Journal of Artificial Intelligence and AI Ethics

SCIVOLVE

Research Article

Open Access Full Text Article

Assessing the Detectability of AI-Generated Phishing Emails by Modern Email Filters

Ruth Imanria Itua[1], Aaron Ogochukwu Okolo[2], Ifelunwa Ada Ikuni[3], Chidinma Anumaka[4], Sebastian Obeta[5*]

[1]*Security Operations Analyst; Epaton LTD.*
[2]*Senior Consultant, Henderson Drake Ltd.*
[3]*Cloud Adoption Manager, Oracle Corporation, UK.*
[4]*University of Nottingham.*
[5]*Cambridge University UK.*

***Correspondence:**

Sebastian Obeta
Digievolve UK.

Received: January 18, 2026; **Published:** January 30, 2026.

How to cite this article:

Itua, R.I., Okolo, A.O., Ikuni, I.A., Anumaka, C., and Obeta, S. (2026), 'Assessing the Detectability of AI-Generated Phishing Emails by Modern Email Filters', *Journal of Artificial Intelligence and AI Ethics*, vol. 1, no. 1, pp. 1–15.

Abstract

The emergence of large language models (LLMs) has significantly altered the phishing threat landscape by enabling the automated generation of linguistically fluent and contextually convincing phishing emails. While prior studies demonstrate the effectiveness of AI-generated phishing and the vulnerability of experimental classifiers, the real-world performance of widely deployed email filtering systems remains insufficiently understood. This study addresses this gap through an empirical evaluation of modern email filters exposed to AI-generated phishing content.

A controlled dataset of 100 phishing emails was generated across multiple attack categories using contemporary LLMs, including ChatGPT, Claude, Gemini, Meta AI, and Qwen 2.5, and evaluated against commonly used email filtering systems such as Gmail, Outlook, Yahoo Mail, Proton Mail, and Spam Assassin. Detection outcomes were analysed using quantitative performance metrics, rule activation analysis, and statistical testing to examine the influence of filtering system, phishing category, and language model on detectability.

The results reveal pervasive detection failures across most evaluated systems, with high false negative rates observed. Statistical analysis shows that detection outcomes are significantly associated with the filtering system employed, but not with the phishing category or the LLM used. These findings demonstrate systemic limitations in current email filtering architectures and highlight the need for adaptive, intent-aware defences against AI-enabled phishing.

Kaywords: AI-generated phishing; Large Language Models; Email filtering systems; Phishing detection; Email security; Cybersecurity

1. Introduction

The rapid development of digital infrastructure globally has placed cybersecurity at the forefront of concerns for individuals, organisations, and governments (Kumar and Patel, 2025; Kumar et al., 2024; Vadisetty and Polamarasetti, 2024). Within this dynamic landscape, phishing remains the most pervasive, widely used, and sophisticated threat (Abdolrazzagh-Nezhad and Langarib, 2025; Jaiswal et al., 2024). Defined as a social engineering tactic, phishing involves manipulating targets to acquire sensitive information or persuade them to undertake harmful actions (Alahmed et al., 2024; Jaiswal et al., 2024). Attackers exploit fundamental human psychological and behavioural weaknesses, such as favouring trust over scepticism and prioritising urgency, to gain access to systems (Heiding et al., 2023, 2024; Qi et al., 2024).

Phishing serves as the predominant entry point for the majority of sophisticated cyber-attacks, leading to severe consequences, including substantial financial losses, data breaches, and

reputational damage worldwide (Heiding et al., 2024; Jaiswal et al., 2024; Vadisetty and Polamarasetti, 2024). Despite immense organisational investments in traditional detection measures and employee training, phishing remains a persistent nuisance (Vishwanath, 2022).

The evolution of generative Artificial Intelligence (AI) and Large Language Models (LLMs), such as GPT-4 and Claude, has ushered in a new era of cyber threats by fundamentally altering the economics and scalability of phishing campaigns (Heiding et al., 2023; Jaiswal et al., 2024). Threat actors are now actively leveraging the Natural Language Generation (NLG) capabilities of these tools for malicious purposes, resulting in phishing attempts that are highly sophisticated and difficult to detect using conventional methods (Alahmed et al., 2024; Kumar and Patel, 2025). LLMs act as a significant force multiplier, drastically reducing the barrier to entry and the skill requirements for cybercriminals, enabling low-skilled attackers to generate convincing emails at scale (Hazell, 2023; Humphreys et al., 2024). Critically, LLMs allow attackers to create content that is highly personalised and contextually relevant, often based on just a few easily collected data points about the recipient, sometimes even mimicking the linguistic style of an acquaintance (Kumar et al., 2024). This enhancement significantly increases the incentives for launching spear phishing attacks, rendering personalised attempts far cheaper than traditional spear phishing, sometimes approaching the cost of arbitrary mass-scale emails (Kumar et al., 2024). The impact is quantifiable: since the introduction of advanced models like ChatGPT in late 2022, the volume of malicious emails has skyrocketed, with reported increases as high as 4151% (Kumar et al., 2024). Empirical studies validate the enhanced efficacy of this threat, showing that AI-generated phishing emails generally achieve significantly higher click-through rates (30–44%) compared to traditional, non-personalised emails (19–28%) (Dash and Sharma, 2023; Heiding et al., 2024).

The primary cybersecurity challenge presented by the rise of LLM-generated phishing lies in its inherent capacity to circumvent established security defences. Historically, mail servers often relied on identifying identical or nearly identical emails to categorise them as spam and filter them out (Eze and Shamir, 2024). Generative AI bypasses this mechanism entirely by creating a large number of unique email messages automatically, ensuring that each generated email reads as if it were written by a person, making it difficult for spam detection systems to identify repetitive messages sent from external sources (Hazell, 2023). Furthermore, AI-generated emails are typically polished, grammatically correct, and lack the obvious flaws and overtly suspicious language often characteristic of manually crafted phishing attempts (Kumar et al., 2024).

Eze and Shamir (2024) argue that AI-generated phishing emails are stylistically different from manually-generated phishing scam emails and regular emails (Hazell, 2023). Their research, based on examining numerical text descriptors, established identifiable writing differences in AI-generated content, such as a higher frequency of verbs and pronouns, significantly longer average word lengths, and a tendency to express more positive sentiments (Eze and Shamir, 2024). The implication of this view encourages automatic identification tools, such as machine learning systems, can classify AI-generated emails with high accuracy (up to 99-100% in controlled tests against older datasets), provided they are specifically trained on corpora of AI-generated emails (Qi et al., 2024). This constitutes a significant strength, establishing that the threat is not inherently undetectable, but rather relies on distinct stylistic metadata produced by generative models (Hazell, 2023; Qi et al., 2024).

