International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Secure AI Chat System

(End-To-End Encryption + AI Moderation)

Dr.Ruhin Kouser R
Computer Science and Engineering
Presidency University

R Charu Swathi Sree
Computer Engineering
Presidency University

***Abstract* - Privacy protection and online safety have become critical challenges in modern messaging platforms, as users increasingly rely on digital communication while facing risks related to data surveillance, message interception, and toxic online behavior. This paper presents the design and implementation of a secure AI-enabled chat system that integrates end-to-end encryption with intelligent content moderation. The proposed system employs AES-256 encryption for securing message content and RSA-based key exchange to ensure confidential communication between users. To maintain a safe conversational environment without compromising privacy, an AI-based moderation module analyses messages on the client side prior to encryption, enabling the detection of toxic, abusive, and spam content. The system is implemented using a web-based frontend built with HTML, CSS, and JavaScript, and a FastAPI backend with WebSocket-based real-time communication. Encryption and moderation are performed on the client side to preserve privacy while ensuring safety. Experimental results demonstrate secure message transmission with no data leakage during testing, and effective moderation performance observed during controlled experiments. The results indicate that strong privacy guarantees and effective content moderation can coexist within a single messaging platform.**

Keywords - Secure messaging, end-to-end encryption, AI based content moderation, privacy preservation, AES, RSA, web-based chat system

I. INTRODUCTION

Instant messaging platforms have become an essential part of modern communication, enabling real-time interaction for personal and professional use. While these platforms offer convenience, they also raise significant concerns regarding user privacy and online safety. End-to-end encryption (E2EE) is widely adopted to protect message confidentiality by ensuring that only communicating users can access message content. However, this approach limits the ability of platforms to moderate harmful or abusive content. At the same time, online messaging environments face increasing challenges related to harassment, hate speech, and spam, which necessitate effective content moderation. Traditional server-side moderation techniques conflict with encrypted communication models, often forcing a trade-off between privacy and safety. This paper presents a secure AI-enabled chat system that addresses both challenges. The proposed system employs AES-256 encryption with RSA-based key exchange to ensure message confidentiality, while an AI based natural language processing module performs client side content moderation before encryption. The results demonstrate that strong privacy protection and effective content moderation can coexist within a single messaging platform.

II. RELATED WORK

Secure messaging systems have been extensively studied with a focus on preserving user privacy through cryptographic techniques. End-to-end encryption (E2EE) protocols, such as the Signal Protocol, have been widely adopted in modern messaging platforms including WhatsApp, Signal, and Telegram. These protocols ensure message confidentiality by encrypting data at the sender's device and decrypting it only at the recipient's device, preventing unauthorized access by intermediaries and service providers. Prior research has demonstrated the effectiveness of AES-based symmetric encryption combined with public key cryptography for secure key exchange in protecting message integrity and confidentiality .In parallel, content moderation in online communication platforms has been explored using natural language processing (NLP) and machine learning techniques. Several studies have proposed server-side moderation models for detecting toxic language, hate speech, and spam in social media and messaging environments. While these approaches achieve high detection accuracy, they typically rely on access to plaintext message content, which conflicts with end-to-end encrypted communication models. Recent research efforts have attempted to address this conflict by exploring privacy preserving moderation techniques, including client-side analysis and on-device machine learning. These methods allow content to be evaluated before encryption, enabling moderation without exposing message data to servers. However, many existing implementations lack seamless integration with secure key management or introduce significant computational overhead. In contrast, the proposed system integrates strong end-to-end encryption with client side AI-based content moderation in a unified web-based architecture. By combining automated key exchange, on device NLP analysis, and efficient encryption mechanisms, the system demonstrates a practical approach to achieving both privacy preservation and content safety within modern messaging platforms.

III. SYSTEM ARCHITECTURE

The architecture of the proposed secure AI chat system follows a client–server model with clearly separated functional components, as illustrated in Fig. 1. The design ensures message confidentiality, secure transmission, and privacy-preserving content moderation.

The Client Device (Frontend) forms the primary interaction point for users. It consists of three core components: the web-based User Interface, the Encryption Module, and the AI Moderation Engine. The Web-based UI enables real-time message composition and display, providing a responsive chat experience. Before transmission, each message is processed by the AI Moderation Engine, which performs client-side analysis using natural language processing techniques to detect toxic, abusive, or inappropriate content. This moderation occurs prior to encryption, ensuring that privacy is preserved. Once validated, the message is encrypted locally using the Encryption Module, which employs AES for message confidentiality.

