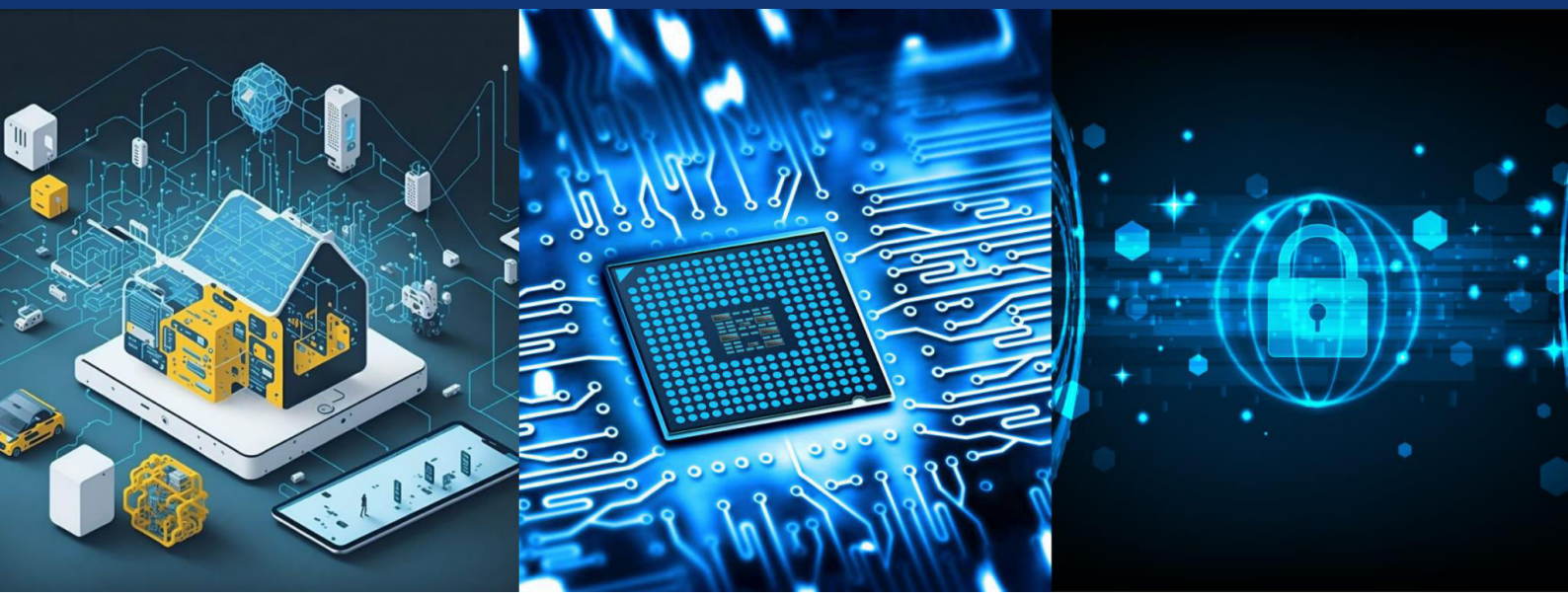
IJIRCCE

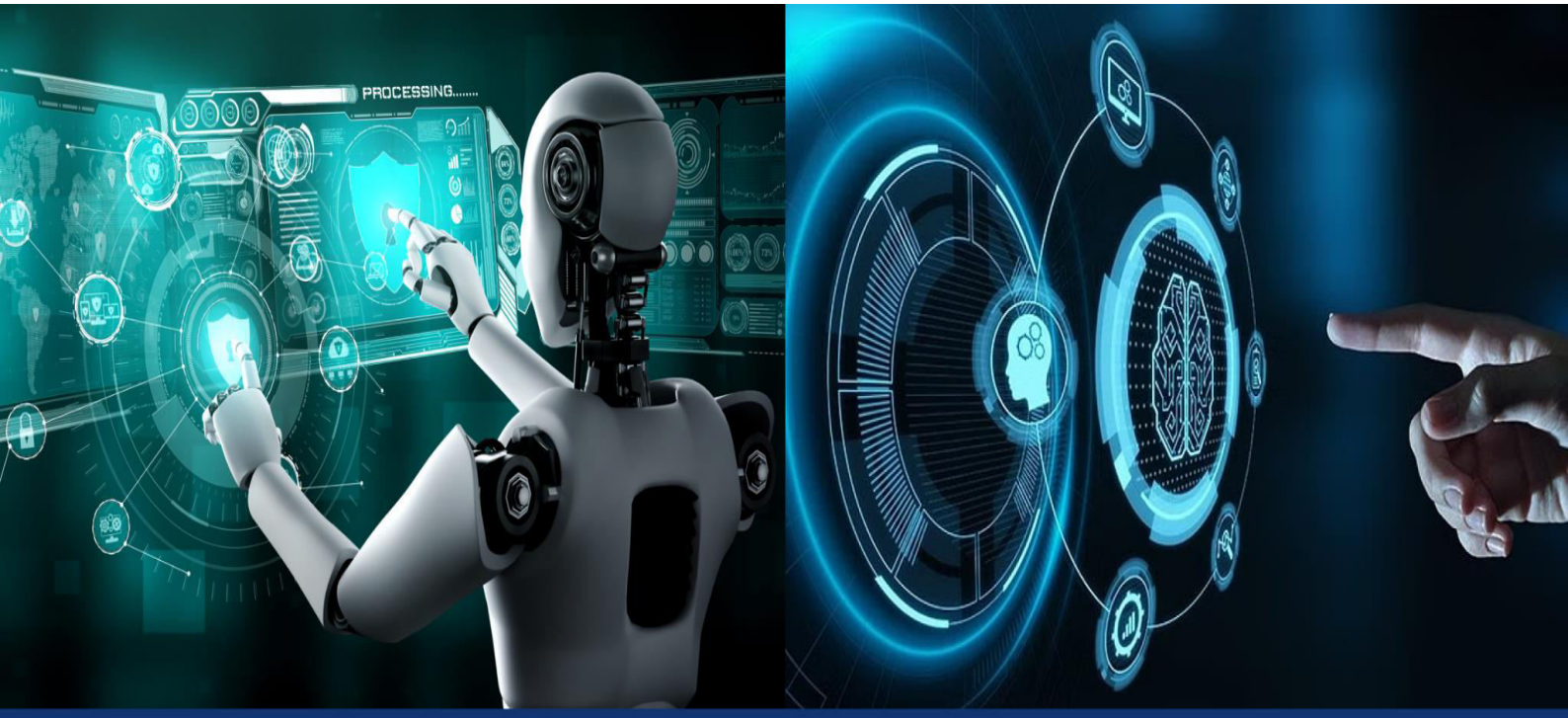
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 5, May 2025

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI: 10.15680/IJIRCCE.2025.1305021

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Natural Language Processing and Cryptography Approaches to Combat Document Theft

Dr.B.C.Brightlin, Dhanush S, Kamaraj P, Sathrapathimaaraj M

Professor, Department of Cyber Security, Muthayammal Engineering College, Rasipuram, Tamil Nadu, India

Student, Department of Cyber Security, Muthayammal Engineering College, Rasipuram, Tamil Nadu, India

ABSTRACT: Document theft and unauthorized access are major concerns in today's digital world. Traditional security methods like access control and encryption often fall short. This project proposes a dual-layered solution using Natural Language Processing (NLP) and Cryptography. NLP analyzes and classifies content, creating unique document fingerprints and detecting sensitive data or unauthorized changes. Simultaneously, cryptographic techniques such as Blowfish encryption and digital signatures secure documents during storage and transmission. This hybrid model enhances content monitoring and access protection, offering a proactive, adaptive system. It aims to improve document security across industries by preventing theft, unauthorized access, and tampering with sensitive information.