The SpearBot study demonstrated that traditional Machine Learning (ML) defenders (e.g., XGBoost) registered extremely low accuracy (e.g., just 21.70%), and even sophisticated Pre-trained Language Model (PLM) defenders experienced performance crashes, with accuracy plummeting to as low as 1.00% to 3.00% when tested against new, optimised AI content (Qi et al., 2024). Similarly, Heiding et al. (2024) found that while LLMs show promise for detection, models often surpass human detection rates primarily when they are specifically "primed for suspicion," implying that general, untrained detection systems struggle (Heiding et al., 2023). The core implication synthesized from this contradiction is alarming: while the findings of Eze and Shamir (2024) offer a route to future resilience, the empirical results from SpearBot (2023/2024) confirm that current systems, largely utilizing legacy models trained on outdated datasets, are fundamentally inadequate and overfit to previous attack patterns, thus failing to repel this new, personalized generation of threats reliably (Qi et al., 2024). The LLM-generated content thus poses a critical challenge because it evades traditional vertical feature-space analysis by creating novel, unique, and compelling social engineering cues (Abuadbba et al., 2022).

The convergence of AI's capability to generate highly deceptive content and the demonstrated failure modes of current machine-based defences highlights a critical and urgent research gap. Despite laboratory studies showing that laboratory-trained ML and PLM defenders exhibit low accuracy against tailored AI-generated content (Qi et al., 2024). A pervasive lack of empirical testing remains, focused directly on widely deployed, commercial email filtering solutions (e.g., anti-phishing tools and enterprise-grade mail gateways) against a purpose-built corpus of LLM-generated phishing emails. Traditional detection techniques often rely on predefined datasets or heuristic rules, which struggle with zero-day phishing attacks and evolving tactics (Abdolrazzagh-Nezhad and Langarib, 2025; Abuadbba et al., 2022). Failure to proactively quantify the efficacy of these commercial systems against AI-driven campaigns represents a critical unknown risk, necessitating urgent, real-world validation to inform organisational cyber resilience efforts. Therefore, this study, titled "Assessing the Detectability of AI-Generated Phishing Emails by Modern Email Filters," directly addresses this gap by subjecting commercially utilised detection infrastructure to this novel, sophisticated threat, thereby quantifying the measurable limitations of current security postures against exponentially growing AI-powered cyber threats (Chien and Khethavath, 2023; Jaiswal et al., 2024).

1.1 Research Problem

The central research problem addressed in this study is the potential inadequacy of contemporary email filtering systems in detecting phishing emails generated by artificial intelligence (AI), coupled with a notable lack of empirical evidence quantifying the extent and severity of this vulnerability. As generative AI systems increasingly produce highly fluent and contextually convincing text, traditional assumptions underpinning phishing detection may no longer hold.

More specifically, several interrelated challenges contribute to this problem. First, AI-generated phishing emails often lack conventional linguistic and structural indicators that existing filters rely upon for detection. Second, most modern email filters are trained mainly on phishing emails written by humans, which

can make them less effective at detecting AI-generated ones.

Third, there are no standard datasets or testing methods designed to measure how well filters detect AI-generated phishing, which makes it hard to fairly compare different solutions. Finally, as a result of these limitations, organisations may be unknowingly exposed to a novel and increasingly sophisticated class of phishing threats.

This gap in empirical understanding poses a significant risk for cybersecurity practitioners, email service providers, and organisations that rely on automated email filtering systems as their primary defence mechanism. Addressing this research problem is therefore critical for assessing current defensive capabilities and informing the development of more robust detection strategies.

1.2 Research Objectives

This study aims to empirically evaluate the effectiveness of contemporary email filtering systems in detecting phishing emails generated by large language models (LLMs). Specifically, the study seeks to: (i) generate a representative dataset of AI-generated phishing emails using multiple LLMs; (ii) assess the detection performance of widely used email filtering systems, including Gmail, Outlook, Yahoo, Proton.me and Spam Assassin; and (iii) analyse false negative cases to identify attack-type patterns and linguistic or structural characteristics associated with detection failures.

1.3 Research Questions

The study is guided by the following research questions:

1. How effectively do modern email filtering systems detect phishing emails generated by large language models?
2. Are certain categories of AI-generated phishing attacks more likely to evade detection than others?
3. Do phishing emails generated by different large language models differ in their detectability?
4. Which linguistic or structural feature(s) of AI-generated phishing emails contribute to false negative detections?

1.4 Justification and Significance of the Study

The increasing use of generative artificial intelligence has introduced a new class of phishing threats that may challenge existing email filtering mechanisms. Despite this shift, there is limited empirical evidence evaluating the performance of real-world email filters against AI-generated phishing content. Given that email filtering systems constitute a primary defensive layer for individuals and organisations, assessing their effectiveness in this emerging threat context is both timely and necessary.

This study contributes to the cybersecurity literature by providing a systematic, empirical evaluation of widely deployed email filtering systems using AI-generated phishing emails. The findings enhance understanding of how generative language models affect phishing detectability and offer insights relevant to the development of more resilient and adaptive filtering strategies.

1.5 Scope of the Study

This study is limited to the content-based detection of AI-generated phishing emails. It evaluates phishing emails generated using selected large language models and examines the detection performance of Gmail, Outlook, and Spam Assassin. The analysis focuses exclusively on the textual content of emails and excludes attachments, embedded links, and malware payloads.

Non-email phishing vectors (e.g., SMS, voice, and social media), real-world phishing campaigns, proprietary enterprise-grade filtering systems, and multi-stage phishing attacks are outside the scope of this study.

2. Evolution of Phishing

The trajectory of phishing, an enduring form of cybercrime involving the fraudulent extraction of sensitive information by deception, reflects a continuous arms race between motivated attackers and evolving security measures (Ghazi-Tehrani and Pontell, 2021; Osamor et al., 2025). Historically, phishing attacks have undergone a fundamental methodological shift, transitioning from simple, template-based mass campaigns to highly sophisticated, targeted operations that now leverage advanced automation technologies (Ghazi-Tehrani and Pontell, 2021; Liesnaia and Malakhov, 2023; Osamor et al., 2025). This evolution establishes a clear historical rationale for why large language model (LLM)-powered phishing represents the definitive next major step in the cyber threat landscape.

The earliest era of phishing, often dated between 1995 and 2005, was characterised by "wide" attacks relying primarily on basic email spoofing and generic mass mailings (Osamor et al., 2025). These campaigns involved sending identical messages to thousands or even hundreds of thousands of recipients, famously including the grammatically incorrect "Nigerian prince" scams (Aldam, 2025; Eze and Shamir, 2024). Attackers employed basic social engineering (SE) tactics that relied more on the sheer volume of attempts than on sophistication, aiming for a low but profitable response rate, sometimes hovering around 0.1% (Ghazi-Tehrani and Pontell, 2021; Osamor et al., 2025). However, this brute-force approach was gradually rendered ineffective due to the widespread adoption of standardized email security measures and robust spam filters (Ghazi-Tehrani and Pontell, 2021).

