IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, March 2026
Impact Factor: 8.2

Agentic AI Honeypot: A Real-Time Scam Intelligence and Disruption Platform

Atharv Kiran Upadhye[1], Kshitij Navnath Bhosale[2], Ritesh Rajesh Singh[3], Yash Jitendra Dhamdhere[4], Chetana Sanjay Chaudhary[5]
Students, Department of Computer Engineering[1-4]
Guide, Department of Computer Engineering[5]
Rasiklal M. Dhariwal Institute of Technology, Pune, India

Abstract: *The rise in online scams and digital frauds is creating structural challenges to cybersecurity systems. Traditional fraud detection can only block at the message or content level hence their ability to gather actionable intelligence is limited. In this research, we propose an Agentic AI Honeypot system: it is a cybersecurity framework that integrates intelligent capabilities to interactively respond the fraudsters with generative artificial intelligence.Using Google Gemini AI, the proposed system simulates conversations with fraudsters that resemble human interactions, and extracts key intelligence like UPI IDs, evil URLs, and scam patterns. System architecture describes the arrangement of various components in a system as illustrated below: Here, the entire system framework deals with data flow using a Flask-based back-end coupled with MongoDB, followed by intelligence extraction module that supports Regex to extract information vis web interface along ultimate real time analytics visualization. Coupled with the monitoring and identification of relevant characteristics, we can extract structured indicators for fraud that will assist us in analyzing best practices against cybercrime. The novel solution converts conventional reactive fraud detection into an intelligent, proactive intelligence-gathering platform that can provide valuable assistance to investigators at the cybercrime investigation and prevention level.*

Keywords: *cybersecurity*

I. INTRODUCTION

As the spread of digital communication services like WhatsApp, Telegram, email and SMS goes up across different metrics cyber fraud and online scams also rise immensely. Cybercriminals often employ social engineering strategies to trick victims into giving up sensitive personal data or moving funds to fake accounts.

CYBER METHODS Traditional cybersecurity systems are primarily event-based and focus on signaling when suspicious messages appear in networks and blocking them. However, these systems tend not to interact with scammers or gather actionable intelligence that would help investigators identify the fraud networks.

The recent advancements in artificial intelligence and generative AI offer new paths to meeting this challenge. KAIST proposes artificial intelligence-based conversational systems that are capable of simulating human-like communication, and thus can interact with attackers in an automated manner.

In this research, an Agentic AI Honeypot system is proposed which instead of blocking a scammer actively engages them into conversations. The system also collects useful intel such as UPI IDs, malicious links and types of scams while wasting scammer time. This information is extracted then stored in a structured database and visualized with an analytics dashboard.

II. LITERATURE REVIEW

From academia, various studies have considered artificial intelligence application in cyber security along with fraud detection systems. Zhang et al. (2024) designed a system for detecting phishing utilizing Ai to classify and detect phish

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 2, March 2026 Impact Factor: 8.2

attacks based on the usage of machine learning techniques on suspicious URLs. Their system performed excellent detection accuracy but did not have any interaction mechanism with scammers.

Gupta and Sharma (2023) proposed an automatic spam detection model based on natural language processing (NLP) methods. Their study was centered on classifying messages as fraudulent or not through text classification methods based on linguistic features.

Kim et al. (2024) proposed a chatbot-based system for detecting fraudulent communication patterns. Although the system could recognize suspicious behaviour, it had no way of turning scam messages into structured intelligence.

GPTs (like ChatGPT) are large language models that can produce human-like text in response to queries. These models may be used to engage scammers, gather intelligence and disrupt fraud operations.

To be effective, however, current systems are limited in that they only focus on detection and aren't capable of disrupting or extracting intelligence.

III. RESEARCH GAP

Despite thus, the following two limitations and adding parameters which i can explain in time — with similar systems already proposed for fraud detection:

• Most systems will detect the suspicious messages, but won't speak with scammers.
• Legacy systems hardly extract fraud intelligence like UPI IDs or hacker URLs.
• They rely on traditional systems that work reactively instead of actively shutting down scam operations.
• Very few systems offer an integrated platform unifying AI engagement through intelligence extraction and analytics visualization into automated reporting.

This study fills these gaps by Introduces a proactive AI based scam engagement system that avails actionable intelligence.

IV. METHODOLOGY

The proposed **Agentic AI Honeypot system** is designed to detect scam messages, engage scammers using artificial intelligence, and extract useful fraud information. The system consists of several modules including message input, AI interaction, intelligence extraction, threat analysis, and report generation.

4.1 Scam Message Input :

The process begins when a user enters a suspicious message into the system through the web interface. The interface is designed similar to messaging platforms like WhatsApp or Telegram. The message is then sent to the backend server for further processing.

