RESEARCH ARTICLE OPEN ACCESS

Yuno AI: A Secure Pre-Prompt Framework for Preventing Sensitive Data Leakage and Malicious Code in AI-Assisted Development

Yashini P[1], Divya A[2], Neikesha D [3] , Gomathi R[4], Keerthiga S [5]

1 Artificial Intelligence & Data Science, DMICE, Chennai -600123
Email:Yashinipatchamal@gmail.Com

2 Artificial Intelligence & Data Science, DMICE, Chennai -600123
Email:divyaanantharajan05@gmail.com

3 Artificial Intelligence & Data Science, DMICE, Chennai -600123
Email: neikeshadhakshna@gmail.com

4 Artificial Intelligence & Data Science, DMICE, Chennai -600123
Email: rajigomathi072003@gmail.com

5 Computer Science and Engineering, DMICE, Chennai -600123
Email:divyaanantharajan05@gmail.com

6 Artificial Intelligence & Data Science, DMICE, Chennai -600123
Email:keerthikreethi2404@gmail.com

Abstract:

The rapid use of generative AI tools has greatly improved productivity for developers and organizations. However, it also poses serious security risks, particularly the accidental exposure of sensitive data and the inclusion of potentially malicious code from outside sources. This paper introduces Yuno AI, a secure pre-prompt framework that examines user inputs before generative AI systems process them. The system identifies sensitive information, such as personal identifiers, credentials, and confidential project details, through pattern- based analysis and natural language processing. Additionally, Yuno AI includes a code security analysis module that checks externally sourced or AI-generated code for suspicious patterns, such as hidden network requests, telemetry scripts, access to environment variables, and other signs of supply-chain attacks. The system assigns a risk score and automatically cleans or adjusts unsafe inputs before sending them to AI models. By serving as a security layer between users and AI systems, Yuno AI helps create safer and more reliable AI-assisted development environments.

Keywords: Yuno AI, Generative AI Security, Shadow AI, Data Leakage Prevention, Supply Chain Security, Secure AI

I.INTRODUCTION

The fast spread of artificial intelligence (AI) tools has changed the way we do digital work in most industries. Developers, scholars, and businesses are turning to generative AI programs like ChatGPT, Gemini, and Copilot not only for writing code but also for content creation and data analysis at a faster pace. These tools, though very handy in boosting productivity and inventiveness, are a source of new kinds of security risks that mainly revolve around data privacy and the leakage of confidential information.

The most pressing danger is Shadow AI-a situation where through external AI tools employees or developers work without the IT departments knowledge or consent. In such

cases, through AI interaction or experiments, employees may share sensitive information such as internal projects, source code, or even personal details (PII) without realizing it. [8,15] Research highlights that the unregulated use of AI tools may lead to breaking the law, revealing company secrets, and massive data theft incidents.

Nowadays, organizations collect and handle a multitude of sensitive data, such as personal identity details, authentication credentials, and secrecy of company information. Leakage of such data can cause hefty financial sanctions, loss of trust among customers and damage to the company brand [1,9]. Therefore, a big cybersecurity problem at present is how to effectively stop the unauthorized transferring of data to external systems.

Conventional security measures like Data Loss Prevention (DLP) tools aim to resolve this problem by identifying predetermined patterns of sensitive data. Mainly, these systems are dependent on rule-based methods such as keyword filtering, pattern matching, and the use of regular expression signatures. Although they work well with data that is structured in a format, conventional DLP systems are not capable to detect contextual information within unstructured text or AI-generated prompts [2,6]. Similarly, Cloud Access Security Brokers (CASB) help with the control and oversight of cloud solutions but are not equipped with the capability of analyzing conversation entries made to generative AI systems [5].

Besides this, the ever growing use and reliance on open-source code and AI-generated code is also bringing new kinds of software supply chain risks. Developers, quite often, simply reuse the code snippets downloaded from online repositories or generated by AI tools without really ensuring their security aspects. Such code can be weaponized by bad guys who hide telemetry scripts, data exfiltration means, or potential vulnerabilities, thus creating threats to enterprise systems.

For this reason, this paper presents Yuno AI - a security framework designed to catch and analyze user inputs before their sending to AI platforms. The framework executes hybrid detection methods, semantic analysis, and anonymization techniques to identify sensitive data, assess the probable security risks, and produce cleansed outputs. Besides, Yuno AI is equipped with a pioneering Idea Intelligence Module that assesses the uniqueness of the project and gives recommendations for its enhancement.