KEYWORDS: Natural Language Processing, LLM, Cryptography, Document Theft, Cyber Security

I. INTRODUCTION

In the digital era, the theft and unauthorized access to sensitive documents pose significant threats to individuals, businesses, and government entities alike. Whether it's financial data, intellectual property, or personal records, the loss or tampering of critical documents can have severe consequences, including financial loss, reputational damage, and legal implications. As digital document sharing and storage continue to increase, traditional methods of document security, such as passwords, access control lists, and basic encryption, are becoming inadequate to address the growing sophistication of cyber threats and the dynamic nature of document management. Document security, or document access security, is the process of safeguarding documents and files from unwanted access or theft. It also refers to procedures carried out to prevent data from being manipulated or reproduced wrongfully. Examples of document security policies include encrypting document, controlling access to confidential information, and monitoring the use of document and files. In addition, documents can be secured by restricting usage to prevent document damage, using secure computer systems and networks, and proper removal of unattended documents and records. With the significant increase in data being produced by businesses each year, a clear, well-defined document security plan is required to secure critical business information. To secure sensitive data, document security is crucial for businesses of any size.

Implementing document security can minimize or prevent any data breaches or misuse. Organizations have to ensure that only authorized individuals have access to the documents, which is the aspect of document security that businesses across the world often struggle with. Document security resolves this by enabling enterprises to easily monitor access and authorization. It secures the maintenance of files in their entire lifecycle of storage, backup, processing, and delivery through features like encryption, watermarking, and data rights management. Additionally, document security can lower the risk of data corruption. Corruption can occur when users have damaged sections in their hard drives or storage media that might contain viruses or malware. Document damage can also be caused maliciously by hackers that install dangerous malware in order to change or destroy information, in an attempt to conduct ransomware.

Cryptography is the practice and study of techniques for securing communication and protecting data from unauthorized access or tampering. In the context of document security, cryptography plays a pivotal role in ensuring the confidentiality, integrity, and authenticity of digital documents. By transforming data into an unreadable format (encryption), cryptography prevents unauthorized parties from accessing or altering sensitive information. Cryptography encompasses a wide range of techniques and algorithms, each suited to different security needs. Document security is an important part of the overall security effort for businesses. Thankfully, there are tools and technologies that can help organizations to manage and protect their data. Protecting numerous sensitive information types should be the top priority because failing to do so would threaten the effectiveness of the whole enterprise.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Document security uses digital rights management as a way to enact document encryption and watermarking. This ensures that documents are highly secured and can only be seen by those with proper authorization. Digital Rights Management (DRM) is the process of managing, controlling, and securing data from unauthorized users. The purpose of DRM solutions is to protect the author's rights and restrict unauthorized distribution or modification. By requesting authorization (automatically or manually) prior to getting access, DRM will increase the security of assets through persistent file protection. This will guarantee that only those with the necessary access rights can view the information. Document/file encryption can be used as a way to enforce DRM. When team members use on-premises or cloud documents for data transfer, or document share, encryption can prevent documents from being viewed by unintended individuals. This way, if the document gets into the wrong hands, the unintended user will not be able to view it even if they may have access to the folder in which the document is stored.

II. EXISTING SYSTEM

Deceptive repositories are an innovative cybersecurity defense mechanism specifically designed to protect intellectual property (IP) from theft. These repositories utilize deception techniques to mislead and engage potential attackers, diverting them away from valuable assets. Unlike traditional cybersecurity defenses, which primarily rely on reactive strategies like firewalls or intrusion detection systems, deceptive repositories adopt a proactive approach. Deceptive repositories leverage active defense mechanisms by creating and placing decoy assets—such as fake documents, credentials, or databases—that mimic genuine resources.

These assets are strategically crafted to attract attackers. When attackers interact with these decoys, it triggers alerts for cybersecurity teams and enables them to monitor and analyze malicious behavior in a controlled environment. By presenting attackers with decoy data that appears authentic and contextually relevant, they effectively disrupt malicious activities while gathering intelligence about the threat actors. WE-Forge is a tool that uses word embeddings—vector representations of words in a multi-dimensional space—to generate fake documents. Word embeddings capture the semantic and syntactic relationships between words, allowing WE-Forge to produce content that appears contextually accurate and relevant to the targeted domain. This ensures that the decoy documents are convincing enough to engage attackers while containing no sensitive or actual intellectual property. This innovative strategy not only strengthens IP security but also enhances an organization's overall resilience against sophisticated attackers.

