Global Journal of Engineering Innovations & Interdisciplinary Research

Original Article

Conversational AI Based Assistant For Detecting And Responding To Phishing Attacks

G Raju, Panga Sravani, Gudisili Suresh, Shaik Sameera Jasmine, Kamurthi Vardhan, Chinthalarevu Supriya

Department of Computer Science and Engineering Gates Institute of Technology , Andhra Pradesh, India.

Correspondence

G Raju

Department of Computer Science and Engineering, Gates Institute of Technology, Andhra Pradesh, India.

- Received Date: 11 Feb 2026
- Accepted Date: 02 Apr 2026
- Publication Date: 25 Apr 2026

Keywords

Conversational AI, Cybersecurity Awareness, Large Language Models, Phishing Detection, QR Code Phishing ..

Copyright

© 2026 Authors. This is an open- access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Abstract

Phishing attacks have become increasingly sophisticated, exploiting human vulnerabilities through social engineering techniques such as urgency, authority, impersonation, and fear to deceive users into revealing sensitive information. Conventional phishing detection mechanisms, including blacklist-based filters and rule-driven heuristic systems, depend largely on predefined signatures, known malicious patterns, and surface-level technical indicators. While effective against previously identified threats, these approaches struggle to detect zero-day attacks, obfuscated URLs, context-aware impersonation, and emerging threats such as QR-code–based phishing, commonly referred to as quishing, and they offer limited transparency or user awareness. To overcome these limitations, this work proposes a conversational AI–based, human-centric assistant that actively supports users in detecting, understanding, and responding to phishing attacks through interactive dialogue. The system leverages Large Language Models integrated via LangChain to analyze phishing attempts across multiple vectors, including email text, embedded hyperlinks, and QR codes, with a strong emphasis on semantic and contextual analysis rather than simple pattern matching. Instead of producing a binary classification, the assistant identifies and explains key phishing indicators such as psychological manipulation cues, urgency triggers, authority abuse, and threat-based consequences using clear natural language. In addition, the system prioritizes data privacy through local processing and incorporates secure user authentication, administrative control, and OTP-based account recovery mechanisms. Experimental observations indicate that the proposed approach adapts effectively to evolving phishing strategies, enhances user awareness, and improves decision-making, demonstrating its potential as an explainable, adaptive, and practical AI-driven solution for addressing modern cybersecurity challenges in real-world environments across diverse digital communication platforms and user environments.

Introduction

Among the various forms of cybercrime, phishing stands out as one of the most prevalent and harmful threats in modern times [5]. According to the Anti-Phishing Working Group (APWG), over 1.2 million phishing attacks were recorded in the second quarter of 2023 alone [6]. Every day, an estimated 3.4 billion phishing emails are sent by cybercriminals, amounting to over one trillion emails per year [7]. Phishing serves as the entry point for approximately 91% of all cyberattacks and is involved in about 36% of all data breaches, making it one of the most common causes of security incidents.

Social engineering lies at the core of phishing attacks, exploiting inherent human tendencies to trust and respond to certain stimuli [1], [3], [19]. Attackers use psychological manipulation to deceive individuals into performing actions that compromise security, such as clicking on malicious links or providing confidential information including login credentials, financial details, or personal data [2], [21]. This manipulation can have severe consequences, leading to identity theft, financial losses, and compromised personal or organizational security.

Human error plays a significant role in the success of phishing attacks, contributing to nearly 95% of successful cybersecurity breaches [8]. Factors such as lack of awareness, inadequate training, stress, fatigue, and cognitive overload can impair an individual's ability to recognize and respond appropriately to phishing attempts [9], [10], [17]. Attackers exploit these vulnerabilities by crafting messages that take advantage of distraction, curiosity, or the tendency to comply with authority figures [4], [11], [20], [37]. These challenges highlight the importance of implementing automated detection systems that act as a first line of defense against human weaknesses.

