bots learn to adapt without centralized data
- Pilot the framework in real industry deployments (e.g., banking, mental health chatbots) and come up with Trail and Test methods to compare the results and strengthen the overall capabilities of the chatbot[8].

6. Conclusion

Chatbots equipped with AI are becoming more popular in the critical sectors like healthcare, finance etc., the need for secure, privacy-respecting communication has never been more urgent. This paper brings a novel secure messaging framework that primarily focuses to address the vulnerabilities inherent in chatbot communication systems. The tests conducted under simulated environments proved that this framework has the capability to mitigate the security threats, but there will be some overheads like using high-end hardware for performance and throughput[7]. The framework offers an industry-ready framework by integrating end-to-end encryption, zero-knowledge authentication, context-aware trust scoring, and real-time anomaly detection. The experiments and evaluations gave us a picture that the system maintains good levels of latency and user satisfaction.

It also proved that this system significantly reduced the risk of data breach, prompt injection attacks and impersonation. This paper also stresses on giving importance to aligning the security architecture with real-time adaptability, user experience and regulatory compliance. The framework offers a solid foundation to a secure chatbot system that prioritizes user trust and data sovereignty, but there are limitations like computational overhead, need for a broader deployment testing etc., This paper provides a framework that can serve as a blueprint for developers and researchers who are aiming to build AI-driven conversational systems that are highly intelligent, secure, resilient and align with ethics and regulatory compliance.

7. Conflicts of Interest

The author declares that there is no conflict of interest concerning the publishing of this paper.

References

[1] Vechev et al., “AI Chatbots Can Guess Your Personal Information From What You Type,” Wired, Sep. 2023. [Online]. Available: https://www.wired.com/story/ai-chatbots-can-guess-your-personal-informationWIRED

[2] J. Giordani, “Mitigating Chatbots AI Data Privacy Violations in the Banking Sector: A Qualitative Grounded Theory Study,” Eur. J. Appl. Sci. Eng. Technol., vol. 2, no. 4, pp. 1–12, 2024. [Online]. Available: https://ejaset.com/index.php/journal/article/view/77EJASET

[3] Hariri, “Privacy and Data Protection in ChatGPT and Other AI Chatbots,” SSRN, Jul. 2023. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4454761Academia+2SSRN+2IGI Global+2

[4] S. K. Sharma, “Evaluating Privacy, Security, and Trust Perceptions in Conversational AI Systems,” Comput. Hum. Behav., vol. 150, pp. 107–118, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563224002127ScienceDirect

[5] Følstad et al., Chatbot Research and Design: 7th International Workshop, CONVERSATIONS 2023, Oslo, Norway, Nov. 2023. Cham, Switzerland: Springer, 2023. [Online]. Available: https://link.springer.com/book/10.1007/978-3-031-54975-5SpringerLink

[6] M. Vechev et al., “A Systematic Literature Review of Information Security in Chatbots,” Appl. Sci., vol. 13, no. 11, p. 6355, 2023. [Online]. Available: https://www.researchgate.net/publication/370998368ResearchGate

[7] Fujitsu, “Fujitsu Launches New Technologies to Protect Conversational AI from Hallucinations and Phishing,” Press Release, Sep. 2023. [Online]. Available: https://www.fujitsu.com/global/about/resources/news/press-releases/2023/0926-02.htmlFujitsu

[8] D. Miller et al., “2023 Conversational AI Intelliview: Decision-Makers Guide to Enterprise Intelligent Assistants,” Opus Research, Oct. 2023. [Online]. Available: https://opusresearch.net/pdfreports/2023_ConversationalAI_Intelliview_leadup.pdfopusresearch.net

[9] M. Jo et al., “AI Privacy in Context: A Comparative Study of Public and Institutional Perceptions,” Social Media + Society, vol. 10, no. 1, pp. 1–12, 2024. [Online]. Available: https://journals.sagepub.com/doi/pdf/10.1177/20563051241290845