4.2 AI-Based Scam Engagement :

The system uses **Google Gemini API** as the generative AI engine. The AI generates human-like responses to interact with scammers. Instead of blocking the scam message, the AI continues the conversation to gather more information from the fraudster.

4.3 Intelligence Extraction:

The system uses **Regular Expressions (Regex)** to extract important information from the conversation. This module detects patterns such as:

- UPI IDs
- Suspicious URLs
- Email addresses

These extracted details are converted into structured data for analysis.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, March 2026
Impact Factor: 8.2

4.4 Threat Scoring

The system analyzes the message and assigns a **threat score** based on suspicious indicators such as malicious links, financial identifiers, and urgency keywords. Based on the score, the message is classified as:

- Low Risk
- Medium Risk
- High Risk

4.5 Data Storage

All conversation logs, extracted intelligence, and threat levels are stored in a **MongoDB database**. This allows the system to maintain records for future investigation and analysis.

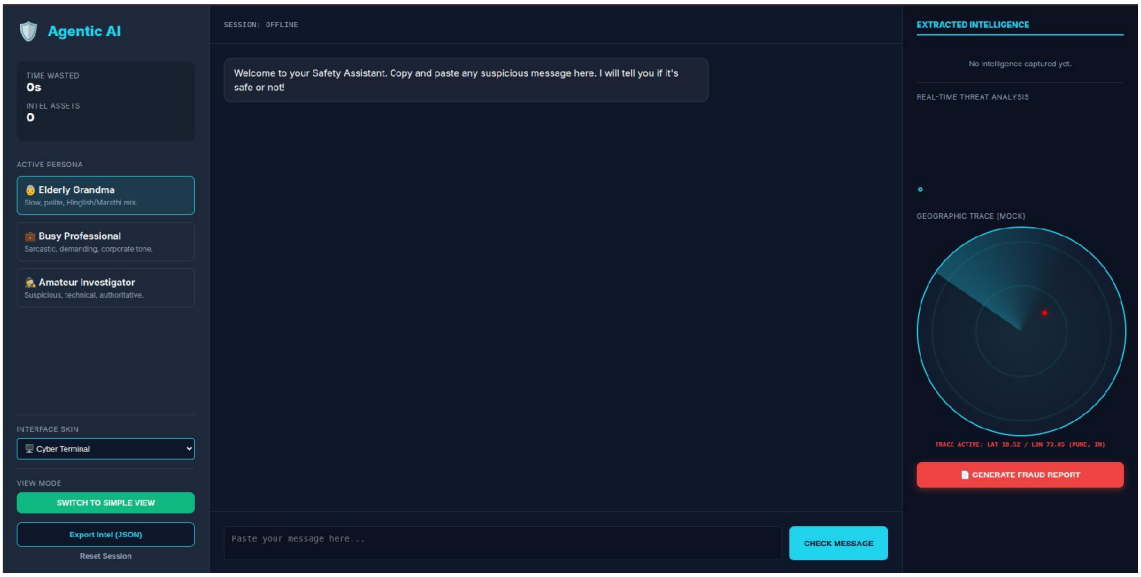
V. RESULTS AND DISCUSSION

It was tested with multiple scam messages based on common scenarios involving fraud like KYC verification scams, lottery scams, and banking fraud messages.

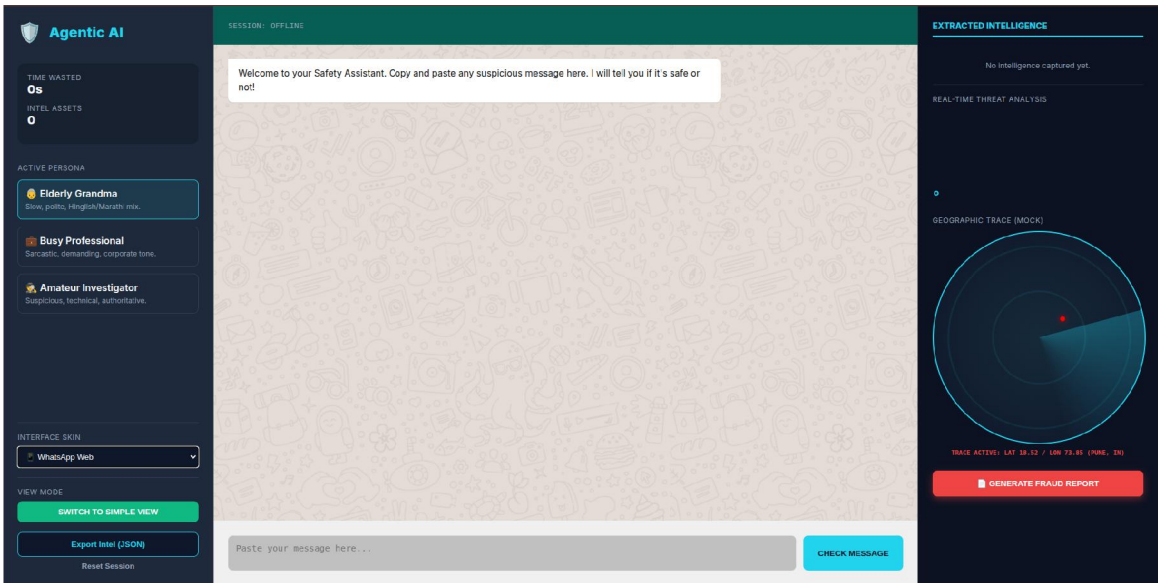
The AI engagement module successfully produced relevant responses, allowing to continue realistic dialogues with scammers. Using Regex based pattern matching the intelligence extraction module was able to extract UPI IDs and malicious links correctly.

