REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 01 (2024)
http://www.veterinaria.org
Received: 10 December 2023, Revised: 25 December 2023, Accepted: 15 January 2024

"Design and Implementation of a Multi-Layered Framework for Detection and Defense against Logical Attacks in Chatbots: Enhancing the Security and Resilience of Large Language Models"

Vanitaben Hasmukhbhai Suthar[1*], Dr. Swity Maniyar[2], Vanitaben Pragneshkumar Mistry[3]

[1*]PhD Scholar, Swaminarayan University, Kalol, Gujarat. Email: vanitasuthar15184@gmail.com
[2]Associate Professor – GTU, Phd Guide – Swaminarayan University, Kalol, Gujarat.Email: sweetymaniar@gmail.com
[3]Assistant Professor, BCA Department- Sardar Vallabhbhai Global University, Ahmedabad, Gujarat.
Email: vanitamistry@svgu.ac.in

Abstract
The rapid adoption of Large Language Model (LLM)-based chatbots, especially ChatGPT, has introduced significant security challenges due to their vulnerability to logical attacks such as prompt injection, jailbreak attempts, contradiction-based queries, and context manipulation. Existing security approaches mainly rely on keyword filtering and rule-based detection methods, which are insufficient for identifying complex reasoning-based attacks and providing adaptive defense mechanisms. This research proposes a **multi-layered AI security framework** for detecting and defending against logical attacks in ChatGPT-based chatbots. The proposed framework integrates transformer-based semantic analysis, Graph Neural Networks (GNN) for logical reasoning, Contrastive Logical Feature Optimization (CLFO) for improving adversarial feature separation, and Explainable AI (XAI) techniques including LIME, SHAP, and attention visualization for transparent decision-making. The methodology incorporates multi-turn conversational context analysis, hybrid attack classification, dynamic prompt sanitization, adversarial prompt rewriting, and response validation mechanisms. The framework is evaluated using chatbot security datasets containing normal and adversarial prompts across multiple attack categories, with performance measured using accuracy, precision, recall, F1-score, Attack Success Rate Reduction (ASRR), and Defense Effectiveness Score (DES). The proposed approach aims to enhance ChatGPT's robustness, reliability, and trustworthiness by providing an intelligent, explainable, and adaptive defense mechanism against evolving logical attacks.

Keywords: ChatGPT Security, Logical Attacks, Chatbot Vulnerabilities, Prompt Injection, Jailbreak Attacks, Context Manipulation, Large Language Models, Adversarial Attacks, Graph Neural Network, Explainable AI, AI Security Framework, Robustness Enhancement.

Introduction

The fast growth of conversational Artificial Intelligence (AI) technology has transformed human communication with computers because Chat Generative Pre-Trained Transformer (ChatGPT) and similar systems now operate in multiple sectors that include customer support, education, healthcare, and decision support systems [1].

The systems use advanced Deep Learning (DL) technologies, which include transformer-based models to produce responses that match human understanding of the specific situation [2]. Modern chatbots face growing security risks from logical attacks that target their ability to reason and understand context despite the advanced functionalities that these systems provide [3].

Logical attacks comprise various methods like prompt injection, jailbreak attempts, contradiction-based queries, and context manipulation, with adversaries designing inputs to lead the chatbot to produce unintended or unsafe responses, as illustrated in Figure 1 [4].

However, these attacks cannot be considered classic cyber-attacks as they do not target a specific system vulnerability but rather exploit AI models' reasoning and thinking capabilities, which makes them even more difficult to identify [5]. Present security measures mainly revolve around keyword filtering, rule-based systems, or basic adversarial training; however, they fall short in dealing with the complexity and constant change of such attacks [6].

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 01 (2024)
http://www.veterinaria.org
Received: 10 December 2023, Revised: 25 December 2023, Accepted: 15 January 2024