In response to improved technological defences, phishing evolved rapidly into more specialised forms, leading to the period of specialisation and professionalisation (Liesnaia and Malakhov, 2023). This phase, covering roughly 2005 to 2015 and continuing today, is defined by "narrow" attacks such as spear phishing, whaling, and Business E-mail Compromise (BEC). Ghazi-Tehrani and Pontell (2021) observed that while wide-net phishing remained the most common form, spear phishing grew in popularity, as motivated offenders adapted to follow the money and circumvent technological countermeasures that had largely contained simple bulk attacks (Ghazi-Tehrani and Pontell, 2021). Conversely, Osamor et al. (2020) highlighted the dramatic increase in efficacy achieved by this shift, noting that early mass mailings achieved meager success rates of approximately 0.1%, whereas spear phishing, which involved extensive research and incorporating personal details, often achieved success rates exceeding 50% (Osamor et al., 2025). This comparison reveals a critical implication: the effectiveness of traditional security technology against widespread attacks forced attackers to pivot to individualised social engineering techniques, escalating the resource investment per attack but yielding significantly greater returns by compromising high-value targets. This necessary adaptation emphasised the human element, turning the challenge of detection into a matter of psychological and contextual nuance rather than signature matching (Ghazi-Tehrani and Pontell, 2021).

This historical context provides the necessary foundation for

understanding the rise of AI-powered and LLM-driven phishing, which marks the next logical and terrifying stage of this evolution (Eze and Shamir, 2024). The incorporation of AI revolutionises the capabilities of cybercriminals, enabling unprecedented scale, personalisation, and effectiveness (Aldam, 2025). Aldam (2025) argues that AI-powered attacks have rapidly evolved, utilising advanced personalisation, multi-channel deception, and deepfake technology to definitively surpass human-crafted scams in both scale and effectiveness (Aldam, 2025). This claim is supported empirically by research showing that AI-generated phishing campaigns outperformed those created by human red team experts by 24% by March 2025, marking a significant milestone in social engineering capabilities (Aldam, 2025). This viewpoint emphasises the outcome AI's speed and superior efficacy in crafting attacks (Aldam, 2025).

In contrast, Eze and Shamir (2024) provide critical insight into the underlying mechanism by which LLMs achieve this superiority, arguing that generative AI is primarily used to bypass one of the last major technological hurdles to scaling personalised attacks: detection systems relying on identifying identical or repetitive email messages (Eze and Shamir, 2024).

Instead of a single email format sent to many recipients, generative AI can be used to send each potential victim a unique email, making identification more difficult for cybersecurity systems. Their study found that AI-generated emails possess specific style elements such as having a larger average word length, using more diverse vocabulary, and expressing more positive sentiments that make them measurably different from manually written human scam emails. The significance of this pivot cannot be overstated: the resource-intensive, manual personalisation that characterised successful spear phishing (Phase 2) is now automated and scalable, resulting in a dramatic explosion in phishing volume since the widespread adoption of generative AI tools like ChatGPT in 2022 (a reported 4,151% increase in volume) (Kumar et al., 2024). Therefore, the historical evolution of phishing, moving from bulk simplicity to targeted complexity, culminates naturally in the LLM-powered approach, which marries the scale of early wide attacks with the quality and personalisation of spear phishing, fundamentally altering the calculus for both offenders and defenders (Aldam, 2025; Osamor et al., 2025).

2.1 Large Language Models and Phishing Content Generation

The integration of Large Language Models (LLMs) into publicly accessible platforms has fundamentally altered the threat landscape of cybercrime, particularly concerning phishing campaigns (Ferrara, 2024). Traditional phishing attempts were often manually constructed and betrayed their malicious intent through linguistic deficiencies such as poor grammar, misspellings, and inconsistent formatting (Kulal et al., 2025). However, the advent of LLMs like Generative Pre-training Transformer (GPT), Claude, and Gemini has enabled adversaries to overcome these barriers, leading to the creation of highly coherent and contextually relevant phishing messages that closely mimic legitimate organisational or personal communication (Pang et al., 2025).

The ability of LLMs to generate high-quality text humanly, is rooted in their massive parameter sizes and training on vast amounts of data, enabling them to execute complex language-related tasks accurately (Xu and Parhi, 2025). Critically, this capability results in LLM-generated phishing content that is grammatically sound and linguistically natural, often lacking the errors typical of past, human-written phishing efforts (Kulal et al., 2025). This linguistic perfection distinguishes machine-written phishing attacks from their predecessors and makes them considerably more difficult for conventional email filters and traditional detection mechanisms to flag (Pang et al., 2025). Olea et al. (2025) empirically demonstrate this shift, finding that LLM-generated phishing emails are consistently labelled as benign (real) more frequently than human-generated phishing emails (Olea et al., 2025). This effect is theorised to be due to the decreased rate of spelling and grammatical errors, as well as the propensity of LLMs to include professional formalities and boilerplate text that dispels the suspicion typically associated with low-effort phishing adversaries (Olea et al., 2025).

The sophistication of machine-generated content leads to differences that are quantitative and qualitative when compared to past phishing attempts, particularly regarding scalability and targeting precision. The generalised deployment of LLMs allows malicious actors to automate the process of generating deceptive content, enabling them to launch campaigns at a much larger scale than previously possible, as highlighted by multiple researchers (Koplin, 2023). Furthermore, LLMs facilitate the creation of highly personalised and tailored spear-phishing emails (Qi et al., 2024). Researchers have developed frameworks, such as SpearBot, that leverage LLMs (like GPT-4) in a generative-critique paradigm to craft sophisticated spear-phishing messages aimed at specific individuals or entities within an organisation (Qi et al., 2024). This process often requires the use of specialised jailbreak prompts to circumvent the safety alignments (such as Reinforcement Learning from Human Feedback) deliberately embedded in LLMs to prevent the generation of harmful content (Pang et al., 2025; Qi et al., 2024). The generation process within frameworks like SpearBot may involve multiple LLM instances acting as critics, refining the generated email based on feedback until it is no longer recognised as malicious, thus enhancing its deceptive quality (Qi et al., 2024).