However, many existing phishing detection solutions focus mainly on technical indicators

Citation: Raju G, Panga S, Gudisili S, Shaik SJ, Kamurthi V, Chinthalarevu S. Conversational AI Based Assistant For Detecting And Responding To Phishing Attacks. GJEIIR. 2026;6(4):244.

and often fail to provide clear explanations that non-technical users can understand. As a result, users may not fully comprehend why a message is considered suspicious, limiting their ability to learn from the detection process and improve their awareness.

To address this issue, this project proposes an AI-based conversational assistant designed to help users detect, understand, and respond to phishing emails. It leverages Large Language Models (LLMs) to analyze the semantic features of email content and identify psychological manipulation tactics that traditional systems may overlook. By focusing on the meaning and context of messages, it provides accurate threat assessments at the sentence level.Furthermore, users can interact with it through a conversational interface to clarify doubts about the analysis, enabling them to gain deeper insights into why an email is flagged as phishing or safe. The system also integrates a user-centered interface that effectively communicates risks while safeguarding user privacy through local data processing. These features significantly improve the effectiveness, usability, and acceptance of phishing detection systems in real-world environments.

Related Work

Phishing detection has evolved significantly over the years as cyberattacks have become more sophisticated and targeted. Traditional approaches initially focused on blacklist-based detection techniques, where known malicious URLs and domains were stored in centralized databases. These systems are efficient in identifying previously reported phishing websites but fail to detect newly generated or obfuscated attacks, making them ineffective against zero-day threats.

To address these challenges, researchers proposed heuristic and rule-based approaches, which analyze predefined patterns such as suspicious keywords, abnormal URL structures, and mismatched domain names. Although these methods improve detection to some extent, they lack adaptability and require continuous manual updates to remain effective.

The introduction of machine learning (ML) techniques marked a significant advancement in phishing detection. Algorithms such as Support Vector Machines (SVM), Naïve Bayes, Decision Trees, and Random Forests have been widely used to classify phishing emails and malicious URLs. These models rely on features such as URL length, domain age, HTML structure, and lexical patterns. Despite achieving good accuracy, these approaches are heavily dependent on feature engineering and often fail to capture the semantic intent of phishing messages.

Recent advancements in deep learning have further improved phishing detection systems. Techniques such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks have been applied to analyze sequential and textual data. These models can automatically extract features and identify complex patterns in phishing content. However, deep learning models often require large datasets and lack interpretability, making them less suitable for user-centric applications .

With the growth of Natural Language Processing (NLP), research has shifted towards analyzing the linguistic and contextual features of phishing emails. NLP-based approaches can detect psychological manipulation techniques such as urgency, fear, and impersonation. These systems provide better insights compared to traditional ML models but still lack the ability to interact with users and provide real-time explanations.

The emergence of Large Language Models (LLMs), such as transformer-based architectures, has opened new possibilities in cybersecurity. These models can understand context, semantics, and intent at a deeper level, enabling more accurate detection of sophisticated phishing attacks. Studies have shown that LLMs can significantly improve detection performance while also generating human-readable explanations.

In parallel, conversational AI systems have been introduced to enhance user awareness and engagement. Chatbot-based systems allow users to interact with the system, ask questions, and receive guidance on identifying phishing threats. However, most existing chatbot solutions are limited to text-based analysis and do not support multi-modal inputs, such as QR codes or image-based phishing attacks .

Another emerging threat is QR code-based phishing (QRishing), where attackers embed malicious links within QR codes to bypass traditional security filters. Research in this area is still limited, and existing systems often lack effective mechanisms to analyze and detect QR-based threats .

Furthermore, many existing systems focus solely on detection and do not provide explainable outputs, which are crucial for improving user awareness and decision-making. Explainable AI (XAI) has recently gained attention for addressing this issue by providing transparent and interpretable results.

To overcome these limitations, the proposed system integrates Artificial Intelligence, NLP, and conversational AI into a unified framework. Unlike traditional systems, it performs semantic analysis of email content, supports QR code-based threat detection, and provides interactive explanations through a chatbot interface. This combination enhances both detection accuracy and user understanding, making the system more effective and practical for real-world applications.