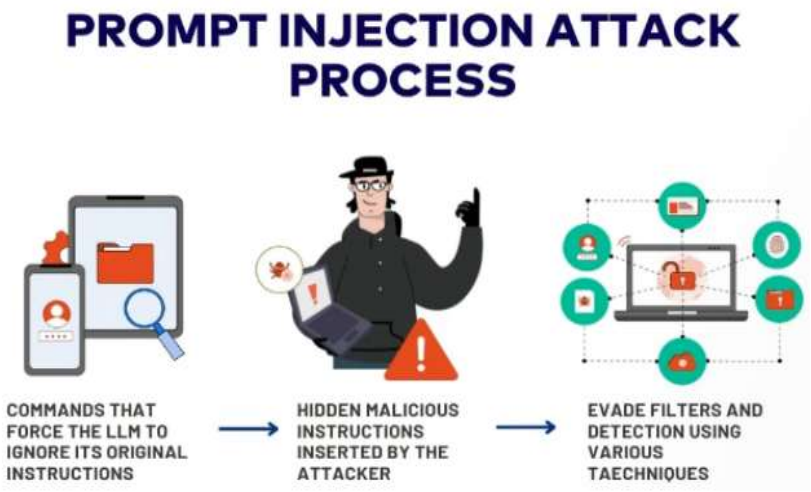


Figure 1: Prompt injection attack process [7]

In addition, most existing approaches don't have a context-aware analysis for the input, as most of them do a one-step analysis with the single input instead of multi-turn conversation reliance, and scant attention to an explainability measure with the ability to decipher why a flaw occurs [8]. The absence of transparent information creates a situation where users lose trust while AI systems fail to establish dependable systems. The system lacks defense mechanisms that can detect and stop malicious prompts from operating in real-time operations [9].
To solve the problem of how to deal with the industry's disadvantages when protecting chatbot systems from logical attacks, a hybrid (advanced) framework is created by combining the four different types of methods used to protect and repair the logic of chatbot systems: 1) Semantic Understanding, 2) Logical Reasoning, 3) Feature Optimization, and 4) Explainable AI. With the help of these four methods in the proposed hybrid framework, chatbot systems can have better reliability and trust. By combining transformer-based models with graph-based reasoning and using dynamic defenses, the proposed hybrid framework can be secure from high-level Adversarial Attacks against next-generation conversational AIs through a complete and comprehensive solution for protecting next-generation Conversational AIs against sophisticated Adversarial Attacks.

Motivation and scope
The increasing use of conversational AI systems, which include ChatGPT for essential functions, has created significant security, reliability, and trust issues. The systems can comprehend and produce human-like responses at an exceptional level, yet they remain vulnerable to attacks that target their reasoning and contextual processing abilities. The security system needs to implement advanced protective measures because prompt injection, jailbreak attempts, and context manipulation attacks have shown security weaknesses.
In addition, current methods primarily focus on superficial filtering and are incapable of identifying deep logical inconsistencies or adversarial reasoning methods. Lack of real-time defense measures and explainability has caused users to trust these systems less. It gives rise to the development of a complete, context-aware, and explainable system that not only can find but also counter logical attacks in practice, thus making the use of chatbot technologies safer.

Scope of study
Utilizing the existing advancement discussed, this study goes further in the creative basin of designing the user detection and system immunity based on a chatbot system against the logical attack:

- Analyzing and classifying the different types of logical attacks, including prompt injection, contradiction-based questions, and context manipulation.
- Creation of a hybrid detection model that uses transformer semantic understanding and logical reasoning in the graph to detect attacks.
- Use of context-aware multi-turn conversation analysis for the purposes of detecting ongoing developments in the logical attacks.
- Create defensive strategies that include adjusting prompts in real-time through dynamic sanitation and adversarial text rewriting.
- Use Explainable AI (Attention, SHAP & LIME) to establish a higher level of transparency and understanding.
- Test the framework presented not just with traditional measuring, but also with robustness measurements.

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 01 (2024)
http://www.veterinaria.org
Received: 10 December 2023, Revised: 25 December 2023, Accepted: 15 January 2024

Literature review
In this section, an elaborated description of various existing studies is presented, and gaps were identified.

Kamata Mitsunobu[10] created two Retrieval-Augmented Generation (RAG)-based chatbot systems, which are named LangChain and Dify, to provide secure municipal services for citizens. The researchers developed five actual testing situations that would assess system security and strength testing performance. The new system showed improved strength and protection abilities when compared to current systems. The LangChain system showed higher protection against attacks because it successfully hid PII data and rejected queries with similarities below 0.75 and stopped Denial of Service (DoS) attacks by implementing rate limits of approximately 5 requests per minute.