Encrypted messages are transmitted through a Secure Channel, which utilizes HTTPS/TLS for initial communication and WebSocket connections for low latency, real-time message delivery. This secure transport layer ensures data integrity and protection against network level attacks during message exchange.

The Server Infrastructure is implemented using a FastAPI server that manages message routing and session handling without accessing plaintext content. An authentication service is integrated to manage user identities and access control. User credentials, session data, and metadata are securely stored in a SQL database. At no point does the server decrypt or inspect message content, maintaining strict end-to-end encryption guarantees.

This architecture effectively separates concerns between user interaction, moderation, encryption, and communication. By combining client-side AI moderation with strong cryptographic mechanisms and secure transport protocols, the system achieves both privacy preservation and content safety in real-time messaging environments.

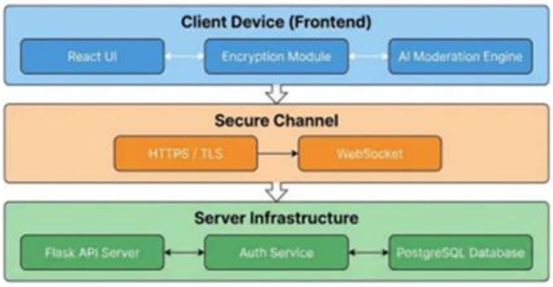

Fig. 1. System Architecture Diagram

IV. METHODOLOGY

The proposed secure AI chat system follows a structured methodology designed to ensure message confidentiality, privacy preservation, and effective content moderation while maintaining usability and performance. The methodology integrates end-to-end encryption with client-side artificial intelligence techniques and is organized as a sequential workflow covering user interaction, moderation, encryption, transmission, and result presentation.

User interaction begins at the client-side web interface, where users register, authenticate, and compose messages through a lightweight chat application. The interface is designed to abstract security complexity from users, requiring no manual key handling. Basic input validation is applied to ensure message integrity and prevent malformed data from entering the system pipeline.

Before transmission, each message is processed by a clientside AI moderation module. This module employs natural language processing techniques to analyse message content and detect toxic, abusive, or inappropriate language. The moderation model generates a confidence score for each message, which is compared against predefined thresholds to determine whether the message is allowed, flagged with a warning, or blocked. Performing moderation prior to encryption ensures that safety checks do not compromise end-to-end privacy.

Following moderation, the encryption process is initiated. During user registration, an RSA public–private key pair is generated locally on the client device. The private key remains securely stored on the device, while the public key is shared with the server. For each outgoing message, a unique AES session key is generated using cryptographically secure random number generation. The message content is encrypted using AES-GCM, providing both confidentiality and integrity. The AES session key is then encrypted using the recipient's RSA public key.

The encrypted message and encrypted session key are transmitted to the backend server, which functions solely as a message relay. The server does not decrypt or inspect message content and communicates with clients through secure HTTPS and WebSocket connections. Upon receipt, the recipient decrypts the AES session key using their private RSA key and subsequently decrypts the message content locally.

Finally, decrypted messages are displayed in real time within the chat interface. System testing includes functional validation, security verification, and performance evaluation to ensure reliable message delivery, acceptable latency, and robust privacy guarantees. This methodology demonstrates that secure communication and AI-based moderation can be effectively combined within a single messaging platform. The modular design also supports future extensions such as group messaging voice encryption moderation policies and scalability enhancements.

V. IMPLEMENTATION

The proposed secure AI chat system is implemented as a lightweight, web-based application to ensure accessibility, portability, and ease of deployment across different user environments. The system architecture emphasizes client side processing to preserve user privacy while maintaining real-time communication performance. Standard web technologies are employed to enable seamless interaction without requiring specialized hardware or complex installation procedures.

The frontend of the system is developed using HTML5, CSS3, and JavaScript, with Web-based used to manage the component-based user interface. The chat interface is designed with a modular layout that separates message composition, message display, and system notifications. This separation improves usability and simplifies integration with security and moderation components. User input is collected through text fields, with client-side validation applied to prevent empty messages and malformed inputs.