Limitations:

Complex or highly specialized fields may require intricate knowledge that word embeddings alone might not fully capture. If the training corpus is not representative of the domain or is outdated, the generated documents may lack the realism required to convincingly deceive attackers. There's a risk that legitimate users within the organization might inadvertently interact with or use fake documents, leading to confusion, operational errors. Careful maintenance and balance between realism and simplicity in the fake documents

III. PROPOSED SYSTEM

The proposed system introduces a comprehensive approach to document security by integrating Natural Language Processing (NLP) and cryptography to effectively combat document theft. This hybrid system establishes a robust defense mechanism that not only protects sensitive information but also actively monitors and analyzes user interactions with documents to detect potential threats. Leveraging NLP algorithms, the system will analyze document contents and identify keywords using the TF-IDF (Term Frequency-Inverse Document Frequency) mechanism. This analysis aids in understanding the context and relevance of document content, enhancing the ability to safeguard critical information. Additionally, the system incorporates a deception layer by creating fake documents that are stored in alternative repositories. These decoy documents mislead unauthorized users, diverting them away from genuine assets. To further strengthen security, the system employs Blowfish encryption, a robust encryption protocol, to ensure that only authorized users can access and decrypt sensitive information. By combining advanced content analysis, deception strategies, and strong cryptographic measures, the proposed system provides a proactive and multi-layered defense against document theft.

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI: 10.15680/IJIRCCE.2025.1305021

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. EXPECTED MERITS

By combining NLP, cryptography, and deception techniques, the system provides a comprehensive defense mechanism. This multi-faceted approach ensures that even if one layer is compromised (e.g., encryption), other layers (e.g., document monitoring and fake documents) still offer protection. The creation of fake documents provides a proactive defense mechanism by luring attackers into engaging with false information. Fake documents stored in alternative repositories add an extra layer of confusion for attackers, increasing the time and effort they spend distinguishing between real and fake data. By securing documents with Blowfish encryption, only authorized users with the correct decryption keys can access the contents, significantly reducing the risk of data leakage. This makes it a cost-effective solution for organizations looking to enhance document security.