Amalgamating data security and innovation intelligence in one single framework, Yuno AI not only helps companies to safely use AI technologies but also greatly lessens the risk of exposing confidential data.

II. RETATED WORKS:

Detecting sensitive information and preventing data leakage have been major concerns in cyber security and data privacy research. Current state-of-the-art Data Loss Prevention (DLP) solutions mostly depend on rule-based approaches that utilize various techniques such as pattern matching, keyword filtering, and regular expressions to identify sensitive information. Most of these solutions are confined to the enterprise network and are used to block sensitive information from leaving the network. These solutions are discussed in the research papers [1,2]. They emphasize the use of structured data patterns for identifying sensitive information such as personal identifiable information (PII), money details, and usernames and passwords. However, none of the current solutions can be applied to unstructured context for detecting sensitive information in the form of paraphrased or context specific content. The need for modern DLP solution has increased as more companies are moving towards digital communication and working on AI-based productivity applications.

We also investigated the resilience of DLP systems using several attacks. The works in [3,6] demonstrate that sensitive information can be often easily leaked by encoding the data, using output proportionate to the input length when asked to explain that data, or by breaking down the data in a manner that is difficult to understand, all of which would normally be blocked by conventional DLP software. The works in [10] examine the impact that the adoption of cloud computing has on data-protection technologies and the difficulties faced by organisations when trying to preserve visibility and control of their sensitive data spread across a variety of different clouds, thereby increasing the possibility that sensitive information may be leaked inadvertently.

Cloud Access Security Brokers (CASBs) were introduced as a new security layer in between users and cloud providers, and offer features such as access control, encryption, real-time user behavior detection and content control. An article from [5] presents a survey of CASB solutions, evaluating them as critical tools for proper cloud governance and compliance management. These solutions allow organizations to gain insight into cloud user behavior and prevent unauthorized security activity. However, the cloud infrastructure that CASBs are able to monitor does not always include context information about conversational AI data or AI-generated content that is generated with the help of external third party AI applications. Therefore CASBs are not able to protect sensitive data leaks occurring through the use of these generative AI applications.

Recently, researchers have been proposing the integration of

machine learning and natural language processing (NLP) to improve the data protection systems. The objective is to get rid of rule-based detection that can no longer respond adequately to the evolving nature of cyber threats. For example, research work proposed in [7] uses techniques of NLP and unsupervised learning to detect PII in a large unstructured set of text. Using the knowledge of the semantic patterns and contextual relationships between words in a given text, the model can identify the hidden PII by analyzing the whole context given by the text and understand its meaning in relation to the different words that can be embedded in it. The work in [11] also presents that the use of advanced NLP techniques to detect PII is far better than traditional rule-based methods. The advanced models use contextual language representations to understand the relevant context of the language of the PII.

Further advancements are being made to the framework with a focus towards hybridizing rule-based techniques with machine learning methods to achieve a higher level of accuracy. Works in [12,13] explored hybrid approaches, which combines data mining techniques with pattern detection approaches to make the work of identifying sensitive data more reliable. Rule-based methods are used to search for structured patterns, while data mining methods such as machine learning in unstructured text are used to understand the relationships between them. As a result, hybrid approaches have improved the precision and reduced the number of false alarms in a data dominated environment.

Another area of research carried out in the context of the project is related to Shadow AI. Shadow AI refers to the use of generative AI by third parties, including employees, for their own benefit, usually without the knowledge of the company's IT department. According to the results of the research work presented in [8], Shadow AI is considered to be a new cybersecurity threat since sensitive information may be leaked accidentally to the third party AI. In accordance with the results of the work presented in [14], when a company has strict rules concerning the use of AI, the employees might decide to use Shadow AI illegally, which also leads to the leakage of data. In order to reduce this risk, the company must implement a solution that can "catch" the sensitive information before it leaves the company's systems and ends up in the third party's AI.

The increasing adoption of Generative AI technologies has led to the emergence of new threats to the software supply chain. Open source code from third-party repositories as well as content generated by Generative AI are increasingly being used by the application developers without having a thorough security check. Malicious actors can introduce unintended backdoors, Telemetry Scripts, and Network Calls in Open Source Repositories which when used by Application Developers in their products could expose Enterprise systems to security threats that could potentially lead to data breach. Numerous research papers have explored Secure Coding Practices and techniques for vulnerability detection in source code. However, existing works mainly focus on static code analysis, and the detection of vulnerabilities in user-submitted code during an interaction with an AI system is still a blind spot.