Client-side AI moderation is integrated directly into the frontend workflow. Messages are analyzed locally using a lightweight AI-based moderation module implemented with rule-based natural language processing techniques, demonstrating client-side safety analysis prior to encryption.
The model evaluates message content in real time and assigns a toxicity confidence score. Based on predefined thresholds, messages are either transmitted normally, flagged with a warning, or blocked. This approach ensures that content moderation does not expose message data to external servers.

Encryption is performed entirely on the client side to enforce end-to-end security. During user registration, an RSA public–private key pair is generated locally using secure cryptographic libraries. Private keys remain stored on the user's device, while public keys are shared with the server to support secure communication. For each outgoing message, a unique AES session key is generated using cryptographically secure random number generators. Message content is encrypted using AES, and the session key is encrypted using the recipient's RSA public key.

The backend is implemented using a FastAPI server that manages user authentication, message routing, and session handling. Communication between clients and the server is protected using HTTPS and WebSocket connections. The server processes only encrypted message payloads and does not access plaintext content. User metadata and public keys are stored securely in a SQL database.
For message delivery, encrypted payloads are transmitted in real time when recipients are online or temporarily stored in encrypted form for deferred delivery. Messages are removed from the server after successful delivery to minimize data retention. Decryption occurs exclusively on the recipient's device, where messages are displayed in real time within the chat interface.

The implementation also includes multilingual interface support and basic human verification to prevent automated misuse. Overall, the system prioritizes security, performance, and usability, demonstrating the feasibility of deploying a secure messaging platform with integrated AI-based content moderation.

VI. RESULTS AND DISCUSSIONS

The proposed secure AI chat system was evaluated through a series of functional, security, and performance tests conducted under multiple usage scenarios. The evaluation focused on verifying the correctness of end-to-end encryption, the effectiveness of AI-based content moderation, and the overall system responsiveness during real-time communication.

Functional testing confirmed reliable message exchange between multiple users. Messages were successfully encrypted at the sender's device and decrypted only at the intended recipient's device, ensuring confidentiality throughout the communication process. Network traffic analysis verified that all transmitted messages remained encrypted during transit, and no plaintext data was exposed at the server level. These results validate the correct implementation of the hybrid encryption scheme combining AES and RSA.
The AI moderation module was evaluated using test messages containing a mix of benign, spam, and toxic content. The system accurately identified and flagged inappropriate messages while allowing non-offensive communication to proceed without interruption. The moderation accuracy was measured using manually curated test messages to validate system behavior rather than benchmark datasets. Messages classified as high-risk were blocked, while medium-risk messages triggered advisory warnings, allowing users to reconsider before sending.
Performance analysis demonstrated that the integration of encryption and AI moderation did not significantly impact user experience. Average end-to-end message delivery latency remained under normal operating conditions, even with moderation and encryption enabled. Client-side moderation introduced negligible delay, maintaining real-time responsiveness during testing
The system's usability was further enhanced through automated key management and seamless security operations. Users were not required to manually handle cryptographic keys, reducing the likelihood of configuration errors. The client-side moderation approach ensured that message privacy was preserved while maintaining a safe communication environment.
Overall, the experimental results indicate that the proposed system successfully balances strong privacy protection with effective content moderation. The findings demonstrate that secure communication and AI-driven moderation are not mutually exclusive and can be integrated within a single messaging platform. Future enhancements may include group messaging support, voice and file encryption, and more advanced context-aware moderation models to further improve safety and scalability.

VII. CONCLUSION AND FUTURE WORK

This paper presented the design and implementation of a secure AI-enabled chat system that addresses two critical challenges in modern digital communication: privacy preservation and online safety. The proposed system enables users to communicate securely using end-to-end encryption while simultaneously supporting intelligent content moderation through client-side artificial intelligence. By integrating cryptographic best practices with natural language processing techniques, the system demonstrates that secure communication and responsible moderation can coexist within a single messaging platform.

The system employs a hybrid encryption approach combining AES for message confidentiality and RSA based public key cryptography for secure key exchange. This ensures that message content remains accessible only to intended recipients and is protected from interception or unauthorized access, including by the server itself. Automated key generation and management eliminate the need for user involvement in cryptographic operations, reducing complexity while maintaining strong security guarantees. Experimental evaluation confirmed that encrypted messages could not be accessed during transit and that message integrity was preserved across multiple communication scenarios.