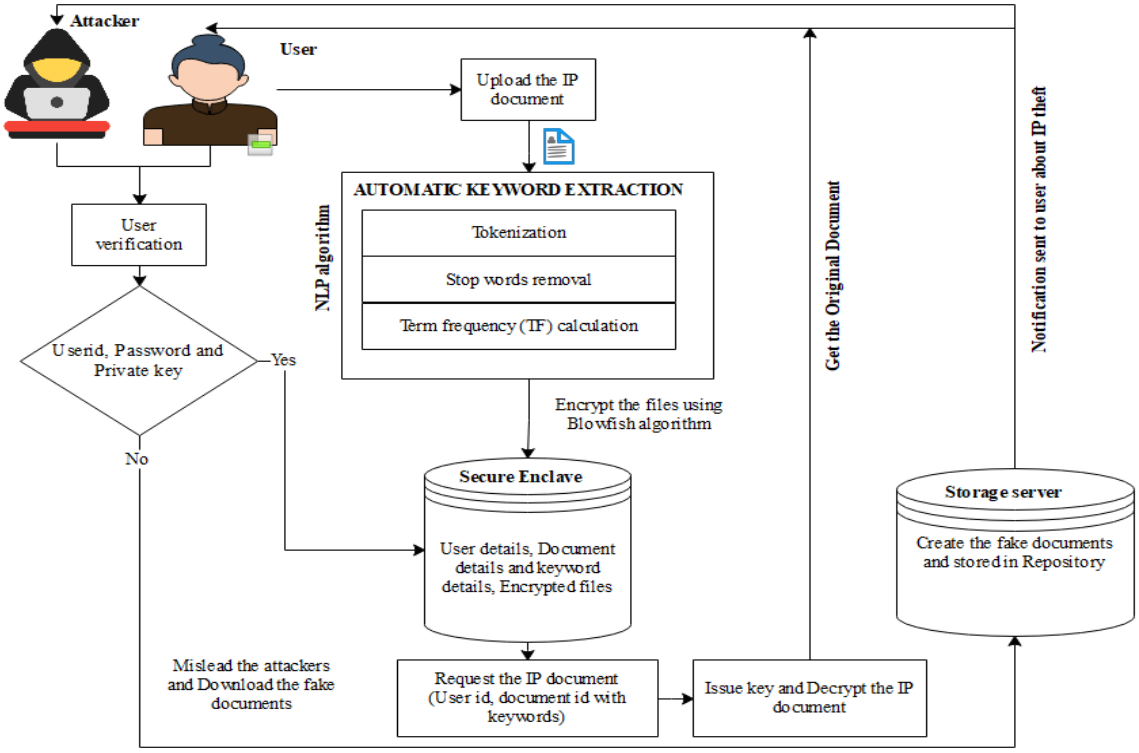


FIGURE: SYSTEM ARCHITECUTRAL DESIGN

V. SOFTWARE DESCRIPTION

Python

Python is a high-level, interpreted programming language that is widely used in various domains such as web development, scientific computing, data analysis, artificial intelligence, machine learning, and more. It was first released in 1991 by Guido van Rossum and has since become one of the most popular programming languages due to its simplicity, readability, and versatility.

MySQL

MySQL is an open-source relational database management system (RDBMS) widely used for managing and storing structured data. It is based on the Structured Query Language (SQL) and supports a wide range of applications, from small-scale projects to large, complex enterprise systems.

PyCharm

PyCharm has a highly customizable user interface, allowing users to tailor the IDE to their specific needs and preferences. This includes customizing the color scheme, key mappings, and even the appearance of the code editor.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

PyCharm also supports various plugins and extensions, enabling users to add new functionality to the IDE or integrate with external tools and services. In addition to its development features, PyCharm also includes tools for project management, such as version control integration with Git, Mercurial, and Subversion. It also provides support for task management and issue tracking through integration with tools like Jira and Trello. PyCharm has a strong focus on code quality and maintainability, providing tools for code inspections, unit testing, and code coverage analysis. This can help developers catch errors and ensure that their code is maintainable and scalable over time. PyCharm also supports multiple Python versions and virtual environments, allowing users to switch between different versions of Python or create isolated environments for different projects. This can help ensure compatibility and prevent version conflicts between different projects. Overall, PyCharm is a comprehensive IDE that can greatly improve productivity and code quality for Python developers. Its extensive feature set, customization options, and focus on code quality make it a popular choice for Python development.

VI. EXPECTED OUTCOME

The proposed system is anticipated to deliver several key outcomes, significantly improving the security and management of sensitive documents. By integrating Natural Language Processing (NLP) and Blowfish encryption, it ensures robust protection against unauthorized access and theft. The system's proactive approach to monitoring user interactions and analysing document content will enable early detection of potential threats, reducing the risk of data breaches. Leveraging TF-IDF for keyword identification enhances contextual understanding, facilitating intelligent classification and prioritization of sensitive documents. Furthermore, the deployment of decoy documents in alternative repositories will mislead attackers, safeguarding genuine assets while providing actionable insights into malicious activities. The incorporation of Blowfish encryption ensures secure access control, restricting decryption capabilities to authorized users and maintaining confidentiality. This multi-layered approach, combining strong encryption, advanced content analysis, and deceptive strategies, enhances organizational resilience against cyber threats, fostering trust in the system's ability to safeguard critical information

VII. CONCLUSION

The increasing frequency and sophistication of cyber threats have exposed the limitations of traditional document security approaches, making it essential to explore innovative, multi-layered solutions for protecting sensitive information. This project addresses these challenges by proposing a hybrid security model that integrates Natural Language Processing (NLP) and Cryptography. By leveraging NLP, the system can intelligently analyze, classify, and monitor document content in real time, enabling the detection of unauthorized changes, extraction attempts, or anomalies that traditional systems often overlook. Through the generation of unique document fingerprints and identification of sensitive content, NLP enhances situational awareness and facilitates proactive threat mitigation. Complementing this, the application of cryptographic techniques such as Blowfish encryption and digital signatures ensures that data remains confidential and unaltered throughout its storage and transmission lifecycle. The combined use of content-aware monitoring and robust encryption forms a powerful defense mechanism, addressing both access control and content protection in a unified manner. Furthermore, incorporating Digital Rights Management (DRM) strategies like watermarking and controlled access adds an additional layer of protection, enhancing traceability and ensuring accountability. This comprehensive approach offers significant advantages across industries that handle sensitive documents, such as finance, healthcare, legal, and government sectors. By delivering a system that is adaptive, secure, and capable of intelligent decision-making, this model not only protects data from theft and misuse but also builds resilience against evolving digital threats. Ultimately, the integration of NLP and Cryptography represents a forward-thinking advancement in document security, offering a reliable, scalable, and proactive framework to ensure the integrity, confidentiality, and availability of critical information in an increasingly digital and interconnected world.