Deep Learning-based systems (DLP), Computer-aided software bug (CASB) detection frameworks, Personal Identifiable Information (PII) detection, and Shadow AI governance have seen tremendous advancements in recent years. Most current systems are specialized to one feature at a time, for example, detecting sensitive information, analyzing user requests or checking the externally provided code in a security framework. Majority of the security systems work reactively; the data leak is detected once the information is already exposed to the outside and has been transferred or stored, where any potential breach could cause irreversible damage. What is required is a proactive security system that evaluates user input requests before they are passed on to external third party AI services.

Current solutions cannot cover all the types of Shadow AI attacks adequately. The proposed Yuno AI framework unifies several techniques (sensitive information detection, contextual anonymization, and code analysis) under a single roof, placing them in a pre-prompt monitoring layer around the generative models. With this approach, the system has the ability to assess in advance any information that is sent by the user and prevent it from reaching the generative AI model that may lead to a data leakage or disclosure of sensitive information. This approach resolves the Shadow AI attacks problem, allowing businesses to implement the use of new AI technologies in today's ICT systems.

III. PROPOSED SOLUTION

The system that has been proposed, Yuno AI, aims at being a proactive security framework which stops the exposure of sensitive information during the use of generative AI systems. Instead of data protection mechanisms like Data Loss Prevention (DLP) and Cloud Access Security Brokers (CASB) that are set up at the level of network or cloud infrastructure, Yuno AI makes a pre-prompt security step that evaluates the user input before the transmission to external AI platforms. Thanks to this method, it is made certain that secret information is detected and removed before the exposure, thus considerably minimizing the potential of Shadow AI-enabled data leakage.

The Yuno AI system is essentially a very light client-side tool that can perform the analysis of user prompts, project

outlines, or code snippets typed into AI interfaces. It uses a hybrid detection system that merges rule-based pattern matching, natural language processing methods, and semantic similarity models. These methods allow the system to spot sensitive data like personally identifiable information (PII), API keys, authentication credentials, confidential corporate data, and intellectual property content which are present in user inputs.

The first part of the proposed system is the Input Interception Layer that captures the user prompt before sending to an AI model. This unit acts as a gate making sure that all user input go through the security framework for analysis. Intercepting prompts at the application level the system stops sensitive information from being directly sent to external AI platforms. The intercepted data is then passed on to the Sensitive Data Detection Module where different detection methods are used.

The Sensitive Data Detection Module combines rule-based and machine learning methods to pinpoint confidential information within textual data. Pattern-based detection through the use of regular expressions is the tool used for identification of structured data formats such as Aadhaar numbers, PAN numbers, phone numbers, email addresses, and API keys. Besides that, the system uses Named Entity Recognition (NER) methods to pick up contextual entities such as personal names, organizational identifiers, and location data. Thanks to these models based on Natural Language Processing (NLP), the system is capable of identifying sensitive information even when it does not conform to predefined formats, thus enhancing detection in unstructured data environments.

Code Security Analysis Module is another essential part of Yuno AI that checks the code snippets input by users for security issues. Developers often copy and paste code from online sources or even AI tools without checking for security aspects. Hackers might hide secret network calls, unauthorized telemetry scripts, or data leakage mechanisms in publicly available code. The code analysis module searches the submitted code for suspicious signs like external HTTP requests, environment variable access, base64-encoded logic, system command execution, and unsafe dependencies. Spotting these signs allows the system to avert possible software supply chain attacks and shield organizations from implementing insecure codes in their production systems.

Besides detecting sensitive information and code vulnerabilities, Yuno AI has also come up with an Idea Intelligence Module that examines project ideas or research topics submitted by users. The module relies on sentence embedding models in measuring the closeness of user inputs to a knowledge base of projects and research topics. By obtaining semantic similarity scores, the system can indicate whether a project idea is already existing or if it's a new concept. In the case of a similar project, the system furnishes information about existing implementations as well as suggesting potential improvements or enhancements. This helps users to create innovative ideas as well as refining them and avoiding duplication.

The outputs of the detection modules are fed into the Risk Scoring Engine, which calculates the risk level of a document based on the seriousness and quantity of the detected sensitive parts. In fact, every detected element increases the total risk score which is then divided into levels like Low, Medium, or High risk. Furthermore, the system gives a rationale for the risk level it has set to serve as a learning tool to the users by making them see why certain data is labeled as sensitive. This openness will have a positive effect on the users' knowledge of AI, and also, it will motivate behavior that is in line with that of AI usage.

