Machine learning and watermarking for accurate detection of AI generated phishing emails.

*electronics*

MDPI

*Article*

# Machine Learning and Watermarking for Accurate Detection of AI-Generated Phishing Emails

**Adrian Brissett and Julie Wall ***

School of Computing and Engineering, University of West London, London W5 5RF, UK; 32138694@student.uwl.ac.uk
* Correspondence: julie.wall@uwl.ac.uk

**Abstract**

Large Language Models offer transformative capabilities but also introduce growing cybersecurity risks, particularly through their use in generating realistic phishing emails. Detecting such content is critical; however, existing methods can be resource-intensive and slow to adapt. In this research, we present a dual-layered detection framework that combines supervised learning for accurate classification with unsupervised techniques to uncover emerging threats. In controlled testing environments, our approach demonstrates strong performance. Recognising that human users are often the weakest link in information security systems, we examine historical deception patterns and psychological principles commonly exploited in phishing attacks. We also explore watermarking as a complementary method for tracing AI-generated content. Together, these strategies offer a scalable, adaptive defence against increasingly sophisticated phishing attacks driven by Large Language Models.

**Keywords:** phishing detection; large language models; AI-generated content; watermarking; techniques; paraphrasing detection; hybrid detection models



Academic Editors: Wei Ji, Hao Fei and Fei Li

Received: 28 February 2025
Revised: 6 June 2025
Accepted: 10 June 2025
Published: 27 June 2025

**Citation:** Brissett, A.; Wall, J. Machine Learning and Watermarking for Accurate Detection of AI-Generated Phishing Emails. *Electronics* **2025**, *14*, 2611. https://doi.org/10.3390/electronics14132611



## 1. Introduction

Artificial Intelligence (AI), first formally articulated in the 1950s, has historically been regarded as a future technology, perpetually just around the corner. However, with the advent of ChatGPT in 2022 and other generative AIs capable of producing innovative content and ideas, that future has arrived. In this era of rapid advancements in Large Language Models (LLMs), it is evident that this transformative technology offers significant benefits. These advanced language models have made remarkable strides in generating coherent and persuasive prose, tackling complex technical questions, and giving millions of people access to high-quality writing and imagery [1].

However, as LLMs evolve, so too do the complexities of cyberthreats. LLMs can be used to emulate a particular person's writing style [2]. Social engineering involves manipulating users into performing specific actions or revealing sensitive information [3]. It remains a critical concept that demands careful attention due to its potentially severe consequences, which can lead to substantial financial losses. Phishing is one of the most popular methods of social engineering, where an attacker seeks to achieve their nefarious aims by sending convincing messages to unsuspecting individuals to steal passwords, credit card numbers, or to elicit a transfer of funds [4]. The term refers to where unsuspecting users are "phished" or lured into a trap, akin to fish to a baited hook [4]. Phishing can also involve the procurement of information, intellectual property, or other sensitive information [5].

Other phishing tactics include tricking recipients into installing malware or voluntarily disclosing private system information. Cybercriminals often exploit a false sense of trust or urgency to infiltrate networks and access devices and user accounts [6].

Detecting these AI-generated phishing emails or messages is critical, particularly when they impersonate banks or other financial organisations. Furthermore, accurately identifying whether responses in identity verification processes are human-generated or produced by automated systems is vital for fraud prevention. The challenge of paraphrasing, which involves rewriting text to mask AI-generated features, further complicates distinguishing AI-generated content from human-authored text [7]. As will be discussed in the literature review, prior studies have evaluated the capabilities of classical machine learning in phishing detection [8], and others have examined the limitations of watermarking techniques under paraphrasing attacks [9]; existing approaches often lack adaptability to evolving tactics. Notably, detecting paraphrased or contextually manipulated content remains a significant challenge, as highlighted in recent surveys [10,11]. Moreover, many systems overlook the psychological strategies that underlie the effectiveness of phishing emails, such as those described by Cialdini's principles of influence [12].

To address these gaps, this study proposes a hybrid framework that combines supervised learning, unsupervised clustering, and watermarking to enhance detection robustness. It is guided by two central research questions:

1. How can machine learning and watermarking be effectively used to detect AI-generated phishing emails?
2. How can detection systems be enhanced to identify paraphrased or contextually manipulated content?

By addressing these questions, the study aims to develop a scalable, cognitively informed detection system capable of identifying AI-generated phishing content even when obfuscated. This work also explores the evolution of phishing techniques, evaluates the use of watermarking for content provenance, and empirically assesses classical machine learning models under varied conditions. The subsequent sections explore the evolution of AI-driven phishing threats and the psychological principles that enhance their effectiveness, including a demonstration of how AI-generated emails can exploit Cialdini's principles of influence. This is followed by a review of related research, a detailed methodology, empirical findings, and practical recommendations for strengthening cybersecurity against these emerging threats.

### *1.1. Evolution of AI-Driven Phishing Threats*

Historically, phishing attacks were rudimentary and overtly greedy, often requesting a wide range of sensitive information like ATM PINs, and they frequently wrote emails poorly [13]. These early phishing emails often contained obvious grammatical errors and inconsistencies, making them easier to detect. Bad actors further refined their approach by employing individuals adept in the English language to craft persuasive communications and mislead recipients [13]. Such a role has become superfluous with the introduction of LLMs, due to the advancements in linguistic capabilities, which have increased the efficacy of phishing attacks.

From the perspective of a bad actor, the earlier AI systems were relatively limited, often lacking the capabilities needed to produce convincing text at scale. State-of-the-art LLMs demonstrate significant advancements in the creation of highly convincing, personalised, and human-like phishing attacks [14]. LLMs made publicly accessible near the end of 2022 have ushered in a qualitatively new era for cybercriminals. The widespread availability of powerful LLMs has now made it both feasible and inexpensive for bad actors to use AI systems to generate phishing emails [14]. Advanced models such as OpenAI's ChatGPT

have been shown to generate personalised and realistic spear phishing emails at scale for mere pennies, merging the worst aspects of both generic and more targeted phishing tactics [6].

The introduction of spear phishing marked a significant shift, adding a layer of deception by impersonating trusted entities to lure victims [6]. A notable instance of its application is the deployment of whaling phishing, a cyberattack specifically aimed at high-ranking executives, such as CEOs and CFOs. The goal is to trick their prey into divulging sensitive information or authorising unauthorised transactions. This type of attack is also known as CEO fraud or Business Email Compromise (BEC) [2]. Pretexting is closely associated with spear phishing or whaling, as it involves creating a web of falsehoods that persuades the target to act [15].

Context-aware attacks become particularly powerful when executed by specialised AI models such as WormGPT (WormGPT is an AI model designed for generating phishing content [16]; ), an LLM fine-tuned for malicious content generation, which is intentionally designed for malicious purposes. In contrast, ethical AI models such as OpenAI's GPT incorporate safeguards to mitigate misuse, though these protections are not entirely foolproof against sophisticated prompt engineering. The prominence of WormGPT underscores that the effectiveness of LLMs is largely influenced by the proficiency of their users. In this context, while fluently written English can be a persuasive tool, the integration of contextual awareness and psychological principles, explored in the next section, can enhance the potential for manipulative success.

### 1.2. *Psychological Principles Behind Phishing Success*

Robert Cialdini, a renowned psychologist, introduced six fundamental principles of influence in his seminal work, Influence: The Psychology of Persuasion (Cialdini, 2007) [12], see Figure 1. Authority, which reflects the tendency to follow figures perceived as credible or powerful. Reciprocity, the impulse to return favours, is leveraged through reward-based lures. Commitment and consistency, which describe how individuals strive to align future actions with prior commitments, underpin multi-step phishing tactics. Social proof, the reliance on others' behaviour in uncertain situations, is used to legitimise fraudulent requests through fabricated statistics or peer actions. Liking, the preference for complying with familiar or appealing sources, enables the impersonation of trusted individuals or brands. Finally, scarcity, which increases perceived value under limited availability, is used to manufacture urgency and pressure immediate responses. These principles, originally studied in contexts like marketing and negotiation [15], also apply to online manipulation techniques. It could be argued that each of Cialdini's six principles of influence aligns with a specific phishing strategy.