VIII. LITRETURE SURVEY

TITLE	:	The Role of Cybersecurity in Protecting Intellectual Property (2024)
AUTHOR	:	Mavani, Chirag
CONCEPT	:	This research seeks to establish how the cybersecurity measures interact with the management of IP assets across different domains. They analyse existing forms of threats that include hacking, data breaches, and internal threats that are a major concern to the IP's integrity and confidentiality. Furthermore, this research focuses

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

	on how current and emergent technologies and initiatives can be utilized in the prevention and management of the threats; with relevant aspects discussed including encryption, access controls, and consciousness monitoring as key principles in IP shield
LIMITATIONS:	There is no capacity to build and enhance the defensive tools, to encourage innovations and to protect the value and credibility of the IPs
REFERENCES:	Mavani, Chirag, et al. "The Role of Cybersecurity in Protecting Intellectual Property." International Journal on Recent and Innovation Trends in Computing and Communication 12.2 (2024): 529-538.
TITLE :	A Psycholinguistics-Inspired Method to Counter IP Theft using Fake Documents (2024)
AUTHOR :	Denisenko, Natalia
CONCEPT :	By combining psycholinguistic-based surprisal scores and optimization to generate two bilevel surprisal optimization problems (an Explicit one and a simpler Implicit one) whose solutions correspond directly to the desired set of fakes. As bilevel problems are usually hard to solve, we then show that these two bilevel surprisal optimization problems can each be reduced to equivalent surprisal-based linear programs. We performed detailed parameter tuning experiments and identified the best parameters for each of these algorithms
LIMITATIONS:	One interesting hypothesis is that modern LLMs might be better than other models at suggesting plausible replacement words because they use a much larger context than other kinds of language models and encode much more world knowledge because they are trained on vastly more text
REFERENCES:	Denisenko, Natalia, et al. "A Psycholinguistics-Inspired Method to Counter IP Theft using Fake Documents." ACM Transactions on Management Information Systems (2024).
AUTHOR :	Zhu, Hongyu
CONCEPT :	To counteract this, we leverage diffusion models to synthesize unrestricted adversarial examples as trigger sets. By learning the model to accurately recognize them, unique watermark behaviors are promoted through knowledge injection rather than error memorization, thus avoiding exploitable shortcuts. In this paper, we identify the dilemma that poisoning-style model watermarks increase susceptibility to evasion while protecting against theft
LIMITATIONS:	Dual effectiveness underscores its potential as a comprehensive solution to protect deep learning models from a spectrum of complex threats
REFERENCES:	Zhu, Hongyu, et al. "Reliable Model Watermarking: Defending Against Theft without Compromising on Evasion." arXiv preprint arXiv:2404.13518 (2024).

REFERENCES

[1] Mavani, Chirag, et al. "The Role of Cybersecurity in Protecting Intellectual Property." International Journal on Recent and Innovation Trends in Computing and Communication 12.2 (2024): 529-538.
[2] Denisenko, Natalia, et al. "A Psycholinguistics-Inspired Method to Counter IP Theft using Fake Documents." ACM Transactions on Management Information Systems (2024).
[3] Zhu, Hongyu, et al. "Reliable Model Watermarking: Defending Against Theft without Compromising on Evasion." arXiv preprint arXiv:2404.13518 (2024).
[4] Seethala, S. C. (2024). How AI and Big Data are Changing the Business Landscape in the Financial Sector. European Journal of Advances in Engineering and Technology, 11(12), 32–34. https://doi.org/10.5281/zenodo.14575702
[4] Zhang, Ruisi, and Farinaz Koushanfar. "EmMark: Robust Watermarks for IP Protection of Embedded Quantized Large Language Models." arXiv preprint arXiv:2402.17938 (2024).
[5] Wang, Zhenyi, Yihan Wu, and Heng Huang. "Defense against Model Extraction Attack by Bayesian Active Watermarking." Forty-first International Conference on Machine Learning, (2024)

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details