Proposed Methodology

The proposed system presents a Conversational AI-based phishing detection framework designed to identify and respond to phishing attacks in email communications and QR-based inputs. The methodology integrates Artificial Intelligence, Natural Language Processing (NLP), and conversational interfaces to provide both accurate detection and user-friendly explanations. Unlike traditional systems that rely on static rules or blacklists, the proposed approach focuses on semantic analysis and user interaction, making it more effective against modern and evolving phishing techniques.

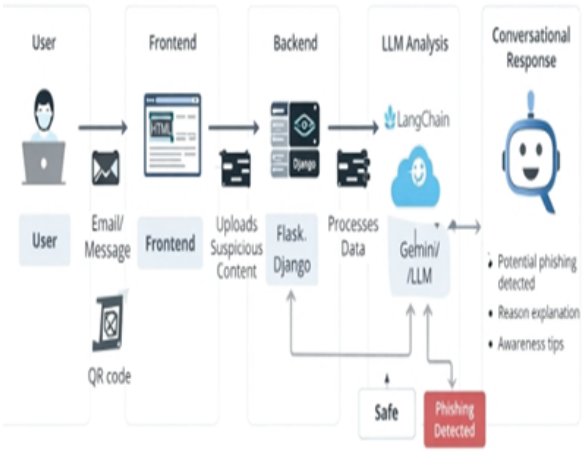

Fig 1. *Architecture of User Interaction and Query Processing Flow*

The system follows a web-based client–server architecture, where users interact through a browser interface developed using HTML, CSS, and Django. The methodology begins with user authentication, where users register and log in to access system functionalities. Once logged in, users can submit suspicious email content or upload QR code images for analysis. This input is processed by the backend system, which acts as the central controller for all modules.

In the initial stage, the system performs data preprocessing, which includes cleaning the input text, removing unnecessary characters, and extracting relevant information such as URLs from email content. In the case of QR code inputs, the system uses image processing techniques (OpenCV) to decode the QR code and extract the embedded URL. This ensures that all inputs are converted into a structured format suitable for further analysis.

The core component of the methodology is the AI-based phishing detection module, which utilizes a Large Language Model (LLM) integrated through Lang Chain. The model analyzes the semantic content of the email rather than relying only on surface-level features. It identifies phishing indicators such as urgency, impersonation, suspicious requests, misleading language, and abnormal link structures. Based on this analysis, the system classifies the input as phishing or legitimate.

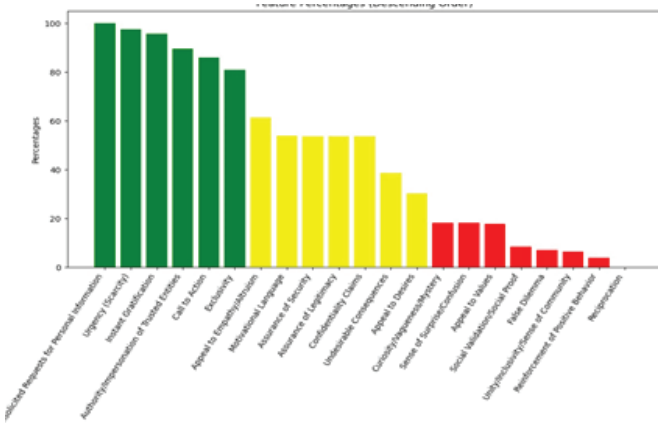

Fig. 2. Final configuration for semantic features detection accuracy

This approach improves detection accuracy, especially for zero-day and sophisticated phishing attacks.

To enhance the reliability of detection, the system also considers link-based analysis by evaluating extracted URLs for suspicious patterns. In QR code analysis, the decoded URL is directly passed to the AI model for classification. The system categorizes QR-based inputs into safe, suspicious, or phishing, thereby addressing emerging threats like QRishing.