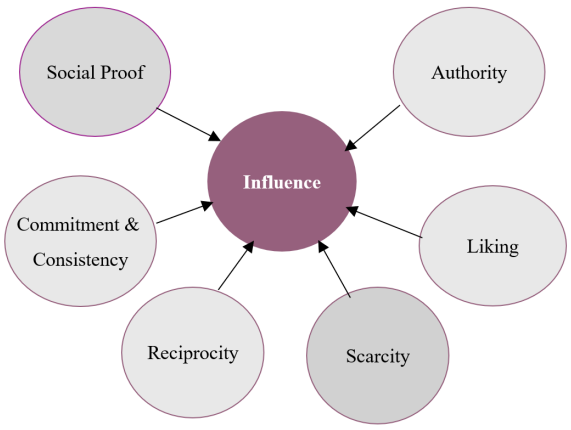


**Figure 1.** Cialdini's six principles of influence, adapted to illustrate how these psychological tools can be exploited in phishing attacks by social engineers [15].

Ethical Disclaimer: The following demonstration has been conducted solely for educational and research purposes, to expose potential vulnerabilities and support the development of effective countermeasures. The authors of this work do not endorse or condone the use of AI tools for deceptive, manipulative, or malicious purposes, including phishing, impersonation, or fraud. Misuse of language models violates both ethical standards and OpenAI's use policies. While ChatGPT is built with ethical safeguards and intended for responsible use, its outputs are shaped by the inputs it receives, a concept known as Garbage In, Garbage Out (GIGO). This principle highlights that even advanced systems can produce harmful outputs when given unethical or malicious prompts. For example, on 19 July 2024, the authors, acting under controlled and ethical conditions, prompted ChatGPT 4.0 to generate phishing messages using Cialdini's six principles of influence. The exact responses from ChatGPT 4.0 are shown in Table 1.

**Table 1.** Cialdini's six principles guided the prompt design used to generate the LLM-generated phishing emails in our dataset.

| Principle | Message |
| --- | --- |
| Authority | "As part of our ongoing commitment to protecting your financial information, we are conducting an immediate security update on all customer accounts. This action is being taken under the direct supervision of our Chief Security Officer, Dr. Emily Richards, who has over 20 years of experience in cybersecurity." |
| Scarcity | "Please note that this security update is available only for a limited time and must be completed within the next 24 h to ensure uninterrupted access to your account. Failure to act within this timeframe may result in temporary suspension of your account for security purposes." |
| Social Proof | "Our records indicate that over 95% of our customers have already completed this security update, ensuring the highest level of protection for their accounts. Join them now in safeguarding your personal and financial information." |
| Reciprocation | "As a token of our appreciation for your prompt attention to this matter, we will credit your account with a bonus of $10 upon completion of the security update process." |
| Commitment and Consistency | "To ensure your account remains secure, please follow the link below and confirm your information. Once you start the process, we encourage you to complete it to uphold the integrity of your account's security." |
| Liking | "We value your trust and are committed to providing you with the best service possible. Our dedicated customer service team, whom many of you have positively rated in recent feedback, is here to assist you 24/7 with any questions or concerns you may have." |

It could be argued that, of all Cialdini's principles, authority can pose the greatest threat in phishing because it exploits entrenched social and organisational norms. As discussed in Section 1.1, whaling, phishing tactics that target senior executives, illustrates the especially dangerous use of authority, in which attackers impersonate executives to issue seemingly legitimate and urgent directives. The perceived credibility of the sender, coupled with workplace norms that discourage questioning authority, often leads employees to act without verification. Unlike other principles that target individual impulses, authority subverts hierarchical structures, making it more pervasive and difficult to resist. While CEO fraud involves impersonating an executive and whaling targets executives directly, both exploit positions of authority within an organisation to manipulate key decision-making processes. Combined with AI's ability to mimic tone and style, this makes authority a particularly powerful weapon in high-stakes phishing attacks. The following section presents a comprehensive review, examining various approaches to distinguishing AI-generated content from human-generated content. Hazell's study [14] assessed the capabilities of LLMs, such as GPT-3.5 and GPT-4, in generating spear phishing and whaling-style emails. The research produced over 600 realistic phishing messages targeting British Members of Parliament, using public biographical data to craft emails that mimicked constituents or assistants. These messages invoked authority and civic responsibility, hallmarks of CEO

fraud, demonstrating how LLMs can replicate the psychological and hierarchical dynamics exploited in high-level phishing attacks. Importantly, no emails were actually sent; the study was a simulation designed to show the feasibility, scalability, and persuasive quality of such attacks if carried out. In the context of MPs, the impersonation strategy emphasised political alignment and civic duty, illustrating how language models can weaponise personal and institutional authority to increase the success rate of phishing campaigns.

## 2. Background and Related Research

OpenAI has developed advanced detection systems to address challenges such as automated misinformation and the misuse of AI in academic settings [17]. These systems employ classifiers designed to differentiate between synthetic and human-generated content, thereby enhancing the capability to identify and mitigate risks associated with AI-generated text. However, the OpenAI classifier currently demonstrates a prediction accuracy rate of only 26% highlighting its limited effectiveness, especially since it has not been evaluated under rigorous adversarial conditions. Additionally, the classifier's reliability decreases significantly when analysing texts shorter than 1000 characters. OpenAI has grappled with the inherent limitation, stating that "it is impossible to reliably detect all AI-written text." This admission was underscored on 20 July 2023, when OpenAI decided to discontinue its AI classifier, citing its "low accuracy rate" as a critical shortcoming. This strategic pivot reflects a remarkable shift in priorities: instead of persisting with an imperfect solution, OpenAI has channelled its expertise toward more pressing challenges, namely, detecting misinformation and plagiarism. These objectives not only demand sophisticated detection mechanisms but also represent a clear divergence from the complex and evolving threat of phishing attempts.

There are two main categories of LLM detectors: specific and general. Specific detectors focus on identifying certain types of language or contexts, such as spam or hate speech. On the other hand, general detectors are designed to recognise a wide array of problematic language, including misinformation and propaganda [17]. General-purpose detectors often perform poorly in practical applications, as they struggle to accurately identify problematic language in various contexts [17]. In contrast, specific-purpose detectors have shown more promising results, yet they face significant challenges. These include sophisticated adversarial techniques, such as paraphrasing and adversarial prompt engineering, which manipulate model output to bypass detection. Additionally, the growing diversity of language models, particularly those supporting multiple languages beyond English, complicates accurate detection [9]. These evolving tactics and technologies underscore the ongoing complexity of effectively identifying AI-generated content.

To mitigate these challenges, several techniques have been proposed, including storing user conversations with language models for retrospective analysis and employing watermarking methods to trace and identify AI-generated content. However, these methods also have limitations and require further development to enhance their reliability and effectiveness. It should be noted that both types of detectors typically rely on supervised learning [17]. Unsupervised learning approaches for detecting AI-generated text use statistical measures like entropy, perplexity, and n-gram frequencies to differentiate human-written from AI-generated content without labelled data [11]. Techniques such as the GLTR framework and zero-shot detection use outlier detection and thresholding to identify AI-generated text [18]. Although these methods generalise effectively across various language models, they are susceptible to adversarial tactics like paraphrasing, highlighting the need for continued research to enhance their robustness. These challenges are explored in depth in [11]. Despite ongoing advancements, developing effective detection

systems remains a complex task due to the rapid evolution of language models and their increasingly diverse applications.

The main research problem addressed by [19] explores the difficulty of distinguishing between text generated by ChatGPT and that written by human experts across various domains. The authors present the Human ChatGPT Comparison Corpus (HC3), a dataset designed to compare human and ChatGPT responses. It includes nearly 40,000 questions with answers from both human experts and ChatGPT across various domains and is used to develop several detection models. According to OpenAI, ChatGPT is fine-tuned from the GPT-3.5 series using Reinforcement Learning from Human Feedback (RLHF), which helps it excel at text-based tasks like classification, information extraction, translation, and even generating code and stories. This meticulous fine-tuning also allows ChatGPT to acknowledge mistakes, challenge incorrect assumptions, and refuse inappropriate requests. The researchers substantiate their claims through evaluations and linguistic analyses, revealing significant differences between human and AI-generated texts, despite ChatGPT's advanced capabilities. Critical linguistic differences exist between texts generated by LLMs and those written by humans, as shown by [19]. AI-generated content generally maintains a formal and neutral tone, marked by high grammatical accuracy and coherent, well-organised structures. In contrast, human-authored texts exhibit greater linguistic diversity and informality, featuring a wider range of lexical choices and more flexible syntactic patterns.