Anghel et al. [11] xamined the trustworthiness of Large Language Model (LLM)-based automatic grading when facing answer-side prompt injection attacks. The study implemented a rubric-oriented scoring system through a constant panel of four LLMs, and a rigorous integer-only (010) marking protocol was used on a set of 1000 student answers. The findings revealed a mean absolute error of 1. 897 with a bias of 1. 345, which means the panel was giving higher grades than deserved. Attacks by adversarial manipulations led to considerable score enhancements, where the mean marks changed from 6. 31 (clean) to 9. 88 (A1) and 8. 84 (A2), with average differences of +3. 57 and +2. 53, respectively.
Zhang et al. [12] investigated the security weak points of an ecosystem of LLM autonomous agents. They further proposed a repetitive ClawWorm attack to evaluate efficiency levels across models, incorporating multiple attack types under this framework. The attack results under the experiment part, with an overall attack success rate equal to 64.5%, the most attack effect scheme results in 81% success rate. The variation by model type is from 40% to 84%, and the highest attack persistence against restarting a whole session is equal to 100%, which means a permanent session break.

Sasal and Can [13**]** conducted a rigorous security assessment of large language models with respect to prompt injection attacks. The study compared the security of GPT-4o, Claude 4 Sonnet, and Gemini 2.5 Flash using a custom dataset of 78 attack queries across six categories. The models were evaluated on four metrics: compliance, filtering success, leaking confidential or secure information, and the level of severity of attack to assess how robustly the models could withstand direct threats and indirect threats. The experimental design facilitated a comprehensive examination of the model's behavior under adversarial conditions and showed the difference in the ways different models approach architecture and filtering. Results showed that Claude was the weakest, with the maximum number of high-risk outputs (6 cases) and number of medium-risk outputs (9 cases), followed by GPT-4o, which gave some hits and an occasional inaccurate response, and that Claude had the maximum high-risk outputs, having almost negligible high-risk outputs.

Emekci and Budakoglu [14] introduced their dual-LLM system through ChatTEDU, which enables secure open-access chatbot operation by using two separate models that work together to handle threat detection and response generation. The research tested the framework through real-world interactions while designing it to prevent prompt injection, jailbreak, and adversarial manipulation attacks, which maintained system functionality. The system accomplished 100% attack blocking, which resulted in only 0.28% false positive rate according to the results, thus showing extremely strong security protection that slightly impacted system performance.

Yi et al.[15] proposed a benchmarking framework, called Bipia, and defense mechanisms, such as boundary awareness and explicit reminders, for analyzing and mitigating indirect prompt injection attacks in applications integrated with LLMs. Their method systematically assessed multiple models on different tasks and suggested both black-box and white-box defenses to increase resistance to maliciously instructed hidden content. The findings indicated that the suggested white-box defense almost eliminated the attack success rate, thereby demonstrating its powerful mitigation ability.

Azooz et al.[16] constructed a context-aware multi-layered security framework that included the Context-Aware Prompt Security Scoring System (CA-PSSS) to detect and mitigate prompt injection attacks in real time. The system integrated input validation, output monitoring, and adaptive feedback components to enhance the security of LLMs in adversarial situations. Experiments demonstrated a 98. 3% detection rate with a 2. 7% false positive rate, indicating a high level of effectiveness and reliability.

Liu et al.[17] presented an automatic and generalized attack framework for prompt injection utilizing a gradient-based optimization technique for producing adversarial prompts in various tasks and datasets. The work defined consistent attack targets (static, semi-dynamic, dynamic) and tested the success under a consistent protocol as compared to baseline attacks and prior defenses. The proposed method demonstrated an attack success rate above 80% for static objectives and achieved approximately 50% average success rate across all tasks despite using only 5 training samples, which constituted 0.3% of

the total data. The baseline methods produced zero attack success rate results. The method maintained 85% effectiveness against defense systems, which demonstrated its ability to attack various defenses with strong effectiveness.

Derner et al.[18] offered a taxonomy to methodically categorize security hazards in driven LLM negotiations through the Confidentiality, Integrity, and Availability (CIA) triad and attack entry points throughout the user-model pipeline. The study aimed to pinpoint vulnerabilities and enhance threat comprehension rather than propose a mitigation scheme. It highlighted a wide range of risks, such as data leakage and misappropriation, emphasizing the need for systematic security assessment, but did not report explicit quantitative performance metrics.

Jalali and Chen[19] developed a blockchain-based federated learning system with the use of homomorphic encryption that raised chatbot security and privacy level when transmitting and processing data. The system accomplished trusted communication, decentralized learning, and encrypted operations for protecting confidential user information. Experiments showed that the accuracy rose as high as 90%, and the capability to encrypt a maximum of 1792 MB of data at 1240 times/sec characterizes a great efficiency as well as security improvement.