These capabilities elevate the discussion to critical ethical, adversarial, and misuse concerns surrounding LLM technology, which is fundamentally considered dual-use in nature (Koplin, 2023). Koplin (2023) argues that LLM text generation should be conceptualised as a dual-use technology, capable of both immense benefit and profound harm by undermining individual autonomy and democratic institutions through the proliferation of disinformation at scale (Koplin, 2023). Koplin notes that carelessly used LLMs can result in a flood of low-quality, inaccurate content, characterising the output as a "fluent spouter of bullshit" that distorts the public's understanding of the world. Ferrara (2024), however, focuses on the practical taxonomy of GenAI abuse, detailing specific types of harm such as financial loss, information manipulation, and societal damage caused by the scaled creation of targeted scams and malicious content (Ferrara, 2024). While Koplin provides the crucial philosophical framework for intervention by weighing values like security against scientific openness (Koplin, 2023), Ferrara outlines the specific, manifest cyber threats that require immediate technical mitigation, including the use of LLMs to generate malware and bypass traditional security measures (Ferrara, 2024).

Moreover, the integration of LLMs into applications introduces novel adversarial threats beyond direct malicious prompting (Greshake et al., 2023). Specifically, the blur line between data and instructions enables Indirect Prompt Injection (IPI), where malicious prompts are stealthily embedded into retrieved data (like a web page or document) and executed by the LLM application without the user's knowledge (Greshake et al., 2023). This technique allows adversaries to gain remote control over

the model, facilitating data theft, persistent compromise, and content manipulation (Greshake et al., 2023). This distinction is critical: while IPI exploits architectural vulnerabilities, the inherent linguistic fluency of LLMs simultaneously enables the creation of highly effective phishing payloads that evade detection based on surface-level textual flaws (Olea et al., 2025). Therefore, confronting LLM-enabled phishing requires a multi-faceted approach that addresses both the philosophical dual-use dilemma and the immediate technical challenges posed by sophisticated, automated, and architecturally manipulative attacks (Valencia, 2024).

2.2 Email Filtering Mechanics

The proliferation of unwanted bulk email (UBE), commonly referred to as spam, and malicious phishing emails necessitates a robust and constantly evolving technological defence infrastructure (Anitha et al., 2021). (Awad and ELseuofi, 2011). The core objective of modern email filtering systems is to accurately categorise incoming messages as benign ('ham') or malicious ('spam') (Jeeva and Khan, 2023). These mechanics historically rely on a multi-layered approach, drawing upon conventional detection signals rooted in observable characteristics and learned patterns (Anitha et al., 2021). However, this established reliance on feature engineering tailored for human-written content lays the foundation for vulnerability against sophisticated, generative threats.

The mechanics of email filtering employ several established techniques to assess an email across its content, metadata, and sender behaviour. Early approaches relied heavily on knowledge engineering, which involved manually specifying rules to categorise emails (Awad and ELseuofi, 2011). This rule-based methodology, epitomised by systems like Spam Assassin, detects threats by searching for spam-like patterns based on content-matching rules, each assigned a numerical score (Vu et al., 2015). Anitha et al. (2021) specify that many spam filters utilise blacklists, Bayesian review, keyword matching, and mail header analysis to recognise incoming messages (Anitha et al., 2021). Conversely, Awad and ELseuofi (2011) argue that the knowledge engineering approach is fundamentally problematic because its rules must be constantly updated and maintained, deeming the process a waste of time and inconvenient for most users. This limitation led to the widespread adoption of machine learning (ML) systems, which avoid the need for manual rule specification by learning classification rules directly from pre-classified training samples (Jeeva and Khan, 2023).

Current classification mechanisms incorporate both structural and content-based feature analysis. Header and domain analysis involves scrutinising message attributes like the sender's address or IP address, a method essential for establishing reputation (Anitha et al., 2021). For instance, Karim et al. (2020) proposed evaluating anti-spam frameworks heavily reliant on domain and email headers (Murti and Naveen, 2023). Regarding content analysis, Bayesian classification remains a foundational statistical method (Vu et al., 2015). This technique calculates the probability of certain words occurring in spam versus legitimate emails to combine individual token statistics into an overall score, making the filtering decision based on a defined threshold (Awad and ELseuofi, 2011). Raza et al. (2022) affirm this focus by highlighting that most supervised machine learning research in this field is concentrated on content features, particularly the bag-of-words (BOW) model and body text (Murti and Naveen, 2023). Beyond content and headers, systems also attempt to assess sender behaviour; Tang et al. (2008), for example, proposed extracting email sender behaviour data based on global sending distribution to assign a trust value to each IP address. Yoo et al. (2009) developed a system that analysed personal social networks to capture user groups and model personal priorities over email messages (Yoo et al., 2009).

The evolution from simple rule-based systems to complex ML models, including Logistic Regression, Random Forest, and Artificial Neural Networks (ANNs), demonstrates the technological arms race against spammers (Alsuwit et al., 2024). The success of these techniques is evident in studies showing high accuracy rates; Awad and ELseuofi (2011), using the Spam Assassin dataset, found Naïve Bayes achieved 99.46% accuracy. However, Alsuwit et al. (2024), testing a range of methods on a combined corpus, found that an ANN model demonstrated slightly superior performance with 98% accuracy compared to the 97% achieved by traditional ML algorithms like Logistic Regression and Naïve Bayes (Alsuwit et al., 2024). This highlights the continual competition among different classification models to achieve robustness.

Despite the sophistication achieved by models trained on existing spam datasets, a critical vulnerability persists. Spammers are constantly adapting their tactics to bypass detection (Alsuwit et al., 2024; Awad and ELseuofi, 2011), sometimes using seemingly legitimate email addresses or incorporating personalised details to evade generic filters (Alsuwit et al., 2024). This constant evolution necessitates countermeasures against adversarial methods specifically employed to evade classification techniques. The core limitation is that the effectiveness of current filtering systems is derived from patterns learned from historical, primarily human-generated, malicious correspondence (Jeeva and Khan, 2023). Future work ensures identifying the need to explore detection against emails processed through tools designed to circumvent standard email servers and algorithms, such as email warming tools intended to establish a positive sending reputation (Alsuwit et al., 2024). Given that existing filters rely on detecting proxies for human-driven malice, such as poor grammar, predictable keywords, or structural inconsistencies identified during feature extraction (Awad and ELseuofi, 2011), they are not inherently optimised to address the emerging threat of phishing content created by advanced Generative AI. Consequently, the reliance of current filter mechanics on traditional detection signals supports the hypothesis that AI-generated phishing, capable of bypassing these established linguistic and behavioural cues, may easily circumvent existing defence layers.

2.3 Detection Challenges with AI-Generated Emails

The emergence of advanced text generation models, particularly Large Language Models (LLMs), fundamentally alters the threat landscape for email security by enabling the creation of phishing emails that significantly reduce or eliminate the traditional markers utilised by automated detection systems, thereby introducing new blind spots in detection capabilities (Tang and Li, 2025). Traditional spam and phishing filters, often rooted in historical machine learning approaches, primarily relied on identifying low-quality textual indicators, such as grammatical errors, misspellings (typos), and awkward, inconsistent structural patterns, which are now largely bypassed by sophisticated AI-generated content (Josten and Weis, 2024).