A comprehensive evaluation of eight ML models was conducted to assess their effectiveness in distinguishing AI-generated phishing emails from those written by humans [8]. The study employed a diverse set of models, including Random Forest, SVM, XGBoost, Logistic Regression, and neural networks, trained on a dataset comprising both real phishing emails and AI-generated texts created using WormGPT. The researchers systematically reviewed prior literature to identify 30 key textual features that differentiated AI-generated text from human-written content. Their findings revealed that the neural network achieved the highest accuracy (99.78%), followed closely by SVM (99.20%) and Logistic Regression (99.03%). Beyond classification accuracy, their research examined the interpretability of these models—an essential consideration in real-world applications. Simpler models, such as Naïve Bayes and Logistic Regression, offered greater transparency in decision-making, while more complex models, including neural networks, required post hoc explainability techniques such as SHAP and LIME to approximate the reasoning behind their classifications. Given the inherent trade-off between accuracy and interpretability, the researchers concluded that Logistic Regression represented the optimal balance, making it a viable choice for practical cybersecurity implementations [8]. It has been argued that synthetic content poses a significant threat to financial integrity [20]. In response, watermarking has emerged as a promising pre-emptive detection technique by embedding hidden markers within text to enable the verification of AI-generated content [10]. A detailed explanation is provided in Section 3.2.2. While this approach provides an initial defence, it requires ongoing updates and collaboration with AI developers.

Addressing AI-generated phishing emails is critical due to their potential to deceive even well-trained individuals, especially in the financial sector. Dynamic detection methods are essential to adapt to evolving cybercriminal tactics [21]. The research conducted by [10] focuses on the theoretical and practical aspects of identifying text-generated LLMs by scrutinising the strengths and weaknesses of various detection methods, including pre-emptive strategies that involve training neural networks to differentiate between AI-generated and human-authored texts. These systems are fine-tuned using samples of both types of text to enhance their detection capabilities. The study reviews pre-emptive methods that involve training neural networks to recognise AI-generated text. The "pre-

emptive" aspect signifies that the watermark is applied proactively before the text is used or potentially misused. This is the process of embedding a detectable yet invisible marker or signature within AI-generated content at the time of its creation. This marker, often imperceptible to human readers, is designed to enable later identification of the content as machine-generated [10], even after it has been distributed or modified. This approach can be instrumental in preventing the spread of misinformation, detecting plagiarism, or ensuring the authenticity of content in critical sectors. The watermark does not alter the visible properties of the text but leaves a trace that specialised detection systems can identify, helping to ensure transparency and accountability in the use of AI-generated materials.

Various AI content detection tools were evaluated in [7], including OpenAI (OpenAI's AI Text Classifier, available at: https://platform.openai.com/ai-text-classifier, accessed on 6 July 2024), Writer (Writer's AI Content Detector, available at: https://writer.com/ai-content-detector/, accessed on 19 June 2024), Copyleaks (Copyleaks AI Content Detector, available at: https://copyleaks.com/, accessed on 7 June 2024), GPTZero (GPTZero AI Detector, available at: https://gptzero.me/, accessed on 9 July 2024), and CrossPlag (CrossPlag AI Detector, available at: https://www.crossplag.com/ai-content-detector/, accessed on 20 July 2024), using texts from ChatGPT models 3.5 and 4.0 and human-written texts. They found significant performance variability, highlighting the need for combining AI tools with manual review to improve accuracy and uphold academic integrity. Diverse datasets are crucial for training AI models to detect phishing, as they help reduce biases and generalise to real-world scenarios. Although AI models for phishing detection involve significant costs, they are justified by the benefits of preventing data breaches and protecting sensitive information. In [7], the authors investigated various AI content detection tools and found variability in their effectiveness, with ChatGPT 4.0 being harder to detect due to its human-like text generation. However, current detection tools produce false positives (FPs) and false negatives (FNs) and require manual review, which is resource-intensive and time-consuming. They noted that AI tools can be inconsistent and need constant improvement. While their research did not focus specifically on phishing detection, the principles they discussed, especially ML classifiers, are applicable and beneficial for improving phishing detection techniques.

The research in [22] examined a wide array of detectors for AI-generated text. The researchers sought to address potential abuses of LLMs across multiple domains, including the spread of misinformation, fake news generation, plagiarism in academic contexts, intellectual property protection, and enhancing customer support within financial services. They proposed leveraging ChatGPT for automated detection pipelines and evaluating its zero-shot performance in this context. Their investigation was supported by an empirical analysis using publicly available datasets, such as the OpenAI Human-AI Comparison (OpenAI-HAC) Dataset, which includes paired samples of human-written and AI-generated text. This dataset is specifically designed to assess the effectiveness of AI-generated text detection systems by comparing human-authored and machine-generated content across diverse topics. The results indicated that while ChatGPT struggled to identify AI-generated text, it was more accurate at detecting human-written content. This performance disparity suggested potential applications in downstream detection tasks focused on human-authored text. The study concluded that although ChatGPT was not particularly effective at detecting AI-generated content, its reliability in identifying human-written text could be leveraged to indirectly address the detection challenge. This asymmetry could serve as a foundation for developing more advanced detection tools. Focusing on the financial sector, the researchers investigated ways to enhance current fraud detection methods. They explored how AI could effectively identify phishing emails and messages that impersonate banks or financial institutions. To address this challenge, they proposed implementing watermarking techniques for all publicly accessible language models, as well

as developing sophisticated email filtering systems capable of detecting these watermarks in phishing attempts. These systems would flag suspicious content for further verification, thereby strengthening fraud prevention mechanisms.

Detecting AI-generated text poses both theoretical and practical challenges, which can be addressed using two primary methods: pre-emptive approaches like watermarking and post hoc techniques such as zero-shot classifiers. A comprehensive evaluation of these methods has been provided in [9], highlighting their effectiveness and limitations across diverse scenarios, and offering valuable insights into their practical applications. Fourteen detection tools were selected for testing, including free online tools and commercial plagiarism detection systems like Turnitin (Turnitin, "How to Implement Citation and Paraphrasing Into The Writing Process," Turnitin Blog, accessed on 1 February 2025, https://www.turnitin.com/blog/how-to-implement-citation-and-paraphrasing-into-the-writing-process) and PlagiarismCheck (PlagiarismCheck, "Plagiarism Detection Tool," accessed on 1 February 2025, https://plagiarismcheck.org/) are widely used to ensure academic integrity. The study found significant limitations in current tools designed to detect AI-generated text. None of the fourteen tools tested, including Turnitin and PlagiarismCheck, achieved complete accuracy, with the best performing at less than 80%. The tools often misclassified AI-generated text as human-written, particularly when the text was manually edited or paraphrased by machines. Machine-translated texts also posed challenges, leading to a 20% drop in detection accuracy. Overall, the study concluded that these tools were unreliable for making high-stakes decisions in academic settings and should not be solely relied upon to determine academic misconduct.

The study in [8] demonstrated the effectiveness of eight ML models in detecting AI-generated text, particularly in phishing email detection. The models evaluated included Random Forest, SVM, XGBoost, KNNs, Naive Bayes, Neural Networks, and Logistic Regression. All models showed high accuracy, with Logistic Regression achieving the highest accuracy of 99.03% due to its flexibility and simplicity. This result highlighted the potential of Logistic Regression as a reliable tool for phishing detection. To further enhance detection capabilities, the researchers proposed innovative solutions such as watermarking language models and developing advanced email filtering systems capable of detecting AI-generated text. These strategies not only improved detection accuracy but also proactively addressed the evolving threat of evasion techniques like paraphrasing, paving the way for more robust and adaptive security solutions. The study leveraged the "Human-LLM Generated Phishing-Legitimate Emails Dataset," containing both human-written and AI-generated phishing emails. To create realistic phishing scenarios, WormGPT was used, effectively simulating cybercriminal tactics by incorporating Cialdini's principles of influence. The researchers identified thirty key textual features—such as coherence, stylometric properties, and linguistic patterns—that effectively differentiated human and AI-generated texts, significantly improving detection precision.