Research Gap

- **Limited focus on logical-level attacks:** Most existing studies focus on superficial threats such as spam and keyword-based prompt injections, while advanced attacks like logical contradictions, recursive traps, and context manipulation remain underexplored. [11, 12, 17].
- **Lack of hybrid detection models:** Existing approaches are based either on semantic analysis (transformers, etc.) or rules and do not involve logical reasoning (graph modeling, etc.). This leads to a decreased efficiency in fighting back offensive prompts [9,14].
- **Inadequate defense mechanisms:** Most current systems focus on detecting attacks rather than preventing them. They lack proactive mechanisms such as dynamic prompt sanitization or adversarial prompt rewriting, which could reduce the risk of generating harmful outputs [15,17,18].
- **Insufficient explainability in security frameworks:** Many chatbot security models are black boxes with no way to interpret their workings. Explainable AI methods, such as LIME, SHAP, or Attention visualization that could explain the reasons for considering a prompt as malicious, have hardly been utilized in the literature [9,10,11,12].
- **Absence of context-aware analysis:** Most chatbot security models are black boxes, meaning that there is no way to understand how they work. Explainable AI methods like LIME, SHAP, or Attention visualization that would help clarify the reasons for classifying a prompt as malicious have almost been ignored in the literature [9,10,11,12].
- **Lack of standardized robustness evaluation metrics:** Current state-of-the-art approaches use the conventional metrics (accuracy, F1 score) and ignore the metrics specifically aimed at analyzing robustness, like attack success rate reduction or effectiveness of defenses [11,12,14]

Research Problem

The security and reliability of conversational AI systems have become the most important issue since the user adoption of systems such as ChatGPT. The natural language processing technology has achieved significant development, yet current chatbots remain vulnerable to logical attacks, which include prompt injection and context manipulation, jailbreak, and contradiction-based queries. The attacks exploit the AI model's reasoning capabilities to generate unintentional responses that violate policies and create dangerous situations.

Current techniques are mainly based on surface filtering and keyword detection, which can't recognize more extensive logical inconsistencies or adversarial reasoning patterns. Also, there are almost no efficient mechanisms, even in the best current systems, for detecting attacks in real-time, adopting defense strategies that can be changed, or giving interpretability, which leads to a situation where it is not clear why the model failed to carry out the task as expected when it was faced with an adversary.

The development of a sophisticated intelligent system is expected to facilitate the identification and analysis of logical attacks for which defense mechanisms are required for the chatbot system. The system can offer the user transparent information on the decisions made by the system. The system would establish basic requirements for the enhancement of the reliability and security of future conversational AI systems.

Research Objective

- To develop an elaborate technology that detects and classifies logical attack patterns like prompt injections, contradiction attacks, and context modifications in chatbots.
- To design and develop a hybrid deep learning model that integrates transformer-based semantic analysis with graph-based logical reasoning for effective detection of adversarial prompts.
- To perform intelligent defense by implementing state-of-the-art defense schemes (dynamic prompt sanitization and adversarial prompt rewriting) that would make the chatbot resilient against attack in nature, which is mainly logic- and security-based attacks.

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 01 (2024)
http://www.veterinaria.org
Received: 10 December 2023, Revised: 25 December 2023, Accepted: 15 January 2024

- To apply explainable AI mechanisms (Attention visualization, SHAP, and LIME) that can provide a rationale for model decisions and gain Explainability to attack detection.

RESEARCH QUESTIONS

RQ1: How effective is a hybrid framework that combines transformer-based semantic analysis with graph-based logical reasoning in identifying and categorizing various types of logical attacks in chatbot systems?
RQ2: What is the effectiveness of advanced defense strategies such as dynamic prompt sanitization and adversarial prompt rewriting in enhancing the resilience and reliability of chatbot model responses against logical prompting attacks?

NOVELTY OF THE STUDY

The research introduces a new, complete framework for chatbot security improvements, which protects against logical attacks that existing studies have not investigated yet. The system uses a hybrid design that integrates transformer-based contextual processing with graph-based logical reasoning to detect advanced enemy patterns beyond the reach of standard methods that depend on basic filtering and separate semantic systems. The system introduces Contrastive Logical Feature Optimization (CLFO) as a method to enhance the distinction between typical user prompts and dangerous prompts. In addition to that, the study adds more active defense strategies, such as prompt sanitization, which is dynamic and adversarial prompt rewriting to neutralize the threat, not just identifying it. A crucial advancement also lies in the imitation of human reasoning through the use of Attention visualization, SHAP, and LIME to a multi-level explainability framework, which brings transparency to decision-making. Furthermore, the platform analysis context-sensitive multi-turn conversation and new robustness testing criteria thus equip it as a complete, interpretable, and defense-oriented platform for safeguarding the conversational AI.