AI-generated text inherently avoids typical phishing markers because these generative models are proficient at producing fluent and coherent language that maintains high semantic fidelity to legitimate communication (Du et al., 2024; Josten and Weis, 2024).

Whereas traditional adversarial attacks against NLP systems often resulted in visually perceptible perturbations, grammatical errors, or high perplexity, AI-generated text is specifically crafted to overcome these weaknesses, making the malicious data stealthier (Du et al., 2024). For instance, Boucher et al. (2021) demonstrate that earlier text-based attacks struggled to maintain semantic meaning and indistinguishability, forcing them to rely on noticeable artefacts like single-character spelling mistakes or paraphrasing that changed the meaning enough to be detected (Boucher et al., 2022). In contrast, Du et al. (2023) highlight that text generation models can synthesise fluent and content-relevant text that humans often cannot distinguish from authentic text, ensuring both fluency and preservation of original semantics in poisoned data (Du et al., 2024).

The enhanced quality of AI-generated content introduces a critical challenge: the homogeneity and "clean grammar" reduce the efficacy of traditional machine learning (ML) detection models (Josten and Weis, 2024). Phishing emails created by LLMs exhibit superior linguistic structure and polish, effectively masking the malicious intent behind a facade of normalcy (Josten and Weis, 2024). Josten and Weis (2024) demonstrated this effect directly, showing that a widely adopted Bayesian spam filter misclassified up to 73.7% of LLM-modified spam emails as legitimate, while a simpler dictionary-replacement attack achieved only a 0.4% success rate, underscoring the unique vulnerability posed by linguistic sophistication. This challenge extends beyond simple filters; Boucher et al. (2021) observe that even deep neural network pipelines often use tokenisers and sub-word encoding that are unlikely to handle imperceptible perturbations gracefully, an effect amplified by the perfectly constructed grammar of AI-generated text, which presents itself as "clean" input (Boucher et al., 2022). The implication is that filters trained on identifying "bad words" or obvious errors struggle to establish a decision boundary against inputs where the adversarial nature is hidden in the quality of the language itself (Pawar and Patil, 2015).

Furthermore, AI-generated emails achieve high semantic similarity to legitimate business emails, making them particularly effective in spear-phishing and bypassing content-based filters (Du et al., 2024). Josten and Weis (2024) confirmed that LLM-modified spam preserved a high mean cosine similarity (approximately 0.8) to the original text, confirming that the core message and likely the malicious objective remained intact despite evading detection. This is especially significant because the goal of adversarial attacks in this context is to manipulate the input without changing the true class label (e.g., still being spam, but classified as ham) (Zhang et al., 2020). While traditional "Good Word Attacks" manipulated statistical models by adding benign words to skew message statistics (Lowd and Meek, 2005), AI-generated emails take this further by automatically constructing an entire message that perfectly mimics the linguistic characteristics of legitimate communication, reducing the need for crude word insertions and creating camouflaged spam that blends elements of both spam and legitimate content (Tang and Li, 2025). Dionysiou and Athanasopoulos (2021) reinforce this concept by demonstrating that adversarial text produced using visually similar Unicode characters has little impact on human understanding and the original text's semantics, achieving high success rates because the text is easily interpretable by humans while failing machine classifications (Dionysiou and Athanasopoulos, 2021). The difficulty of detection lies in the fact that the text is clear and plausible syntactically and semantically, compelling LLM-based detection systems relying on complex semantic understanding, a dependency that adversarial attacks exploit (Tang and Li, 2025; Ozioma et al., 2026). Therefore, the inherent quality and coherence provided by AI tools remove the noisy, easily filtered signals, demanding that detection systems evolve from policing grammar to discerning hidden intent, a complex and challenging task.

2.4 Research Gap: Real-World Detectability of AI-Generated Phishing

The literature unequivocally establishes that AI-generated phishing is an operational, rapidly escalating threat. LLM-generated emails are linguistically polished, semantically coherent, and more persuasive than traditional phishing, achieving significantly higher click-through and success rates (Dash and Sharma, 2023; Heiding et al., 2024; Olea et al., 2025). At the same time, these emails deliberately evade conventional detection cues such as grammatical errors, keyword anomalies, and repetitive templates (Hazell, 2023; Josten and Weis, 2024; Du et al., 2024). Experimental evidence further shows that both traditional ML models and advanced pretrained language-model defenders suffer severe performance collapse when confronted with optimised AI-generated phishing content (Qi et al., 2024; Heiding et al., 2023). Together, these findings confirm that the threat is real, measurable, and already exceeding the capabilities of many existing detection mechanisms.

However, a critical empirical gap persists between academic detection studies and real-world defensive infrastructure. Current research overwhelmingly evaluates laboratory-trained classifiers, custom ML pipelines, or experimental LLM-based detectors (Eze and Shamir, 2024; Qi et al., 2024), while widely deployed commercial email filtering systems remain largely unexamined. These operational filters continue to rely on historically derived datasets, heuristic rules, and human-crafted phishing indicators (Anitha et al., 2021; Jeeva and Khan, 2023; Abdolrazzagh-Nezhad and Langarib, 2025), despite clear evidence that such signals are systematically neutralised by generative AI (Abuadbba et al., 2022; Josten and Weis, 2024).

Critically, no study identified in the current literature provides a systematic, real-world evaluation of modern commercial email filters against a controlled corpus of AI-generated phishing emails. While prior work demonstrates that AI-modified spam can bypass Bayesian and ML-based detectors at high rates (Josten and Weis, 2024; Qi et al., 2024), these findings stop short of validating the resilience of production-grade filtering systems that currently serve as the primary defence layer for organisations and individuals.

This unresolved gap represents a non-trivial operational risk, as organisations may be unknowingly relying on defences that have not been empirically validated against the dominant emerging threat vector (Aldam, 2025; Kumar et al., 2024). Accordingly, this study directly addresses this deficiency by providing a targeted, real-world assessment of modern email filtering systems under exposure to AI-generated phishing content, thereby delivering empirical evidence where the current literature remains silent.

3. Methodology

This chapter describes the methodological procedures adopted to evaluate the detectability of AI-generated phishing emails by modern email filtering systems. The methodology follows a structured, empirical approach composed of dataset generation, pre-processing, testbed configuration, experimental execution, and quantitative and qualitative analysis. Emphasis is placed on reproducibility, ethical compliance, and methodological rigour.