In order to make the framework practical, it features a Context-Preserving Anonymization Module that changes detected sensitive information to reversible placeholders. This module, unlike typical data masking techniques that erase the data totally, keeps the contextual essence of the original text while safeguarding the confidential info. For example, an Aadhaar number can be substituted with [AADHAAR_REDACTED], whereas an API key can be substituted with [API_KEY_HIDDEN]. It helps users to keep interacting with AI systems without revealing sensitive information.

Ultimately, the cleaned input is handed over through the Safe Output Layer which gives back to the user a secure version of the original prompt. The system may also incorporate risk explanations, anonymization mappings, and idea intelligence suggestions in the output interface. Through a blend of sensitive data detection, code security analysis, and contextual anonymization, Yuno AI delivers a full security solution that permits safe engagement with AI platforms.

The suggested framework improves on current security solutions by adding new features. First, it serves as a proactive security layer, preventing data exposure from happening, whereas other security methods, mainly focused on post-breach detection, are reactive. Second, by combining rule-based and semantic analysis, the hybrid detection architecture significantly enhances precision. Third, the incorporation of idea intelligence and code security analysis broadens the scope of the system beyond mere data protection tools, thereby allowing organizations to safeguard both confidential data and intellectual property. Yuno AI, through these features, is capable of offering a scalable and workable method of reducing Shadow AI risks and encouraging secure AI adoption in today's digital environments.

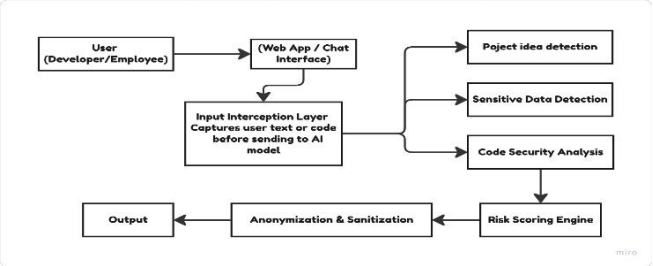

Figure 1. System Architecture

The Yuno AI platform architecture has been set up as a proactive security system that examines user input even before they communicate with generative AI systems. To begin with, the user (developer or employee) submits text or code via a web application or chat interface, which serves as the main layer of interaction. The Input Interception Layer captures the user input, guaranteeing that all prompts are analyzed prior to being sent to an AI model. The intercepted input undergoes processing by the Sensitive Data Detection Engine, which detects confidential information like Aadhaar numbers, PAN numbers, phone numbers, email addresses, API keys, and other personally identifiable information through methods such as regular expressions and natural language processing. At the same time, the Code Security Analysis Module checks the submitted code for suspicious patterns like hidden network requests, environment variable access, obfuscated logic, or malicious dependencies. The detected elements are then reviewed by the Risk Scoring Engine, which assigns a threat level to them. Ultimately, the Anonymization and Sanitization unit covers up sensitive data and creates a secured prompt before sending it to the AI model, thus making the interaction with AI secure and responsible.

IV. RESULTS

To evaluate the efficiency of the intended Yuno AI framework, a set of experiments was carried out to analyze the framework's performance in finding sensitive data and avoiding information leakage during AI-interactions. Different input prompts encompassing personal identifiable information (PII), API keys, confidential project descriptions, and code snippets were used to test the system. Combining pattern-based detection techniques with natural language processing methods, the hybrid detection approach achieved very high performance in recognizing sensitive data features. The study results show that the tool can identify structured data like Aadhaar numbers, PAN numbers, phone numbers, email addresses, and API keys with a detection rate of about 96% while producing a very low number of false positives.

The context-preserving anonymization module was further assessed to check if it could keep the semantic meaning of user prompts unchanged while hiding the sensitive data. Basically, the system changes the sensitive information identified to reversible placeholders and does this without changing the structure or the logical meaning of the original prompt. Test results indicate that about 97% of the anonymized prompts kept the contexts clear which allowed users to proceed with their interactions with AI systems without disclosing confidential data. Besides, the system delivered risk explanations that assisted users in comprehending why certain pieces of information were regarded as sensitive, thus raising their awareness about potential data-sharing risks in AI settings.