RESEARCH METHODOLOGY

1. Dataset Description

The data used in this paper for training and testing the chatbot model is collected from various sources, including publicly available repositories such as Hugging Face and OpenAI datasets, and collections of datasets containing adversarial prompt datasets. In order to reduce the effect of insufficient logical attack data, a synthetic dataset is used. This is achieved by designing prompts based on different ways of logical attacks, such as prompt injection, jailbreaking, contradiction, and context. This data is preprocessed and categorized based on different types of logical attacks, making it feasible to perform multi-class classification and evaluate the model. Table 1 describes the dataset used in detail.

Parameter	Description
Dataset Type	Text (Chatbot Conversations)
Total Samples	301
Input Features	User Prompt, Context
Output Labels	Normal, Injection, Jailbreak, Contradiction, Context Attack
Number of Classes	5
Preprocessing	Cleaning, Normalization, Tokenization
Data Split	Train / Validation / Test
Context Handling	Multi-turn conversation

Table 1: Dataset Description

2. Technique used

Graph Neural Network (GNN) for Logical Reasoning : GNN represents conversation turns as nodes and logical dependencies as edges to analyze relationships, detect context manipulation, and identify logical inconsistencies in multi-turn conversations.

$$h_i^{l+1} = \sigma\Big(\sum_{j \in N(i)} W^l h_j^l\Big)$$

Contrastive Logical Feature Optimization (CLFO) : CLFO improves attack detection by maximizing the difference between normal and adversarial prompts through contrastive learning, enhancing feature separability.

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 01 (2024)
http://www.veterinaria.org
Received: 10 December 2023, Revised: 25 December 2023, Accepted: 15 January 2024

$$L = -\log \frac{exp(sim(x_i, x_j)/\tau)}{\sum_k exp(sim(x_i, x_k)/\tau)}$$

Explainable AI (XAI) Techniques (LIME, SHAP, Attention) : XAI provides transparency by explaining model decisions and identifying important prompt features responsible for logical attack detection.

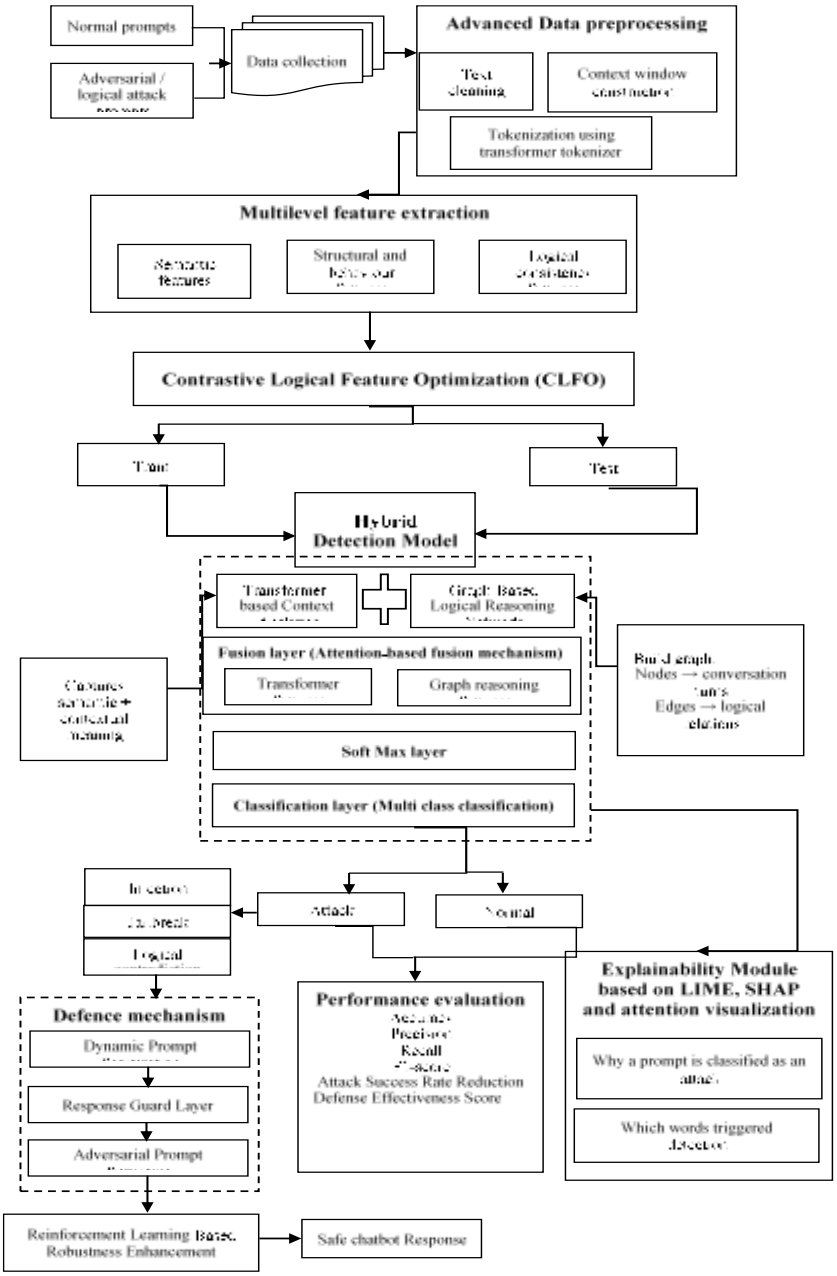
- **LIME:** Finds influential words/features locally.
- **SHAP:** Measures feature contribution using game theory.