3.1 Research Design

This study employs an experimental comparative research design, enabling systematic evaluation of multiple email filtering systems under controlled conditions. The independent variable is the source and type of phishing emails, generated using multiple large language models (LLMs). The dependent variable is the detection outcome (e.g., detected vs. undetected) produced by each filtering system.

The experimental design consists of four main phases:

1. Generation of an AI-based phishing email dataset
2. Preparation of standardised email formats for testing
3. Execution of detection trials across Gmail, Outlook, Yahoo, Proton.me and Spam Assassin
4. Analysis of detection outcomes using defined metrics

An overview of the experimental workflow is illustrated in Figure 1.

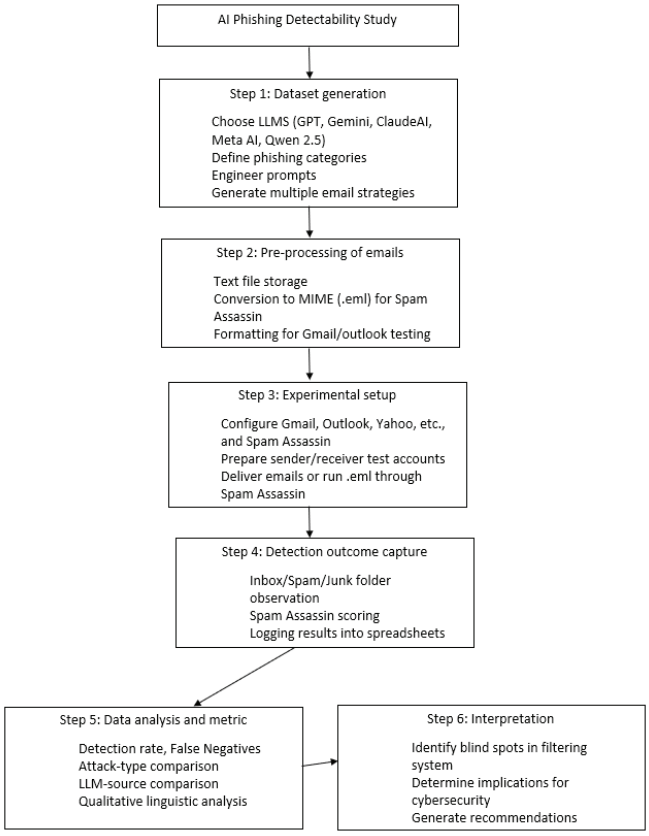

Figure 1. Overall Research Workflow.

This figure illustrates the complete methodological workflow employed in the study, beginning with the generation of AI-generated phishing emails and progressing through pre-processing, experimental setup, detection outcome collection, data analysis, and interpretation. The diagram highlights the sequential flow of activities required to evaluate the detectability of AI-generated phishing emails across multiple email filtering systems.

3.2 Dataset Creation

A custom dataset of phishing emails was generated using 5 prominent large language models: GPT, Claude, Gemini, Meta AI and Qwen 2.5. These models were selected based on their linguistic sophistication, widespread availability, and demonstrated capability to generate human-like text.

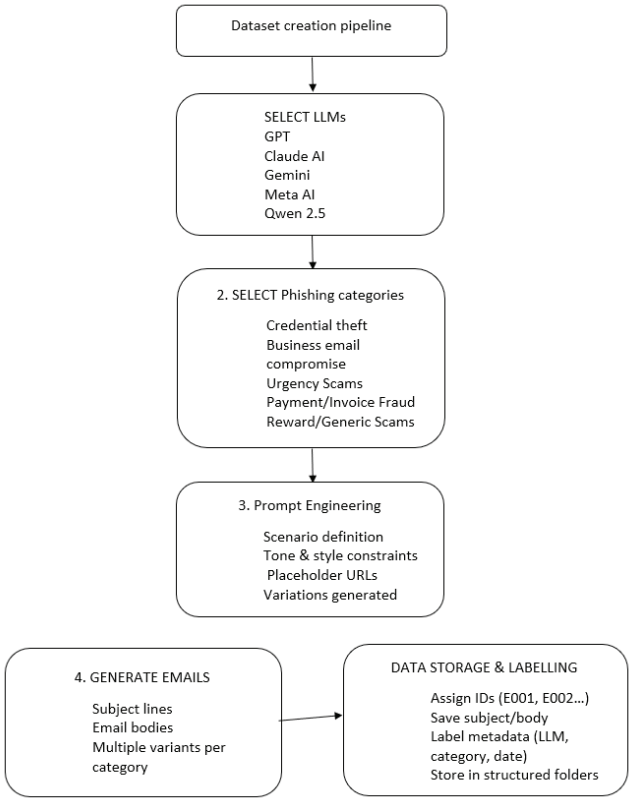

Figure 2. Dataset Creation Pipeline

This figure 2 presents the structured pipeline used to develop the AI-generated phishing dataset. It outlines the selection of large language models, the definition of phishing attack categories, prompt engineering procedures, iterative generation of email variants, and the systematic storage and labelling of samples. The pipeline ensures consistency, diversity, and reproducibility of the dataset.

3.2.2 Prompt Engineering for Email Generation

Phishing emails were produced using structured prompts designed to simulate realistic malicious communication while ensuring ethical compliance (i.e., no real malicious URLs, credentials, or payloads). The requirements are corporate tone, grammatically correct, no real malicious links (use https://example-login.com), and provide subject line and email body.

Each prompt was iteratively refined to produce multiple variants per attack category. A total of 40–60 distinct phishing samples were generated, with balanced representation across LLMs and categories.

3.2.3 Dataset Structuring and Storage

Each generated email was assigned a unique identifier (e.g., *E001, E002*), labelled by LLM source, attack type, subject line and Email body content.

Emails were stored in plaintext format and catalogued in a master spreadsheet (Excel/Google Sheets). An example structure is shown in **Table 1**.

Table 1. Example structure of the AI-generated phishing dataset.

email_ id	llm	attack_ type	subject	body	file- name
E001	**GPT**	Urgency account suspension	Action Required: Identity Verification Needed to Prevent Account Access Restriction		**E001. txt**

3.3 Pre-processing of Email Samples

To ensure compatibility across email filtering systems, phishing emails were prepared in two formats: Copy-paste format for Gmail, Outlook, Yahoo, and Pronto.me testing. MIME-formatted .eml files for Spam Assassin testing.

3.3.1 Preparing Web-Based Email Inputs (Gmail & Outlook, Yahoo, Pronto.me)

Each AI-generated email's subject and body were manually copied into the sender account's compose window to simulate real user-to-user communication. No recipients outside the controlled test accounts were used.