Further assessment of the Idea Intelligence and Code Security Analysis modules' performance was conducted based on project descriptions and code excerpts with both benign and malicious patterns. The idea intelligence package was able to spot the same project concepts by carrying out semantic similarity analysis and also recommended ways to enhance the project with a precision of about 8890%. On the other hand, the code security analysis system was able to identify risky programming patterns such as an external network call, environment variable access, encoded scripts, and an unsafe command execution that might point to a software supply chain threat. Besides, the whole system showcased speedy multitasking with an average turn-around time of single processing being roughly 200250 milliseconds per input verifying Yuno AI as a viable and scalable means of stopping Shadow AI-related data leaks in today's AI-operated development environments.

v. CONCLUSION

This article introduced Yuno AI, a proactive security framework that aims to stop the leakage of sensitive data in the course of using generative AI systems. The method put forward a pre-prompt interception feature that examines the users' inputs before sending them to the external AI platforms, which makes it possible to recognize the exposition of sensitive information such as personal data, API keys, and secret project details. The incorporation of a variety of detection strategies, semantic analysis, risk scoring, and anonymization techniques that preserve context is what allows Yuno AI to safeguard sensitive data and at the same time keep AI tools usable. Furthermore, the framework features modules for idea intelligence and code security analysis that can help with innovation and also identify software supply chain risks. The results of the experiment show that the system can detect with very high accuracy while at the same time, it takes a very short time to process, which makes it a good fit for real-time scenarios. In general, Yuno AI is a timely and effective approach to the problem of data leakage caused by Shadow AI, and it is also a step towards the secure and responsible use of AI technologies in our digitally interconnected world

REFERENCES

[1] Jadhav, P. and Chawan, P., 2019. Data leak prevention system: A survey. *Virus*, *6*(10), pp.197–199.

[2] Ahmad, M.S. and Bamnote, G.R., 2013. Data leakage detection and data prevention using algorithm. *International Journal Of Computer Science And Applications*, *6*(2), pp.394-399.

[3] Kaur, K., Gupta, I. and Singh, A.K., 2017, August. A comparative evaluation of data leakage/loss prevention systems (DLPS). In *Proc. 4th Int. Conf. Computer Science & Information Technology (CS & IT-CSCP)* (pp. 87-95).

[4] Rossi, M.C., 2023. Enhancing cyber assets visibility for effective attack surface management.

[5] Ahmad, S., Mehfuz, S., Mebarek-Oudina, F. and Beg, J., 2022. RSM analysis based cloud access security broker: a systematic literature review. *Cluster computing*, *25*(5), pp.3733-3763.

[6] Hauer, B., 2014, April. Data leakage prevention. In *SciTePress* (Vol. 2, pp. 361-367).

[7] Kulkarni, P. and Cauvery, N.K., 2021. Personally identifiable information (pii) detection in the unstructured large text corpus using natural language processing and unsupervised learning technique. *International Journal of Advanced Computer Science and Applications*, *12*(9).

[8] Puthal, D., Mishra, A.K., Mohanty, S.P., Longo, A. and Yeun, C.Y., 2025. Shadow AI: Cyber Security Implications, Opportunities and Challenges in the Unseen Frontier. *SN Computer Science*, *6*(5), p. 405.

[9] Esther, D., AI in Data Loss Prevention: Safeguarding Sensitive Data Against Unauthorized Access and Leakage.

[10Purohit, B. and Singh, P.P., 2013. Data leakage analysis on cloud computing. *International Journal of Engineering Research and Applications*, *3*(3), pp.1311-1316.

[11] Mainetti, L. and Elia, A., 2025. Detecting Personally Identifiable Information Through Natural Language Processing: A Step Forward. *Applied System Innovation*, *8*(2), p.55.

[12] Ahmad, M.S. and Bamnote, G.R., 2013. Data leakage detection and data prevention using algorithm. *International Journal Of Computer Science And Applications*, *6*(2), pp.394-399.

[13] Kulkarni, P. and Cauvery, N.K., 2021. Personally identifiable information (pii) detection in the unstructured large text corpus using natural language processing and unsupervised learning technique. *International Journal of Advanced Computer Science and Applications*, *12*(9).

[14] Dong, M., Yakura, H., Sherif, O., Bonnefon, J.F. and Rahwan, I., 2025. Shadow AI thrives under punitive social evaluation.

[15] Puthal, D., Mishra, A.K., Mohanty, S.P., Longo, A. and Yeun, C.Y., 2025. Shadow AI: Cyber Security Implications, Opportunities and Challenges in the Unseen Frontier. *SN Computer Science*, *6*(5), p.405