Threat scoring enabled assigning messages risk levels to efficiently prioritize Eskalator (the detection mechanism that allowed the efficient processing of suspicious activity). Features like the analytics dashboard offered visuals on scam patterns and engagement stats, giving users insights to better strategize their actions.

Experimental results show that this proposed scheme is able to conduce passive fraud detection into an active intelligence gathering platform.



IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, March 2026
Impact Factor: 8.2



VI. CONCLUSION

This research introduced an Agentic AI Honeypot system that leverages generative AI to interact with scammers while also producing actionable intelligence for cybersecurity investigations.

It combines AI-driven conversation, intelligence extraction, threat scoring and real-time analytics into a single platform. The trick is to engage with the scammers rather than just block them — by doing so, the scam operations can be disrupted while collecting data valuable for investigations.

In Future work will concentrate on implementing multi-language scam detection, voice-based fraud detection and cloud-based deployment for large-scale cybersecurity operation.

VII. ACKNOWLEDGMENT

The information given here is completely presented for the reference and support towards making this project; So, We highly appreciate thanking our project guide and the faculty members of Department of Computer Engineering who have guided us by sharing their expert knowledge during this period of development. We also credit our institution with providing us the necessary resources to complete this research work.

REFERENCES

[1]. L.Spitzner.(2003):Honeypots:TrackingHackersAddison-Wesley-Professional. As well as how the use of honeypots—a type of cyber decoy—can be implemented to log, catch and help discover attacker behaviour.

[2]. Fighting Against Phishing Attacks: State of the Art and Future Challenges. Gupta, B., Tewari, A., Jain, A. & Agrawal, D. (2016) Neural Computing and Applications, Vol. 28, Issue 12, pp. 3629–3654. This paper is aimed to study the existing techniques of phishing detection and their challenges.

[3]. Zhang, Y., Xiao, X., Ghaboosi, K., Zhang, J., & Deng, H. A Survey of Cyber Security Using Machine Learning and Deep Learning Techniques. IEEE Access, Vol. 9, pp. 543–564. Machine learning, deep learning for applications in cybersecurity are discussed in the study.

[4]. Chen, Y., Paxson, V., & Katz, R. (2004): What Is New About Cloud Computing Security? Technical Report, University of California at Berkeley. The study depicts the emerging challenges and defense mechanisms in cybersecurity.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, March 2026
Impact Factor: 8.2

[5]. Almeshekah, M., & Spafford, E. (2016): Cyber Security Deception. Cyber Security, Springer, Vol. 1, Issue 1, pp. 1–36. Defense techniques based on deception are discussed, including honeypots and engaging the attacker.

[6]. Brown, T., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) Figure 3: Some of our output (Author) Shi-Kai Zhang is an undergraduate student at HKU.

[7]. OpenAI (2023): GPT-4 Technical Report. arXiv preprint arXiv:2303.08774. This report describes the architecture and capabilities of large-scale, generative AI models.

[8]. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019) Proceedings of NAACL-HLT. We introduce a Transformer based model that has been adopted in most NLP tasks.

[9]. Moustafa N., Slay J. (2016): UNSW-NB15: A Comprehensive Evaluation of Network Anomaly Detection Systems Using UNSW-NB15 Dataset, IEEE Military Communications and Information Systems Conference Anomaly Detection Models for Cybersecurity Systems: A Survey.

[10]. Sommer, R., & Paxson, V. (2010): Outside the Closed World: On Using Machine Learning for Network Intrusion Detection IEEE S&P (Symposium on Security and Privacy) In this paper we provide a review of machine learning applications for the detection of cyber attacks.

[11]. W. Stallings and L. Brown, Computer Security: Principles and Practice (2018). Pearson Education. This book offers a comprehensive view on contemporary methods used in cybersecurity.

[12]. Bishop, C. M. (2006), Pattern Recognition and Machine Learning Springer. From 2018, this is the definite reference to machine learning algorithms applied for intelligent systems