3.3.2 Creating MIME Email Files for Spam Assassin

Each email was transformed into a MIME-compliant .eml file with structured headers, including:

From: testsenderg1@gmail.com

To: testreceiverg1@gmail.com

Subject: Important Notice: Pending Account Credit Disbursement - Reference ID: C21-449X

Date: Sun, 20 Dec 2025 10:00:00 +0000

Message-ID: <E001@gmail.com>

MIME-Version: 1.0

Content-Type: text/plain; charset=UTF-8

Content-Transfer-Encoding: 8bit

Reply-To: replyto@gmail.com

X-Mailer: Gmail 1.0

A pre-processing diagram is shown in **Figure 2**.

3.4 Experimental Setup

The testing environment consisted of two web-based email platforms and one standalone filtering engine.

This figure 3 depicts the architecture of the experimental environment, showing the relationship between sender and receiver accounts, the flow of emails through Gmail, Outlook, cloud-based filtering engines, etc., as shown in Table 2 below. Local processing of MIME-formatted emails by Spam Assassin. It demonstrates how emails were injected into the test environment, filtered, and subsequently logged for analysis.

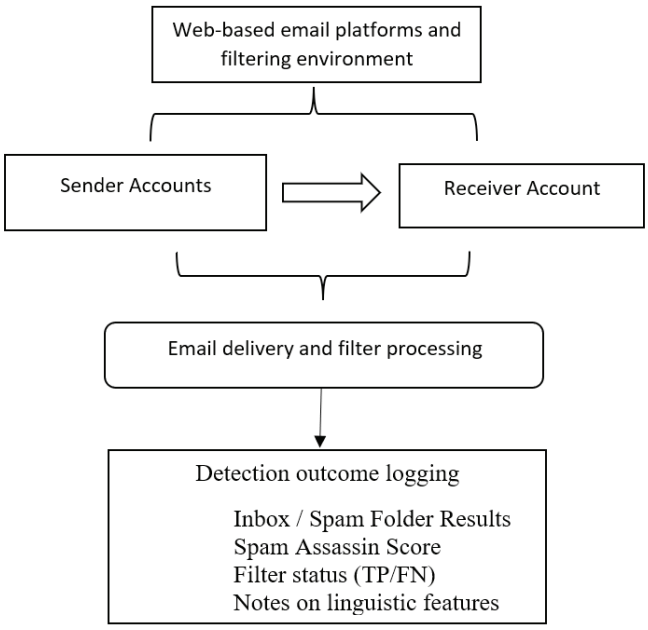

Figure 3. Experimental Setup Architecture

3.4.1 Email Filtering Systems Evaluated

Table 2: Email Filtering Systems characteristics

Filter	Type	Characteristics
Gmail	**Cloud-based**	Machine learning + rule-based filters + sender/domain reputation systems
Outlook	**Cloud-based**	Heuristic analysis + ML + Microsoft threat intelligence engine
Yahoo Mail	**Cloud-based**	ML-driven spam detection + bulk sender analysis + user feedback signals
Pronti.me	Cloud-based	Rule-based filtering + basic ML heuristics + IP/domain reputation checks
Spam Assassin	Rule-based(local)	Score-based textual filtering

3.4.2 Test Accounts and Ethical Isolation

Dedicated sender and receiver accounts were created exclusively for the experiment:

- Gmail sender & receiver
- Outlook sender & receiver
- Yahoo sender and receiver
- Pronti.me sender and receiver

No external users were contacted, and all testing remained isolated.

3.4.3 Email Injection Procedure

For each phishing sample:

1. The email was sent from the designated sender to the corresponding receiver account.
2. The receiver's inbox and spam folder were checked after each delivery attempt.

3. For Spam Assassin, .eml files were processed with:
4. Detection outcomes and spam scores were recorded.

This process was repeated for every email × filter combination, producing a complete matrix of detection results.

3.5 Evaluation Metrics

Two categories of evaluation metrics were used: quantitative performance metrics and qualitative linguistic analysis.

3.5.1 Quantitative Metrics

True Positive (TP) **EQ 1**

Email detected as phishing (moved to Spam/Junk OR marked as spam).

False Negative (FN) **EQ 2**

Email not detected (delivered to the Inbox).

Detection Rate (DR)

$$DR = \frac{TP}{TP+FN} \quad \textbf{EQ 3}$$

False Negative (FNR) **EQ 4**

$$FNR = \frac{FN}{TP+FN} \quad \textbf{EQ 5}$$

Spam Score (Spam Assassin only)

Numerical value indicating rule activation strength.

3.5.2 Qualitative Metrics

Qualitative analysis focused on:

- Linguistic features (tone, style, coherence)
- Structural patterns (greeting, formatting, narrative flow)
- Types of rule-based triggers activated or not activated

This dual approach yields a comprehensive understanding of detection weaknesses.

3.6 Data Analysis Procedure

3.6.1 Quantitative Analysis

Statistical analysis was performed using Excel and Google Sheets:

- Calculation of DR and FNR across filters
- Pivot tables comparing LLMs and attack types
- Visualisations including bar graphs and heatmaps

3.6.2 Qualitative Analysis

Emails that bypassed detection were subjected to:

- Thematic coding
- Structural comparison across LLMs
- Review of Spam Assassin rule activations
- Examination of linguistic signals that might confuse ML-based filters

This analysis identifies the root causes of detection failures.

3.6.3 Triangulation

Findings were cross validated by:

- Comparing multiple filters
- Multiple LLMs
- Multiple phishing categories
- Both quantitative and qualitative evidence

This ensures validity and robustness.

3.7 Ethical Considerations

Ethical compliance was maintained by:

1. Using only researcher-controlled email accounts
2. Avoiding any malicious payloads or real phishing links
3. Ensuring all AI-generated emails served defensive research purposes
4. Not disseminating phishing emails outside the isolated environment
5. Storing data securely and anonymising metadata

This satisfies academic ethical requirements for cybersecurity experimentation

4. Results

This chapter presents the empirical results of evaluating AI-generated phishing emails against modern email filtering systems. The analysis follows the structured data analysis workflow defined in Chapter 3 and reports results in four stages: raw detection outcomes, quantitative performance metrics, comparative analysis across filters, attack types, and language models, and identification of detection blind spots. Interpretive explanations are reserved for Chapter 5.

4.1 Experimental Output Overview

A total of 100 phishing emails generated using large language models were evaluated against five email filtering systems: Gmail, Yahoo Mail, Outlook, Proton Mail, and Spam Assassin. All emails were malicious by design, allowing the analysis to focus exclusively on true positive and false negative detection outcomes without the inclusion of legitimate messages.

For each email–filter combination, detection outcomes were recorded. In the case of Spam Assassin, numerical spam scores and triggered detection rules were also captured. Due to space constraints, the complete per-email detection log is not reproduced in this chapter and is instead provided on request.