In pursuit of computational efficiency, the researchers chose lightweight and interpretable models, demonstrating that high accuracy could be achieved without complex architectures. Logistic Regression was particularly successful in identifying subtle distinctions between human and AI-generated phishing emails. To further strengthen model resilience against paraphrasing attacks, the researchers recommended expanding datasets and testing in adversarial environments. Although the study acknowledged the complexity of the paraphrasing challenge, it laid a strong foundation for future research and highlighted the exciting potential for developing even more advanced detection methods. While the study presented promising solutions, further investigation is needed to enhance adaptability to evolving AI models and to develop more effective countermeasures against sophisticated evasion techniques, such as advanced paraphrasing attacks. Despite these

advancements, several gaps remain. More adaptable and robust detection methods are needed to counter evasion techniques like paraphrasing, which current models struggle to detect. Expanding datasets and testing in adversarial settings would better simulate real-world scenarios. Although watermarking was suggested as a countermeasure, its effectiveness against advanced paraphrasing is uncertain. There is also a need for more interpretable models that provide transparent explanations for non-technical users. The study did not, however, explore unsupervised learning approaches, which could enhance adaptability by detecting novel phishing patterns without labelled data. Methods like clustering or anomaly detection could improve performance against emerging threats. Additionally, multi-class models that distinguish between human and AI-generated text would offer a more comprehensive detection approach. Addressing these gaps is essential for developing more adaptive, resilient, and user-friendly phishing detection systems.

### 2.1. Watermarking Techniques for AI-Generated Text

Watermarking facilitates the detection of text generated by LLMs by embedding identifiable statistical patterns within the output [10]. As a background, watermarking is a widely employed technique for verifying the authenticity and provenance of digital content. It has traditionally involved embedding identifiable signatures, either human-written or cryptographic, into data to assert authorship or ownership [7]. This method has been used across both pre-digital and digital contexts to certify that a given text or media originates from the claimed source. In the context of LLMs, watermarking has been proposed as a means to detect AI-generated content. There are two primary methods: soft watermarking and cryptographic watermarking. Soft watermarking involves subtly guiding the language model to prefer a randomised subset of vocabulary, referred to as "green" tokens, during the generation process. This selection is imperceptible to human readers and does not significantly affect the fluency or coherence of the output. A statistical test can then be applied to detect whether a watermark is present, based on the distribution of these tokens [7]. The watermarking system classifies a passage as AI-generated if it contains a disproportionately high number of tokens from the aforementioned green list [10]. Cryptographic watermarking, on the other hand, leverages the pseudo-random nature of token sampling in language models. By using cryptographically secure one-way functions to generate deterministic random seeds, the generation process can be subtly constrained in a way that embeds a verifiable watermark. This method ensures that content generated under specific seeds can be identified post hoc with high confidence [9,10].

### 2.2. Addressing Research Gaps in Phishing Detection

In our research, we advance the field of AI-generated phishing detection by addressing several key gaps identified in the literature, with a focus on expanding beyond the limitations of supervised learning and leveraging the strengths of hybrid frameworks. Supervised models, while valuable, often rely on historical, labelled attack data, which limits their ability to adapt to novel or AI-enhanced phishing threats. They can struggle with detecting subtle deviations or paraphrased content, and their dependency on labelled datasets can introduce bias and increase the cost and complexity of maintenance. To build on and move beyond these limitations, we introduce a hybrid detection framework that combines supervised and unsupervised learning with watermarking. This integrated approach significantly enhances detection capabilities, particularly for novel or reworded phishing attempts. We contribute a new phishing-specific dataset composed of both human- and AI-generated emails, with embedded watermark tokens that support traceability and detection. Our work also translates watermarking from theory to practice, demonstrating its concrete utility in phishing detection systems. In response to the challenges of lexical

approaches like TF-IDF, we incorporate anomaly detection and clustering to improve the system's sensitivity to subtle variations in language. We bring psychological realism into the discourse, an element often overlooked in previous research. Designed with deployment in mind, our framework includes real-time learning and adaptability, offering a scalable and future-ready defence mechanism.

Kirchner et al. introduce a novel watermarking technique designed to embed subtle, statistical patterns into AI-generated text while preserving its semantic content and readability [23]. Unlike earlier approaches that may compromise fluency or coherence, this method subtly alters word choice probabilities, described metaphorically as giving the text a "digital accent" without impacting the overall meaning or naturalness of the output. The watermark is detectable through statistical analysis rather than explicit token tagging, making it significantly more resilient to paraphrasing attacks compared to traditional methods previously discussed. When contrasted with the TF-IDF-based approach, this method strikes a practical balance between preserving linguistic integrity and ensuring traceability. Its robustness makes it suitable for dynamic environments such as financial communications or academic assessments, where the authenticity of content must be guaranteed. This semantic-preserving watermarking approach expands the capabilities available for preemptive AI content detection, aligning with emerging research that promotes combining statistical signatures with transparency and usability across varied domains [23].

## 3. Proposed Hybrid Detection Framework

This research extends the work of Greco et al. [8] by addressing the challenge of paraphrasing attacks, which pose additional difficulties for phishing detection models, and by proposing a post hoc detection framework to enhance existing systems. We use the dataset introduced by Greco et al., which includes 1000 human-written phishing emails from known malicious email repositories, as well as 1000 AI-generated phishing emails produced using WormGPT. These AI-generated samples were created by Greco et al. using prompts that varied in both topic (e.g., tax scams, job offers) and persuasive strategy, based on Cialdini's principles. Table 1 illustrates these persuasive strategies, which are annotated in the dataset and used as features in our model. Our approach leverages these psychologically grounded cues to flag suspicious emails with interpretable outputs, thereby improving both detection accuracy and user trust. Additionally, we incorporate unsupervised methods, such as K-Means Clustering, to identify novel phishing patterns and support adaptive threat detection.

### *3.1. ModelJustification: Why Logistic Regression?*

Logistic Regression was selected as the primary supervised learning model in our study for several compelling reasons, combining empirical performance with interpretability and practical applicability to phishing detection. First, from a performance perspective, Greco et al. [8] reported that Logistic Regression achieved an accuracy of 99.03% in detecting phishing emails generated by LLMs, placing it among the top three performers alongside SVMs and neural networks. Although neural networks outperformed slightly in raw accuracy (99.78%), the marginal gain did not outweigh the interpretability trade-off. Crucially, Logistic Regression is a white-box model that enables local interpretability by highlighting the features most responsible for classification decisions. This is especially important in phishing contexts, where human–computer interaction (HCI) factors, such as user trust and the ability to explain decisions clearly, play a critical role. Logistic Regression offers a rare blend of high performance, explainability, computational efficiency, and adaptability to engineered signals such as watermarks, making it an ideal choice for our phishing detection framework.

### 3.2. Technical Environment, Data Collection, and Description

The training phase was conducted on a Surface Laptop Studio with an 11th Gen Intel® Core™ i5-11300H processor and 16 GB of RAM. The device was sourced from the Microsoft Store on Oxford Street, London, UK. The Human-LLM Generated Phishing-Legitimate Emails Dataset was used, comprising 1000 human-generated phishing emails from the "Nazario" collection and 1000 AI-generated phishing emails crafted using WormGPT. All AI-generated phishing content was ethically created and used solely within a controlled research environment for academic purposes. These AI-generated emails varied in topic and persuasion techniques to ensure diversity. The dataset also included 30 textual features, such as TF-IDF, to evaluate word importance and capture linguistic patterns. Each category contained 1000 samples, ensuring balanced distribution. The data processing pipeline integrated both supervised and unsupervised learning techniques with watermarking for enhanced detection. The workflow began with data acquisition using the Human-LLM Generated Phishing-Legitimate Emails Dataset, ensuring a balanced mix of human-generated, AI-generated, and legitimate emails. Data cleaning included tokenisation, normalisation, and special character removal. TF-IDF was used for feature extraction, allowing the model to capture vocabulary and structural differences between human-written and AI-generated text. Anomaly detection was enhanced through K-Means Clustering, which identified paraphrased phishing content deviating from known patterns. Watermarking techniques were used to distinguish synthetic from human-written text, ensuring accurate detection of sophisticated phishing attempts.