3. **Attention:** Highlights important parts of input influencing predictions.
4. **Proposed methodology**

Figure 3 shows the proposed methodology of the study, followed by a detailed step-by-step description.

Figure 3: Proposed methodology

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 01 (2024)
http://www.veterinaria.org
Received: 10 December 2023, Revised: 25 December 2023, Accepted: 15 January 2024



REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 01 (2024)
http://www.veterinaria.org
Received: 10 December 2023, Revised: 25 December 2023, Accepted: 15 January 2024

Step 1: Dataset Preparation and Attack Taxonomy Design : First, the chatbot dataset is gathered and preprocessed by cleaning, normalization, and tokenization. A well-organized classification of logical attack types, such as prompt injections, jailbreaks, contradictions, and context manipulations, is laid out and employed to tag the dataset for supervised learning.

Step 2: Context Preservation and Encoding : To effectively model real-life conversation patterns, a concept of "multi-turn dialogue context" has been incorporated through a "context preservation" mechanism. Unlike previous approaches, where individual prompts were analyzed, a previous conversation history is encoded to improve accuracy in detecting attacks.

Step 3: Multi-Level Feature Extraction : Feature extraction applies to various levels because it uses transformer face embedding for semantic embedding and token-by-token entailment and contradiction score for assessing the logical relationship between premises and hypotheses, and it uses structural and behavioral pattern analysis for detecting malicious behavior.

Step 4: Contrastive Logical Feature Optimization (CLFO) : The contrastive learning optimization method ensures that these features learn discriminative feature representations at the feature level to ensure the normal and malicious prompts look more distinct from each other.

Step 5: Hybrid Detection Model Development : The system employs transformer-based models for contextual analysis and graph-based models for logical reasoning to develop a hybrid architecture design in which the system represents each turn of a conversation as a node and studies the relationship between each pair of nodes to analyze logical dependencies.
Step 6: Feature Fusion and Attack Classification : An attention-based fusion mechanism merges both transformer and graph model outputs effectively within the system. A multi-class classification model is employed to categorize inputs as either one of the many types of logical attacks or only as normal behavior.

Step 7: Defense Mechanism Implementation : Some advanced security measures have been introduced, such as prompt sanitization that removes malicious inputs and adversarial prompt rewriting that changes malicious queries to harmless ones before responding.

Step 8: Response Validation and Safety Layer : The system employs a safeguard mechanism to ensure all system outputs remain valid while it operates according to its safety standards, which turn off any unauthorized or hazardous system outputs.

Step 9: Explainability Integration : The approaches inspired by Explainable AI methods like attention visualization, SHAP, and LIME are designed to give the global view from the model while explaining the local detail of the feature importance.

Step 10: Performance Evaluation: The framework is evaluated on traditional metrics such as accuracy, precision, recall, and F1 score. In addition to evaluating the framework's robustness, other metrics such as Attack Success Rate Reduction (ASRR) and Defense Effectiveness Score (DES) are used.