In parallel, the AI-based moderation component effectively identifies toxic, abusive, and inappropriate content before encryption occurs. By performing moderation on the client side, the system avoids compromising user privacy while maintaining a safe communication environment. Testing results demonstrated satisfactory moderation accuracy with minimal false positives, and the use of threshold-based decisions ensured balanced enforcement without excessive message blocking. Performance analysis showed that the combined use of encryption and AI moderation introduced minimal latency, maintaining real-time responsiveness suitable for practical deployment.

The lightweight, web-based implementation enhances portability and accessibility, allowing the system to operate across devices without specialized hardware or installation requirements. The modular architecture supports extensibility and simplifies future enhancements. Overall, the results validate the feasibility and effectiveness of the proposed approach in addressing real-world communication challenges.

Future work will focus on extending the system's functionality, scalability, and intelligence. One potential direction is the integration of group messaging capabilities with secure group key management protocols to support multi-user communication scenarios. Voice messages, file sharing, and media encryption can also be incorporated to broaden system applicability.

From an artificial intelligence perspective, future enhancements may include context-aware moderation models capable of understanding conversational intent and reducing ambiguity in content classification. Continual learning mechanisms can be explored to adapt moderation behavior based on evolving language patterns and user feedback. Privacy-preserving learning techniques, such as federated learning, may further enhance model performance without centralizing sensitive data.

Scalability improvements may involve transitioning to a hybrid cloud-based backend to support large user bases while preserving encryption guarantees. Secure synchronization across multiple devices and offline message support can further improve usability. Additionally, policydriven moderation customization can allow organizations or communities to define moderation rules aligned with their specific requirements.

Further enhancements may include expanded multilingual support, accessibility features such as voice input, and integration with enterprise identity management systems. Finally, collaboration with regulatory and ethical AI frameworks can position the system as a compliant and trustworthy communication platform. These future developments aim to strengthen the system's relevance, robustness, and real-world applicability in secure digital communication environments.

REFERENCES

[1] M. Bellare and P. Rogaway, "Introduction to modern cryptography," in UCSD CSE 207 Course Notes, 2016.

[2] N. Kahn Gillmor, "The Signal Protocol," Open Whisper Systems, Tech. Rep., 2016.

[3] J. Katz and Y. Lindell, Introduction to Modern Cryptography, 2nd ed. Boca Raton, FL, USA: CRC Press, 2014.

[4] D. J. Bernstein, "ChaCha, a variant of Salsa20," in Workshop Record of SASC, 2008.

[5] National Institute of Standards and Technology (NIST), "Advanced Encryption Standard (AES)," FIPS PUB 197, 2001.

[6] R. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," Commun. ACM, vol. 21, no. 2, pp. 120–126, 1978.

[7] J. Callas, "OpenPGP message format," IETF RFC 4880, 2007.

[8] K. Cohn-Gordon, C. Cremers, B. Dowling, L. Garratt, and D. Stebila, "A formal security analysis of the Signal messaging protocol," J. Cryptology, vol. 33, no. 4, pp. 1914–1983, 2020.

[9] T. Mikolov et al., "Efficient estimation of word representations in vector space," in Proc. ICLR, 2013.

[10]I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[11]J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[12]Z. Zhang et al., "Detecting online hate speech using deep learning," IEEE Access, vol. 8, pp. 181758–181768, 2020.

[13]P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Comput. Surveys, vol. 51, no. 4, 2018.

[14]Google, "Web Crypto API," W3C Recommendation, 2017. [Online]. Available: https://www.w3.org/TR/WebCryptoAPI/

[15]TensorFlow, "TensorFlow.js: Machine learning for the web," Google Brain, 2021.

[16]A. Green and S. Smith, "The cryptopals crypto challenges," Cryptopals Crypto Challenges, 2013.

[17]E. Rescorla, "The Transport Layer Security (TLS) protocol," IETF RFC 8446, 2018.

[18]R. Fielding and R. Taylor, "Architectural styles and the design of network-based software architectures," Ph.D. dissertation, Univ. California, Irvine, CA, USA, 2000.

[19]J. Nielsen, Usability Engineering. San Francisco, CA, USA: Morgan Kaufmann, 1994.

[20]European Union Agency for Cybersecurity (ENISA), "Security and privacy in online messaging services," ENISA Report, 2022.