Model training and optimisation followed a dual-layer approach, integrating Logistic Regression for high-accuracy binary classification and unsupervised learning for detecting evasive phishing patterns. K-Means Clustering grouped similar phishing emails based on textual features, effectively identifying paraphrased variations. Combining unsupervised methods with Logistic Regression created a multi-layered defence system, increasing accuracy and resilience against evolving phishing tactics. This strategy effectively utilised TF-IDF for linguistic feature extraction, clustering for anomaly detection, and watermarking for synthetic content filtering. By leveraging both supervised and unsupervised learning techniques, the approach maximised the detection accuracy and adaptability to emerging phishing tactics. The hybrid approach combined watermarking with Logistic Regression to flag watermarked AI-generated content, while non-watermarked emails were further classified. The dataset was split 80:20 for training and testing, with a feedback loop for continuous improvement. Various performance metrics, including accuracy, precision, recall, and F1-score, were calculated. Adversarial testing was conducted to evaluate the model's resilience to paraphrasing attacks. After testing, the model was deployed within an email security framework, with watermarking verification providing an additional security layer. Continuous monitoring and updates enabled dynamic adaptation to new phishing tactics. The multi-layered approach, combining TF-IDF, clustering, and watermarking, ensured a highly accurate and adaptable phishing detection system.

#### 3.2.1. Text Representation Using TF-IDF

We used the Term Frequency–Inverse Document Frequency (TF-IDF) algorithm to convert raw email text into numerical features for machine learning. TF-IDF highlights words that are frequent in a document but rare across the dataset, helping identify terms of greater significance [24]. In our experiments, particularly Logistic Regression in Experiment 1, TF-IDF was used to extract key lexical patterns from unigrams and bigrams. While TF-IDF can also be used for measuring textual similarity [25], we applied it strictly for feature extraction to support classification between human-written and AI-generated emails. It is worth mentioning that TF-IDF captures frequency patterns rather than meaning, and thus

lacks semantic understanding. This makes it vulnerable to paraphrasing, where content is reworded without changing its underlying intent. Ref. [9] show that such attacks degrade the performance of frequency-based detectors, as semantic structure remains intact while surface features vary. To improve robustness, future approaches should incorporate semantic features that capture meaning beyond individual terms. Techniques such as word embeddings (e.g., Word2Vec, GloVe, FastText) and contextual models like BERT or Sentence-BERT can model deeper relationships in language. These representations allow classifiers to generalise across paraphrased or stylistically varied text. Ref. [19] states that WormGPT was used to generate the LLM-based content and describes it as a version of ChatGPT fine-tuned to comply with malicious requests. The specific underlying model version (e.g., GPT-3.5 or GPT-4) is not disclosed.

#### 3.2.2. Watermarking Implementation Details and Impact on Performance

We implemented a watermarking token, written as ##WATERMARK##, by appending it to the end of each AI-generated phishing email. This acted as a simple digital tag, allowing the system to recognise that the content had been generated by an LLM. As discussed in Section 2.1, this form of tagging is known as a soft watermark because it did not affect how the message appeared to human readers, but could still be detected programmatically. To identify these watermarks, we implemented a custom function that scanned each email for the presence of the token. If the token was detected, the email was flagged for further analysis. We also processed the email content using TF-IDF to identify the most important words in each message. This was combined with Logistic Regression. While Logistic Regression by itself achieved an accuracy of 83.62%, the inclusion of watermarking improved the accuracy to 85% (as shown in Section 4.3). Thus, the watermarking did enhance performance, albeit moderately.

### *3.3. Experiment 1—Phishing Detection Using Logistic Regression (No Watermarking)*

The primary objective of this experiment was to train a model capable of effectively distinguishing between phishing emails and legitimate ones. Specifically, it evaluated the ability of Logistic Regression to identify phishing emails generated by LLMs. The approach was based on the analysis of text-based features, using tokenisation and vectorisation techniques. Tokenisation involves breaking down raw text into smaller units (tokens), such as words, subwords, or characters, to facilitate structured input for the model [26] (p. 199). Vectorisation then converts these tokens into numerical representations, allowing machine learning algorithms to analyse and classify the textual data [27].

### *3.4. Experiment 2—K-Means Clustering of Human vs. LLM Emails (No Watermarking)*

This experiment explored the effectiveness of unsupervised learning. It specifically evaluated the use of K-Means Clustering to distinguish between emails written by humans and those generated by AI. The central aim was to assess whether clustering could reveal underlying structural or lexical patterns unique to AI-generated phishing content, without relying on labelled data. Emails were first transformed into numerical representations using TF-IDF vectorisation to capture word usage and n-gram patterns. These feature vectors were then clustered using K-Means, which grouped similar texts based on their linguistic characteristics. To facilitate interpretation and visualisation of the clustering outcome, Principal Component Analysis (PCA) was applied to reduce dimensionality. This unsupervised approach allowed for the identification of potential anomalies where manual labelling was infeasible. The objective was to measure clustering performance and visually inspect the degree of separation between AI- and human-generated content. While K-Means can identify a subset of AI-generated emails with high precision, recall remains a challenge due to the paraphrasing capabilities of modern language models.

### 3.5. Experiment 3—Phishing Detection Combining Logistic Regression and Watermarking

The purpose of Experiment 3 was to build on the work of Greco et al. [8] by embedding identifiable watermark tokens into AI-generated phishing emails and evaluating the feasibility of detecting such content through supervised learning. A watermark-based detection system was implemented by appending a unique token to the end of each phishing email. This modified dataset was used to train a Logistic Regression classifier capable of distinguishing between legitimate and AI-generated phishing emails based on text features. The experiment used a balanced dataset of 2000 emails, comprising 1000 legitimate emails written by humans and 1000 phishing emails generated by LLMs. The text data were preprocessed and transformed into feature vectors using TF-IDF with bigram support. The dataset was then be split into training (80%) and testing (20%) subsets. The model was trained on the training set and evaluated using standard performance metrics and a confusion matrix based on the test set.

### 3.6. Experiment 4—Phishing Detection Combining K-Means Clustering and Watermarking

This experiment evaluated the effectiveness of combining Logistic Regression with K-Means Clustering in a two-tiered defence system. K-Means Clustering served as the first layer, identifying anomalies and grouping similar phishing patterns based on textual features. Logistic Regression functioned as the second layer, classifying messages that lacked clear clustering patterns or exhibited novel phishing techniques. By leveraging both methods, this approach enhanced detection accuracy, ensuring that even if attackers attempted to evade detection through paraphrasing or bypassing standard phishing filters, the system could still identify fraudulent emails by detecting linguistic inconsistencies and behavioural anomalies. The most significant performance improvement was observed when watermarking was combined with unsupervised learning techniques, achieving a perfect 100% accuracy in these controlled tests. This indicates that its true value emerges when used as part of a layered detection system.

### 3.7. Experiment 5—Semantic Watermarking via Lexical Substitution

To address alternative approaches regarding the artificiality of visible watermark tokens, we implemented a semantic-preserving watermarking technique inspired by [23]. Specifically, this method subtly embedded signals into text via lexical substitution—replacing common words with contextually appropriate synonyms using the WordNet lexical database. This approach retained the original meaning while modifying the surface form, mimicking watermarking that would be less detectable by humans and more resilient in adversarial settings. These watermarked texts were then subjected to Logistic Regression classification using TF-IDF features to assess their detectability.

### 3.8. Metric Evaluation

Evaluation strategies were tailored to the design of each experiment. Experiment 1, which used Logistic Regression without watermarking, was assessed using 5-fold stratified cross-validation to enhance robustness and minimise overfitting. Experiment 3, involving Logistic Regression with watermarking, used a standard 80/20 train–test split to evaluate performance on unseen data. Experiment 2 (K-Means without watermarking) and Experiment 4 (K-Means with watermarking) were both unsupervised; for these, evaluation was conducted by mapping the resulting clusters to ground-truth labels post hoc. Notably, Experiment 4 operated on the full dataset without a split due to its controlled conditions and consistent watermarking. Across all experiments, model performance was measured using standard classification metrics: accuracy, precision, recall, and F1-score. These were derived from counts of true positives (TP), true negatives (TN), false positives (FP), and

false negatives (FN). Accuracy indicates overall correctness; precision captures the proportion of correct positive predictions; recall reflects sensitivity to actual positive instances; and the F1-score, as the harmonic mean of precision and recall, offers a balanced view, especially in scenarios with class imbalance [28].

## 4. Results and Discussion

In this section, we evaluate the results for the supervised learning approach using Logistic Regression, an unsupervised learning model, a watermarking approach, and a combined method utilising both watermarking and unsupervised learning. Before presenting the results, it is important to reiterate the primary goal of this research paper, which was to detect phishing messages, particularly those generated by AI.