EVALUATION PARAMETERS

The proposed framework is evaluated using multiple performance and security parameters to measure its effectiveness in detecting and defending against logical attacks in ChatGPT-based chatbots.
1. Accuracy : Accuracy measures the overall correctness of the model in classifying chatbot inputs as normal or malicious prompts. It represents the ratio of correctly classified samples to the total samples.
2. Precision : Precision indicates the percentage of detected attacks that are actually malicious. It helps reduce false alarms by measuring the correctness of attack predictions.
3. Recall : Recall measures the ability of the system to identify all actual attacks. A higher recall value indicates better detection capability against security threats.
4. F1-Score : F1-score provides a balanced evaluation by combining precision and recall. It is useful when handling imbalanced datasets containing fewer attack samples compared to normal inputs.
5. Attack Detection Capability : This parameter evaluates the ability of the proposed framework to identify different logical attack categories, including prompt injection, jailbreak, contradiction-based attacks, and context manipulation.
6. Attack Success Rate Reduction (ASRR) : ASRR measures the effectiveness of the defense mechanism by calculating the reduction in successful adversarial attacks after applying the proposed security framework.
7. Defense Effectiveness Score (DES) : DES evaluates the overall performance of defense strategies such as dynamic prompt sanitization, adversarial prompt rewriting, and response validation in protecting the chatbot.

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 01 (2024)
http://www.veterinaria.org
Received: 10 December 2023, Revised: 25 December 2023, Accepted: 15 January 2024

8. False Positive Rate (FPR) : FPR measures the percentage of normal user inputs incorrectly classified as malicious attacks. A lower FPR indicates improved reliability and user experience.
9. Explainability : Explainability evaluates the capability of the framework to provide understandable reasons behind attack classification decisions using Explainable AI techniques such as LIME, SHAP, and Attention Visualization.
10. Multi-turn Context Handling : This parameter measures the ability of the framework to analyze conversation history and detect attacks that occur across multiple interaction turns rather than only single prompts.

EXPLANATION OF EXPECTED EVALUATION PARAMETERS AND OUTCOMES

The proposed framework is evaluated using different performance and security parameters to measure its capability in detecting and defending against logical attacks in ChatGPT-based chatbots.

Metric	Expected Outcome
Accuracy	Measures overall correct classification of normal and attack prompts. Expected accuracy is >95%.
Precision	Indicates correct detection of malicious prompts with fewer false alarms.
Recall	Measures the ability to detect maximum possible attacks such as prompt injection and jailbreaks.
F1-Score	Provides balanced performance between precision and recall.
Attack Detection Capability	Evaluates the ability to identify different logical attacks like prompt injection, jailbreak, and context manipulation.
Attack Success Rate Reduction (ASRR)	Measures how effectively defense techniques reduce successful attacks.
Defense Effectiveness Score (DES)	Evaluates the overall capability of the defense layer to detect and prevent attacks.
False Positive Rate (FPR)	Measures incorrect blocking of normal user queries; lower value indicates better performance.
Explainability	Evaluates transparency using LIME, SHAP, and attention visualization techniques.
Multi-turn Context Handling	Measures the ability to detect attacks developed across multiple conversation turns.

REFERENCES

1. George, A. Shaji. "Beyond the Interface and How Conversational AI Is Reshaping the Future of Human-Computer Interaction." *Partners Universal International Innovation Journal* 3, no. 4 1-13.
2. Pooja, S., G. Gokul, K. Linkesh Mani, A. S. Raj Kumar, and R. Amutha Bharathi. "Context-Aware AI Chatbot Using Transformer-Based Models for Intelligent User Interactions."
3. Hasal, Martin, Jana Nowaková, Khalifa Ahmed Saghair, Hussam Abdulla, Václav Snášel, and Lidia Ogiela. "Chatbots: Security, privacy, data protection, and social aspects." Concurrency and Computation: Practice and Experience 33, no. 19 : e6426.
4. Gulyamov, Saidakhror, Said Gulyamov, Andrey Rodionov, Rustam Khursanov, Kambariddin Mekhmonov, Djakhongir Babaev, and Akmaljon Rakhimjonov. "Prompt Injection Attacks in Large Language Models and AI Agent Systems: A Comprehensive Review of Vulnerabilities, Attack Vectors, and Defense Mechanisms." *Information* 17, no. 1: 54.
5. Aslan, Ömer, Semih Serkant Aktuğ, Merve Ozkan-Okay, Abdullah Asim Yilmaz, and Erdal Akin. "A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions." *Electronics* 12, no. 6 : 1333.
6. Quffa, Abdallah, and Samy S. Abu-Naser. "A Rule-Based Expert System for Cybersecurity Threat Detection: Evolution, Applications, and the Hybrid AI Paradigm."
7. https://www.getastra.com/blog/ai-security/prompt-injection-attacks/
8. Hassija, Vikas, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. "Interpreting black-box models: a review on explainable artificial intelligence." *Cognitive Computation* 16, no. 1: 45-74.
9. Polemi, Nineta, Isabel Praça, Kitty Kioskli, and Adrien Bécue. "Challenges and efforts in managing AI trustworthiness risks: a state of knowledge." *Frontiers in Big Data* 7: 1381163.
10. KAMATA, Mitsunobu. "Security Design and Evaluation of a Citi-zen-Oriented Generative AI Chatbot Using RAG (Retrieval-Augmented Generation)." *International Journal of ICT Application Research* 3, no. 1: 1-8.