An important feature of the proposed methodology is the integration of Explainable AI (XAI). Instead of providing only classification results, the system generates a clear explanation describing why the input is considered phishing or safe. This explanation is displayed on the user interface, helping users understand the reasoning behind the decision. This improves user awareness and reduces dependency on blind trust in automated systems.

The system further incorporates a conversational chatbot module, which allows users to interact with the system after receiving the analysis results. The chatbot uses the same AI model to answer user queries based on the analyzed content. It provides additional explanations, cybersecurity guidance,

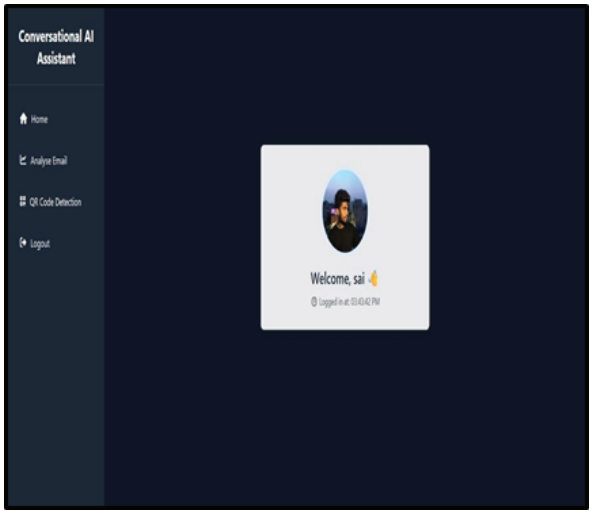

Fig 3: *User Dashboard of the project*

and safe practices, making the system more interactive and educational.

All system data, including user details, analysis results, and session information, are managed using a database through Django's ORM. Security features such as admin approval for user accounts, session management, and OTP-based password recovery are implemented to ensure secure access and data protection.

The overall workflow of the system starts with user input, followed by preprocessing, AI-based semantic analysis, classification, explanation generation, and result display. Finally, users can interact with the chatbot for further clarification. This end-to-end process ensures efficient phishing detection while also improving user understanding and engagement.

The proposed methodology combines semantic analysis, AI-based classification, QR code detection, and conversational interaction to create a comprehensive phishing detection system. By focusing on both technical accuracy and user awareness, the system provides an effective and practical solution for real-world cybersecurity challenges.

The performance of the proposed Conversational AI-based phishing detection system was successfully implemented as a web-based application using Django and Python. The system integrates multiple modules, including email analysis, QR code detection, and a conversational chatbot interface, to provide a comprehensive solution for phishing detection and user awareness.

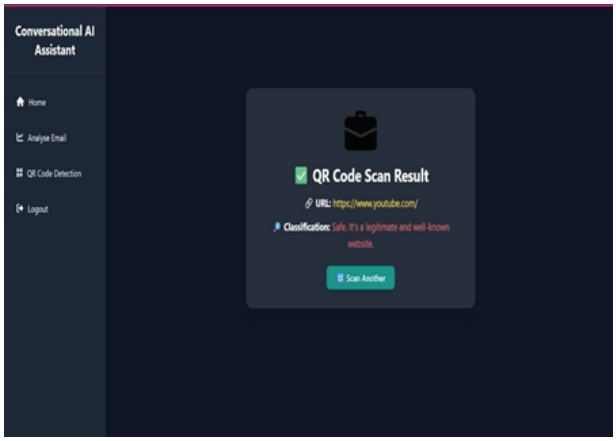

Fig 4 : QR Code Scan Result of our project

The system demonstrated effective performance in detecting phishing emails by analyzing the semantic content of the input. When users submitted suspicious email text, the system was able to classify it as phishing or legitimate and provide clear explanations highlighting key indicators such as urgency, impersonation, and suspicious links. This shows that semantic analysis using a Large Language Model significantly improves detection capability compared to traditional rule-based approaches.