### *4.1. Experiment 1—Phishing Detection Using Logistic Regression (No Watermarking)*

This experiment evaluated a Logistic Regression model trained on TF-IDF features to classify emails as legitimate or AI-generated phishing. The dataset was balanced with 1000 instances per class (2000 total), and text was vectorised using unigrams and bigrams (with a maximum of 5000 features). Using 5-fold stratified cross-validation, the model achieved an overall accuracy of 99.65%.

As shown in Table 2 and Figure 2, the model reached an F1-score of 1.00 for both classes, correctly identifying 995 out of 1000 legitimate emails and 998 out of 1000 phishing emails, with only seven total misclassifications. These results confirm that Logistic Regression, even without watermarking, can deliver near-perfect performance on a clean, balanced dataset.

**Table 2.** Classification report—Logistic Regression (5-fold cross-validation).

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Legitimate | 1.00 | 0.99 | 1.00 | 1000 |
| Phishing | 1.00 | 1.00 | 1.00 | 1000 |
| Accuracy | | 0.9965 (1993 / 2000) | | |
| Macro Avg | 1.00 | 1.00 | 1.00 | 2000 |
| Weighted Avg | 1.00 | 1.00 | 1.00 | 2000 |

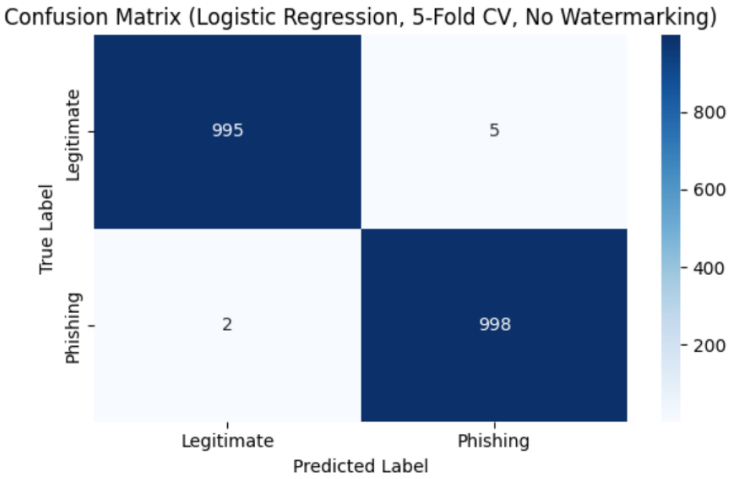


**Figure 2.** Confusion matrix of Logistic Regression using 5-fold cross-validation without watermarking, with minimal misclassifications.

### *4.2. Experiment 2—K-Means Clustering of Human vs. LLM Emails (No Watermarking)*

This experiment assessed the effectiveness of unsupervised learning using K-Means Clustering to distinguish between human-written and AI-generated phishing emails. The dataset included 1000 instances of each type. K-Means achieved an overall accuracy of

82.25%. It demonstrated perfect precision (1.00) for AI-generated emails but lower recall (0.65), yielding an F1-score of 0.78 for that class, as shown in Table 3 and Figure 3.

**Table 3.** K-Means Clustering—performance metrics.

| Category | Precision | Recall | F1-Score |
|---|---|---|---|
| Human—Legitimate | 1.00 | 1.00 | 1.00 |
| LLM—Legitimate | 1.00 | 0.65 | 0.78 |
| Accuracy | | 0.8225 | |
| Macro Avg | 1.00 | 0.825 | 0.89 |
| Weighted Avg | 1.00 | 0.825 | 0.89 |

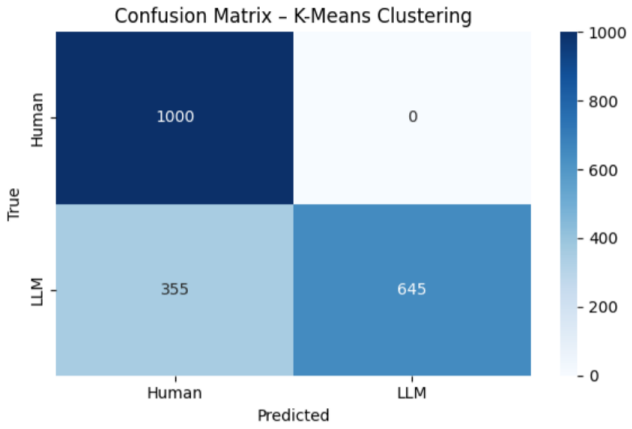


**Figure 3.** Confusion matrix for K-Means Clustering applied to human and LLM-generated emails. While human-written emails are perfectly identified, 35.5% of LLM-generated phishing emails are misclassified.

Thus, these results would suggest that K-Means can reliably detect more distinctive AI-generated content but struggles with subtler cases. See Table 3 for the performance metrics.

### *4.3. Experiment 3—Phishing Detection Combining Logistic Regression and Watermarking*

A watermark-based detection system was integrated with a supervised Logistic Regression classifier to evaluate its effectiveness in identifying AI-generated phishing emails. A balanced dataset of 2000 emails (1000 legitimate and 1000 phishing) was created, with a visible watermark token appended to all AI-generated phishing messages. The model was trained on 80% of the data and tested on the remaining 20% (400 emails). Only one misclassification occurred, demonstrating high model accuracy and watermark effectiveness.

As shown in Table 4 and Figure 4, the classifier achieved an accuracy of 99.75%. All 200 phishing emails in the test set contained the watermark, and 199 were correctly classified. These results demonstrate that watermark-based signals can significantly enhance detection performance in controlled settings, though robustness against obfuscation remains an open challenge.

**Table 4.** Watermarking and Logistic Regression—performance metrics.

| Category | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Legitimate | 1.00 | 1.00 | 1.00 | 200 |
| Phishing | 1.00 | 0.99 | 1.00 | 200 |
| Accuracy | | 0.9975 | | |
| Macro Avg | 1.00 | 1.00 | 1.00 | 400 |
| Weighted Avg | 1.00 | 1.00 | 1.00 | 400 |

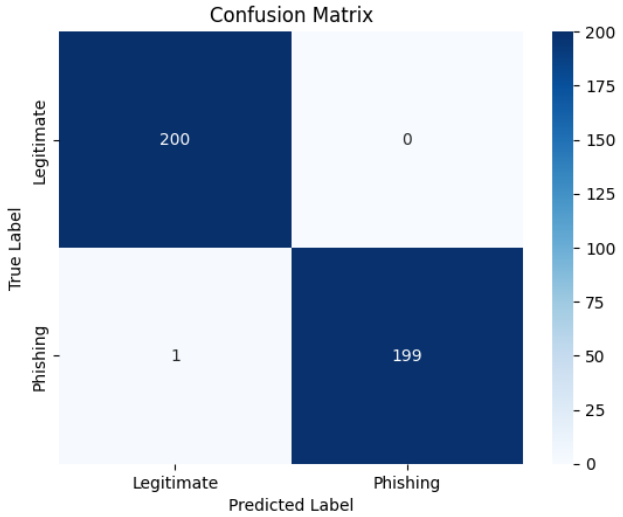


**Figure 4.** Confusion matrix showing near-perfect classification of legitimate and LLM-generated phishing emails using Logistic Regression with watermarking.

### *4.4. Experiment 4—Phishing Detection Combining K-Means Clustering and Watermarking*

This experiment evaluated the effectiveness of combining text watermarking with unsupervised K-Means Clustering. The dataset was balanced and split 80:20 for training and testing, with watermark tokens embedded in all AI-generated samples. After performing clustering on TF-IDF features, each K-Means cluster was mapped to the most common label using majority vote, enabling indirect classification without supervised learning. The model achieved 100% accuracy, with perfect precision, recall, and F1-score for both human and LLM classes, as reported in Table 5. The confusion matrix in Figure 5 confirms that no misclassifications occurred across the 2000 test emails.