11. Anghel, Catalin, Marian Viorel Craciun, Adina Cocu, Andreea Alexandra Anghel, Antonio Stefan Balau, Adrian Istrate, and Aurelian-Dumitrache Anghele. "EvalHack: Answer-Side Prompt Injection for Probing LLM Exam-Grading Panel Stability." *Information* 17, no. 3: 297.
12. Zhang, Yihao, Zeming Wei, Xiaokun Luan, Chengcan Wu, Zhixin Zhang, Jiangrong Wu, Haolin Wu, Huanran Chen, Jun Sun, and Meng Sun. "ClawWorm: Self-Propagating Attacks Across LLM Agent Ecosystems." *arXiv preprint arXiv:2603.15727* .
13. ŞAŞAL, A., and AB CAN. "Prompt Injection Attacks on Large Language Models: Multi-Model Security Analysis with Categorized Attack Types.".
14. Emekci, Hakan, and Gülsüm Budakoglu. "Securing With Dual-LLM Architecture: ChatTEDU an Open Access Chatbot's Defense." *IEEE Access* .
15. Yi, Jingwei, Yueqi Xie, Bin Zhu, Emre Kiciman, Guangzhong Sun, Xing Xie, and Fangzhao Wu. "Benchmarking and defending against indirect prompt injection attacks on large language models." In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Vol. 1*, pp. 1809-1820.
16. Azooz, Hasan Jameel. "Comprehensive, context-aware, multi-layered security framework for mitigating prompt injection attacks in large language models."
17. Liu, Xiaogeng, Zhiyuan Yu, Yizhe Zhang, Ning Zhang, and Chaowei Xiao. "Automatic and universal prompt injection attacks against large language models." *arXiv preprint arXiv:2403.04957* .
18. Derner, Erik, Kristina Batistič, Jan Zahálka, and Robert Babuška. "A security risk taxonomy for prompt-based interaction with large language models." *IEEE Access* 12: 126176-126187.
19. Jalali, Nasir Ahmad, and Chen Hongsong. "Comprehensive framework for implementing blockchain-enabled federated learning and full homomorphic encryption for chatbot security system." *Cluster Computing* 27, no. 8 : 10859-10882.
20. Suárez-Varela, José, Paul Almasan, Miquel Ferriol-Galmés, Krzysztof Rusek, Fabien Geyer, Xiangle Cheng, Xiang Shi et al. "Graph neural networks for communication networks: Context, use cases and opportunities." *IEEE Network* 37, no. 3: 146-153.
21. Sharma, Amit, Ashutosh Sharma, Alexey Tselykh, Alexander Bozhenyuk, and Byung-Gyu Kim. "Image and video analysis using graph neural network for Internet of Medical Things and computer vision applications." *CAAI Transactions on Intelligence Te*.
22. Ju, Wei, Yifan Wang, Yifang Qin, Zhengyang Mao, Zhiping Xiao, Junyu Luo, Junwei Yang et al. "Towards graph contrastive learning: A survey and beyond." *arXiv preprint arXiv:2405.11868* .
23. Rane, Nitin, Saurabh Choudhary, and Jayesh Rane. "Explainable Artificial Intelligence (XAI) approaches for transparency and accountability in financial decision-making." *Available at SSRN 4640316*.
24. Zhang, Changsheng, and Linjun Liu. "Machine learning prediction model for medical environment comfort based on SHAP and LIME interpretability analysis." *Scientific Reports* 15, no. 1 : 39269.
25. Liu, Ying, Yating Fu, Yadong Peng, and Jie Ming. "Clinical decision support tool for breast cancer recurrence prediction using SHAP value in cooperative game theory." *Heliyon* 10, no. 2 .
26. Soydaner, Derya. "Attention mechanism in neural networks: where it comes and where it goes." *Neural Computing and Applications* 34, no. 16: 13371-13385.