The QR code detection feature further enhanced the system by addressing QR-based phishing (QRishing). The system successfully decoded QR codes using image processing techniques and extracted embedded URLs for analysis. Based on this analysis, the system categorized the QR content as safe or malicious, thereby extending protection beyond text-based phishing attacks.

The conversational chatbot module improved the overall usability of the system. Users were able to interact with the system, ask questions about the analysis, and receive meaningful responses. This interactive feature not only improves user engagement but also helps in educating users about phishing threats and safe practices.

The results also highlight the importance of Explainable AI (XAI) in phishing detection. By providing detailed explanations along with classification results, the system increases transparency and helps users understand the reasoning behind decisions. This builds user trust and supports better decision-making.

Despite these advantages, certain limitations were observed. The system's performance depends on the quality of input data and the effectiveness of prompt design for the AI model. In cases where be fully accurate. Additionally, real-time processing may Introduce slight delays due to AI model response time.

Overall, the proposed system demonstrates strong performance in detecting phishing attacks while also enhancing user awareness through explanation and interaction. The integration of AI, NLP, QR code analysis, and conversational features makes the system a practical and effective solution for modern phishing detection challenges.

Conclusion

This work presents an AI-powered conversational framework for enhancing user awareness and protection against phishing attacks. The system integrates multiple components, including secure user authentication with administrator control, phishing email analysis, QR code threat detection, and a chatbot-based guidance mechanism. By leveraging the capabilities of Google Gemini through LangChain and combining them with computer vision techniques using OpenCV, the framework effectively analyzes suspicious inputs and provides meaningful explanations along with security recommendations.

The implementation ensures secure access through verified user authentication and supports efficient user management via an administrative dashboard. Experimental evaluation demonstrates that the framework accurately detects phishing emails, classifies malicious and legitimate content, processes QR-based threats, and handles invalid inputs effectively while generating appropriate responses.

In addition to threat detection, the system emphasizes user awareness through an interactive conversational interface. This feature enables users to understand phishing techniques and make informed decisions when encountering suspicious content. Thus, the framework serves both as a detection tool and an educational platform for improving cyber security awareness.

Overall, the proposed approach provides an intelligent, reliable, and user-friendly solution for addressing modern phishing threats. Future enhancements may include real-time browser integration, multilingual support, improved model optimization, and scalable cloud deployment.