**Table 5.** K-Means on TF-IDF Features with Watermarking – Performance Metrics

| Category | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Human-Generated (Legitimate) | 1.00 | 1.00 | 1.00 | 1000 |
| LLM-Generated (Phishing) | 1.00 | 1.00 | 1.00 | 1000 |
| Accuracy | | | 1.00 | |
| Macro Avg | 1.00 | 1.00 | 1.00 | 2000 |
| Weighted Avg | 1.00 | 1.00 | 1.00 | 2000 |

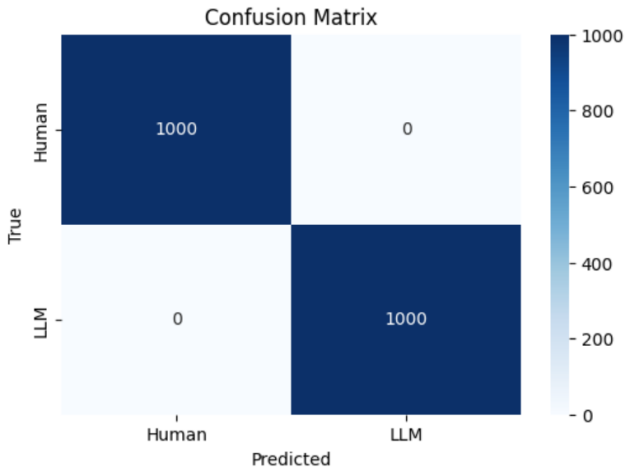


**Figure 5.** Confusion matrix illustrating perfect separation of human and LLM-generated emails using K-Means on TF-IDF features with watermarking.

The 2D PCA projection in Figure 6 reveals a clear spatial separation between clusters, demonstrating strong feature distinctiveness introduced by watermarking. Two distinct groups of points are visible: one cluster (purple) corresponds to human-written emails, while the other cluster (red) represents AI-generated phishing emails.

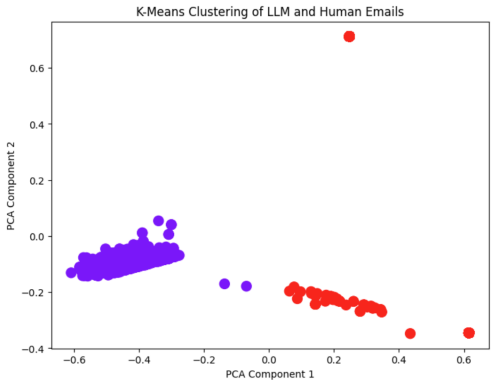


**Figure 6.** 2D PCA projection reveals clear spatial separation between clusters, where purple corresponds to human-written emails and red represents AI-generated phishing emails.

These results indicate that, under controlled conditions, even simple unsupervised methods like K-Means can perfectly separate AI-generated content from human-written text when watermarks are applied consistently.

### *4.5. Experiment 5—Semantic Watermarking Detection*

In this experiment, we evaluated the detectability of AI-generated phishing emails watermarked through synonym-based lexical substitution. A Logistic Regression classifier was trained on TF-IDF features derived from our dataset. The model achieved an accuracy of 98.5%, with precision and recall exceeding 97% across both classes. No LLM-generated phishing emails were misclassified as human-written. This proves that even when the watermark is imperceptible to humans, it can be statistically detected with high reliability.

Table 6 and Figure 7 outline the results. This result enhances the versatility of prior experiments by simulating watermarking techniques that more closely reflect real-world deployment scenarios.

**Table 6.** Classification report for semantic watermarking.

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Human | 1.00 | 0.97 | 0.98 | 197 |
| LLM | 0.97 | 1.00 | 0.99 | 203 |
| Accuracy | | 98.50% | | |
| Macro Avg | 0.99 | 0.98 | 0.98 | 400 |
| Weighted Avg | 0.99 | 0.98 | 0.98 | 400 |

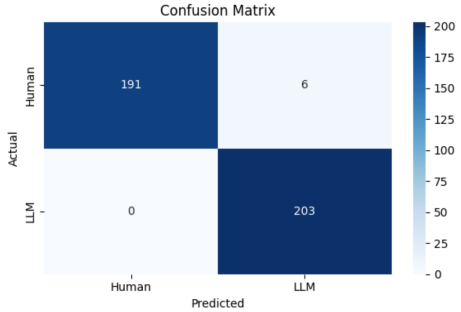


**Figure 7.** Confusion matrix illustrating near-perfect classification of human and LLM-generated emails using Logistic Regression on TF-IDF features with semantic-preserving watermarking.

### *4.6. Discussion*

We conducted five experiments to evaluate the effectiveness of phishing detection models using both supervised and unsupervised learning, with and without watermarking. Experiment 1 used Logistic Regression without watermarking and achieved an accuracy of 99.65%, establishing a strong baseline. Experiment 2 applied K-Means Clustering without watermarking and achieved 82.25% accuracy, highlighting the potential of unsupervised learning to identify AI-generated phishing emails, though with lower recall for more subtly crafted messages. Experiment 3 combined Logistic Regression with visible token watermarking, achieving 99.75% accuracy, showing that embedded cues can enhance detection performance. Experiment 4, which paired K-Means Clustering with visible token watermarking, achieved a perfect 100.00% accuracy under controlled conditions, demonstrating how watermarking can strengthen even unsupervised approaches. Experiment 5 implemented a semantic-preserving watermarking approach using synonym substitution and achieved 98.50% accuracy with Logistic Regression. While slightly below the 99.03% benchmark set by [8], our method was more aligned with realistic deployment conditions. These results suggest that semantic-preserving watermarking strikes a viable balance, delivering high accuracy without sacrificing fluency or semantic integrity. These results, summarised in Table 7, reinforce the role of watermarking, both visible and semantic, in enhancing detection performance across different learning settings.

**Table 7.** Comparison of experiments based on accuracy and standard deviation.

| Experiment | Method | Accuracy (%) | Std. Deviation |
|---|---|---|---|
| Experiment 1 | Logistic Regression (No Watermarking) | 99.65 | 0.0055 |
| Experiment 2 | K-Means Clustering (No Watermarking) | 82.25 | 0.0105 |
| Experiment 3 | Logistic Regression + Visible Token Watermarking | 99.75 | 0.0057 |
| Experiment 4 | K-Means Clustering + Visible Token Watermarking | 100.00 | 0.0000 |
| Experiment 5 | Logistic Regression + Semantic Substitution Watermarking | 98.50 | 0.0068 |
| [8] | Logistic Regression (Baseline Reference) | 99.03 | 0.0062 |

## 5. Conclusions

This research addressed the growing challenge posed by AI-generated phishing emails, particularly those created using LLMs. It investigated two central questions:

(1) How can machine learning and watermarking be effectively leveraged to detect AI-generated phishing emails?, and (2) How can detection systems be enhanced to identify paraphrased or contextually manipulated content that may evade traditional filters?

To explore these questions, we developed a hybrid detection framework that integrated supervised learning (Logistic Regression), unsupervised clustering (K-Means), and watermarking. K-Means Clustering was employed to group similar writing patterns and expose linguistic anomalies common in AI-generated content. This enhanced the model's adaptability to evolving phishing strategies, an area frequently cited as a weakness in traditional, static filters [10,11]. Our experiments showed that TF-IDF was effective at capturing lexical features characteristic of AI-generated text, such as high n-gram frequency and repetitive phrasing. However, its susceptibility to paraphrasing and lack of semantic understanding underscored the need for a complementary mechanism. Watermarking addressed this limitation by embedding identifiable signals directly into the text, significantly boosting detection performance. In combination with clustering, it enabled perfect classification in controlled environments, demonstrating its effectiveness as a robust discriminative feature. Unlike earlier approaches that rely on statistical or syntactic signals [8], our framework incorporates both linguistic structure and psychological manipulation patterns, including authority and urgency cues as outlined by Cialdini [12]. This hybrid model not only improves transparency but also increases resilience to more sophisticated phishing

tactics. Importantly, the value of watermarking extends well beyond phishing detection. Embedding traceable markers in AI-generated text has broader implications for protecting intellectual property, verifying content provenance, and meeting emerging regulatory expectations for responsible AI. As LLMs continue to be adopted in sectors ranging from publishing to software development, watermarking offers a scalable, lightweight mechanism to support attribution, trust, and accountability. This aligns directly with the objectives of standards such as ISO/IEC 42001 [29], which call for transparency, traceability, and risk-based governance in AI systems. Our work contributes to these goals by demonstrating a practical means of identifying AI-generated content in high-risk communication channels.

*Future Work*

Our findings have actionable implications for a wide range of stakeholders. Businesses, especially in high-risk sectors like finance and healthcare, can integrate our detection framework into existing email filters and threat intelligence systems to enable earlier intervention and reduce exposure to fraud. Government agencies may adopt it as part of a layered cybersecurity strategy to protect critical infrastructure and public services against state-sponsored or organised phishing campaigns.