References

1. XF. Salahdine and N. Kaabouch, "Social engineering attacks: A survey," Future Internet, vol. 11, no. 4, 2019, doi: 10.3390/fi11040089.
2. G. Desolda, L. Ferro, A. Marrella, M. Costabile, and T. Catarci, "Human factors in phishing attacks: A systematic literature review," ACM Computing Surveys, vol. 54, 2022, doi: 10.1145/3469886.
3. O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," Expert Systems with Applications, vol. 117, pp. 345–357, 2019, doi: 10.1016/j.eswa.2018.09.029.
4. A. Muneer, R. F. Ali, A. A. Al-Sharai, and S. M. A. Fati, "A survey on phishing email detection techniques," in Proc. Int. Conf. Innovative Computing (ICIC), 2021, pp. 1–6, doi: 10.1109/ICIC53490.2021.9692960.
5. N. Altwaijry, I. Al-Turaiki, R. Alotaibi, and F. Alakeel, "Advancing phishing email detection: A comparative study of deep learning models," Sensors, vol. 24, 2024, doi: 10.3390/s24072077.
6. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
7. A. Vaswani et al., "Attention is all you need," in Proc. NIPS, 2017.
8. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019.
9. F. Heiding, B. Schneier, A. Vishwanath, J. Bernstein, and P. S. Park, "Devising and detecting phishing: Large language models vs. smaller human models," 2023. [Online]. Available: https://arxiv.org/abs/2308.12287
10. S. S. Roy, P. Thota, K. V. Naragam, and S. Nilizadeh, "From chatbots to phishbots? Preventing phishing scams created using ChatGPT, Google Bard and Claude," 2024. [Online]. Available: https://arxiv.org/abs/2310.19181
11. T. Koide, N. Fukushi, H. Nakano, and D. Chiba, "ChatSpamDetector: Leveraging large language models for effective phishing email detection," 2024. [Online]. Available: https://arxiv.org/abs/2402.18093
12. Y. Li et al., "KnowPhish: Large language models meet multimodal knowledge graphs for phishing detection," 2024. [Online]. Available: https://arxiv.org/abs/2403.02253
13. Google Developers, "Google Safe Browsing API." [Online]. Available: https://developers.google.com/safe-browsing
14. AbuseIPDB, "IP Address Abuse Reports." [Online]. Available: https://www.abuseipdb.com/
15. Thunderbird Developers, "WebExtension APIs for Thunderbird." [Online]. Available: https://webextension-api.thunderbird.net/
16. Nagesh, C., Chaganti, K.R. , Chaganti, S. , Khaleelullah, S., Naresh, P. and Hussan, M. 2023. Leveraging Machine Learning based Ensemble Time Series Prediction Model for Rainfall Using SVM, KNN and Advanced ARIMA+ E-GARCH. International Journal on Recent and Innovation Trends in Computing and Communication. 11, 7s (Jul. 2023), 353–358. DOI:https://doi.org/10.17762/ijritcc.v11i7s.7010.
17. S. Khaleelullah, P. Marry, P. Naresh, P. Srilatha, G. Sirisha and C. Nagesh, "A Framework for Design and Development of Message sharing using Open-Source Software," 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2023, pp. 639-646, doi:10.1109/ICSCDS56580.2023.10104679.
18. Ramesh Kumar Ramaswamy, Pannangi Naresh, Chilamakuru Nagesh, Santhosh Kumar Balan, Multilevel thresholding technique with Archery Gold Rush Optimization and PCNN-based childhood medulloblastoma classification using microscopic

images, Biomedical Signal Processing and Control, Volume 107, 2025,107801, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2025.107801.

19. K. R. Chaganti, B. N. Kumar, P. K. Gutta, S. L. Reddy Elicherla, C. Nagesh and K. Raghavendar,"Blockchain Anchored Federated Learning and Tokenized Traceability for Sustainable Food Supply Chains," 2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), Gobichettipalayam, India, 2024, pp. 1532-1538, doi: 10.1109/ICUIS64676.2024.10866271.
20. Mohammed Inayathulla and Karthikeyan C, "Image Caption Generation using Deep Learning For Video Summarization Applications" International Journal of Advanced Computer Science and Applications(IJACSA), 15(1), 2024. http:// dx.doi.org/10.14569/IJACSA.2024.0150155.
21. Mohammed, I., Chalichalamala, S. (2015). TERA: A Test Effort Reduction Approach by Using Fault Prediction Models. In: Satapathy, S., Govardhan, A., Raju, K., Mandal, J. (eds) Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1. Advances in Intelligent Systems and Computing, vol 337. Springer, Cham. https://doi. org/10.1007/978-3-319-13728-5_25
22. Inayathulla, M., Karthikeyan, C. (2022). Supervised Deep Learning Approach for Generating Dynamic Summary of the Video. In: Suma, V., Baig, Z., Kolandapalayam Shanmugam, S., Lorenz, P. (eds) Inventive Systems and Control. Lecture Notes in Networks and Systems, vol 436. Springer, Singapore. https://doi.org/10.1007/978-981-19- 1012-8_18.
23. Mohammed and C. Karthikeyan, "Object detection in video summarization for video surveillance applications," Yugoslav Journal of Operations Research, vol. 35, no. 3, pp. 523–546, 2025. doi:10.2298/YJOR240615010I.
24. Inayathulla Mohammed and K. R. Rao, "Enhancing real-time violence detection in video surveillance using hybrid deep learning model," Journal of Web and User Applications, vol. 16, no. 1, 2025. doi:10.58346/JOWUA.2025.I1.021.