AI developers and platform providers also play a critical role. Embedding watermarking protocols at the model level could facilitate downstream detection of malicious content and align with emerging standards for responsible AI, such as ISO/IEC 42001. Our framework supports this by enabling the traceability of AI-generated content, promoting transparency in detection, and providing interpretable analyses that would aid compliance and inspire trust. Our results show that Logistic Regression on TF-IDF features achieves near-perfect classification performance, validating the feasibility of detecting semantically obfuscated AI-generated text. This finding strengthens the case for embedding sophisticated watermarking at the model level and highlights the importance of evaluating detection systems under adversarial and real-world conditions. Future research should extend this line of work by benchmarking against transformer-based models and testing performance across languages and paraphrasing strategies.

To advance this work, future research should focus on integrating transformer-based models such as BERT and Sentence-BERT, which have shown greater resilience to paraphrasing and deeper contextual understanding [19]. Addressing the English-centric nature of current systems is also essential, as multilingual capability and generalisability remain major limitations in current phishing detection systems [11]. Expanding multilingual datasets, evaluating cross-linguistic performance, and incorporating local explainability techniques such as SHAP and LIME would further enhance insight and build user trust in real-world detection environments.

**Author Contributions:** Conceptualization, A.B. and J.W.; Methodology, A.B.; Software, A.B.; Validation, A.B.; Formal analysis, A.B.; Investigation, A.B.; Data curation, A.B.; Writing – original draft, A.B.; Writing – review & editing, A.B. and J.W.; Visualization, A.B.; Supervision, J.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding

**Data Availability Statement:** The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author(s).

**Conflicts of Interest:** The authors declare no conflict of interest.

## Glossary

Glossary of Core Terms in Phishing Detection and AI-Based Text Analysis.

| Term | Definition |
|---|---|
| **Cybersecurity Concepts** | |
| Phishing | Deceptive communication aimed at tricking users into revealing sensitive information. |
| Spear Phishing | Targeted phishing aimed at a specific individual or organisation. |
| Watermarking | Inserting detectable patterns into AI-generated text for identification. |
| Attack Vector | Method used by attackers to gain unauthorised access to systems. |
| **Machine Learning Concepts** | |
| Classification | A supervised learning task that assigns input data to predefined categories. |
| K-Means Clustering | An unsupervised algorithm that partitions into clusters based on feature similarity. |
| TF-IDF | Statistical method for weighting terms based on their frequency and distinctiveness across documents. |
| Logistic Regression | A linear classification algorithm for binary outcomes. |
| PCA (Principal Component Analysis) | A technique for reducing dimensionality and visualising high-dimensional data. |
| Prompt Engineering | The process of designing inputs to influence LLM outputs effectively. |
| **Adversarial AI Tools** | |
| WormGPT | Malicious LLM used for generating phishing or harmful content. |
| FraudGPT | A dark web LLM for cybercrime, including impersonation and fraud. |

## References


1. Stokel-Walker, C. (Ed.) *The AI Revolution: What the New Age of Artificial Intelligence Means for Humanity*; New Scientist Essential Guide No. 23; New Scientist: London, UK, 2024.
2. Kucharavy, A. From Deep Neural Language Models to LLMs. In *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*; Springer Nature: Cham, Switzerland, 2024; pp. 3–17.
3. Erbschloe, M. *Social Engineering: Hacking Systems, Nations, and Societies*, 1st ed.; CRC Press: Boca Raton, FL, USA, 2019. https://doi.org/10.1201/9780429322143.
4. Jakobsson, M.; Myers, S. *Phishing and Countermeasures*, 1st ed.; Wiley-Interscience: Oxford, UK, 2006.
5. Nunes, V. *The Cyber Skill Gap: How to Become a Highly Paid and Sought After Information Security Specialist*; Self-Published, 2017; Available online: https://www.amazon.com/Cyber-Skill-Gap-Information-Specialist-ebook/dp/B06XJD1W7N (accessed on 10 April 2024).
6. Gallagher, S.K.; Ratchford, J.; Brooks, T.; Brown, B.P.; Heim, E.; Nichols, W.R.; Mcmillan, S.; Rallapalli, S.; Smith, C.J.; VanHoudnos, N.; et al. Assessing LLMs for High Stakes Applications. In Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice, Lisbon Portugal, 14–20 April 2024; pp. 103–105.
7. Elkhatat, A.M.; Elsaid, K.; Almeer, S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *Int. J. Educ. Integr.* **2023**, *19*, 17..
8. Greco, F.; Desolda, G.; Esposito, A.; Carelli, A. David versus Goliath: Can Machine Learning Detect LLM-Generated Text? A Case Study in the Detection of Phishing Emails. In Proceedings of the ITASEC 2024: The Italian Conference on CyberSecurity, Salerno, Italy, 9–11 April 2024.
9. Sadasivan, V.S.; Kumar, A.; Balasubramanian, S.; Wang, W.; Feizi, S. Can AI-Generated Text be Reliably Detected? *arXiv* **2024**, arXiv:2303.11156.
10. Ghosal, S.S.; Chakraborty, S.; Geiping, J.; Huang, F.; Manocha, D.; Bedi, A.S. Towards Possibilities & Impossibilities of AI-generated Text Detection: A Survey. *arXiv* **2023**, arXiv:2310.15264.

11. Tao, Z.; Li, Z.; Xi, D.; Xu, W. CUDRT: Benchmarking the Detection of Human vs. Large Language Models Generated Texts. *arXiv* **2024**, arXiv:2406.09056v1.
12. Cialdini, R.B. *Influence: The Psychology of Persuasion*, rev., ed.; Harper Business: New York, NY, USA, 2007.
13. Gutmann, A. An Analysis of Computer Systems for the Secure Creation and Verification of User Instructions. Ph.D. Thesis, University College London, London, UK, 2020.
14. Hazell, J. Spear Phishing with Large Language Models. *arXiv* **2023**, arXiv:2305.06972.
15. Chapple, M.; Seidl, D. *Cyberwarfare: Information Operations in a Connected World*; Jones & Bartlett Learning: Burlington, MA, USA, 2021.
16. Firdhous, M.F.M.; Elbreiki, W.; Abdullahi, I.; Sudantha, B.H.; Budiarto, R. Wormgpt: a large language model chatbot for criminals. In Proceedings of the 24th International Arab Conference on Information Technology (ACIT), Ajman, United Arab Emirates, 6–8 December 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–26.
17. Da Silva Gameiro, H. LLM Detectors. In *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*; Springer Nature: Cham, Switzerland, 2024; pp. 197–204.
18. Gehrmann, S.; Strobelt, H.; Rush, A.M. GLTR: Statistical Detection and Visualization of Generated Text. *arXiv* **2019**, arXiv:1906.04043.
19. Guo, B.; Zhang, X.; Wang, Z.; Jiang, M.; Nie, J.; Ding, Y.; Yue, J.; Wu, Y. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. *arXiv* **2023**, arXiv:2301.07597.
20. Bateman, J. *Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios*; Carnegie Endowment for International Peace: Washington, DC, USA, 2020.
21. Weber-Wulff, D.; Anohina-Naumeca, A.; Bjelobaba, S.; Foltýnek, T.; Guerrero-Dib, J.; Popoola, O.; Šigut, P.; Waddington, L. Testing of detection tools for AI-generated text. *Int. J. Educ. Integr.* **2023**, *19*, 26 .
22. Bhattacharjee, A.; Liu, H. Fighting Fire with Fire: Can ChatGPT Detect AI-Generated Text? *ACM SIGKDD Explorations Newsletter* **2024**, *25*, 14–21, ACM New York, USA.
23. Kirchner, D.; Reiter, A.; Scholkopf, B. Watermarking LLM-Generated Texts via Synonym Substitution: Preserving Semantics while Enabling Detection. *arXiv* **2023**, arXiv:2306.04634.
24. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008.
25. Demir, M.; Yildirim, M. Efficient Hybrid Movie Recommendation System Framework Based on A Sequential Model. 2023. Available online: https://www.researchgate.net/publication/372490095 (accessed on 21 April 2025).
26. Jurafsky, D.; Martin, J.H. *Speech and Language Processing*, 3rd ed.; Pearson: London, UL, 2021; p. 199.
27. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
28. Trappenberg, T.P. *Fundamentals of machine learning*; Oxford University Press: Oxford, UK, 2019.
29. *ISO/IEC 42001:2023*; Information Technology—Artificial Intelligence—Management System. ISO: Geneva, Switzerland, 2023.