4.2 Detection Outcome Classification

Detection outcomes were classified according to two categories. Emails classified as spam by a filtering system were labelled as true positives, while phishing emails that were not flagged and were delivered to the inbox were labelled as false negatives. True negative and false positive outcomes were not applicable given the absence of legitimate emails in the dataset.

4.3 Overall filter performance

Overall detection effectiveness was quantified using the detection

rate and false negative rate. Detection rate was defined as the proportion of phishing emails correctly classified as spam, while the false negative rate represented the proportion of phishing emails that bypassed detection.

Table 3- Overall Detection Performance across Email Providers.

Email provider	Total email tested	True positives (TP)	False Negatives (FN)	Detection rate	False Negative rate
Gmail	25	20	5	0.8	0.25
Yahoo mail	25	2	23	0.08	0.92
Outlook	25	4	21	0.16	0.84
Proton mail	25	0	25	0	1
Spam Assassin	25	7	18	0.28	0.72

This table provides a system-level comparison of phishing detection performance across commercial, privacy-focused, and rule-based filtering systems.

4.4 Spam Assassin Spam Score Distribution

Spam Assassin assigns a numerical spam score to each evaluated email. Emails with scores at or above the default threshold of 5.0 were classified as spam, while those below the threshold were considered undetected.

Table 4: Spam Score Distribution for Spam Assassin

Spam score range	Number of emails
Below 3.0	**57**
3.0 to 4.9	30
5.0 to 6.9	8
7.0 and above	**5**

The distribution illustrates the proportion of emails that were clearly detected, borderline detectable, or effectively evasive with respect to rule-based scoring.

4.5 Spam Assassin Rule Trigger Frequency Analysis

This section examines the rule activation behaviour of Spam Assassin when processing AI-generated phishing emails. The objective of this analysis is to identify which detection mechanisms were most frequently triggered and to assess whether detection decisions were primarily influenced by message structure, sender authentication inconsistencies, or semantic indicators of phishing intent.

Spam Assassin evaluates messages using a large collection of heuristic rules, each corresponding to a specific property of email construction, sender authenticity, or message content. When triggered, a rule contributes to the overall spam score assigned to the message. Given the extensive number of available rules, analysis was restricted to a subset of rules that were consistently observed across evaluated emails and that directly reflect phishing-relevant behaviours.

For each evaluated email, the set of triggered rules was extracted from the X-Spam-Status header. Each selected rule was counted once per email when triggered, independent of its numerical score contribution. Rule frequencies were then aggregated by phishing category.

Table 5 Spam Assassin Rule Trigger Frequency by Phishing Category

Rule Name	BEC	Credential Theft	General	Payment / Invoice	Urgency / Account
NO_RECEIVED	5	5	5	5	5
NO_RELAYS	5	5	5	5	5
DKIM_ADSP_CUSTOM_MED	5	5	5	5	5
NML_ADSP_CUSTOM_MED	5	5	5	5	5
MSGID_SHORT	5	5	5	1	5
FORGED_GMAIL_RCVD	5	5	5	5	5
FREEMAIL_FROM	4	4	4	4	4
FREEMAIL_REPLYTO	5	4	5	0	5
HDRS_MISSP	5	0	0	0	1
URG_BIZ	3	0	0	0	0
TVD_PH_BODY_ACCOUNTS_PRE	0	4	0	0	1
TVD_PH_SEC	0	1	0	0	1
LOTS_OF_MONEY	4	0	2	3	0
MONEY_FREEMAIL_REPTO	4	0	0	0	0
XFER_LOTSA_MONEY	5	0	0	3	0
ADVANCE_FEE_2_NEW_MONEY	1	0	0	0	0
ADVANCE_FEE_4_NEW_MONEY	2	0	0	0	0
MONEY_FRAUD_5	1	0	0	0	0
URI_PHISH	0	3	0	0	0
INVALID_DATE	0	1	0	0	0
MISSING_MID	0	1	0	0	0
PP_MIME_FAKE_ASCII_TEXT	0	1	0	0	0
TVD_PH_BODY_META	0	0	1	0	0
UNCLAIMED_MONEY	0	0	2	0	0
T_FILL_THIS_FORM_SHORT	0	0	1	0	0
ADVANCE_FEE_5_NEW	0	0	1	0	0

This table summarises the frequency with which selected rules were activated across phishing categories and highlights dominant detection cues associated with different attack strategies.

The rule trigger frequency analysis demonstrates that a common

set of structural and sender-authentication rules was activated across all phishing categories examined. Rules related to message routing, sender domain alignment, and header formatting appeared consistently in every category. In contrast, semantic rule activation varied by phishing type. Financial manipulation rules were predominantly observed in business email compromise and payment-related emails, while account-related and phishing URL rules occurred more frequently in credential theft and urgency-based emails. General Phishing emails exhibited a broader and less consistent distribution of semantic rule activations.

This section has presented an aggregated view of Spam Assassin rule activations across phishing categories, providing a descriptive account of rule frequency patterns observed during evaluation.

4.6 Comparative Detection Performance across Filtering Systems

To facilitate direct comparison across filtering systems, detection and false negative rates were examined side by side.

Table 6. Cross-Filter Detection Rate Comparison

Filter System	Detection Rate	False Negative Rate
Gmail	0.8	0.25
Yahoo Mail	0.08	0.92
Outlook	0.16	0.84
Proton Mail	0	1
Spam Assassin	0.28	0.72

This comparison highlights differences in detection effectiveness between provider-managed and rule-based filtering approaches.

4.7 Detection Performance by Phishing Attack Category

Detection outcomes were grouped by phishing category to assess whether certain attack strategies were more likely to evade detection.

Table 7: Detection Performance by Phishing Category

Phishing Category	Total Emails	True Positives	False Negatives	Detection Rate
Business Email Compromise	20	5	15	0.25
Credential Theft	20	6	14	0.3
Payment and Invoice Fraud	20	4	16	0.2
Urgency and Account Suspension	20	5	15	0.25
General	20	4	1	0.8

This table enables identification of category-specific detection weaknesses.

4.8 Detection Performance by Large Language Model

To evaluate whether phishing emails generated by different language models exhibited varying detectability, detection outcomes were grouped by model.

Table 8: Detection Performance by Large Language Model

Language Model	Total Emails	True Positives	False Negatives	Detection Rate
ChatGPT	20	6	14	0.3
Claude	20	7	13	0.35
Gemini	20	5	15	0.25
Meta AI	20	4	16	0.2
Qwen 2.5	20	4	16	0.2

This analysis provides a model-level view of phishing detectability.

4.9 Statistical Significance of Detection Differences

To evaluate whether observed variations in phishing detection outcomes were statistically associated with the language model used to generate the emails, the email filtering system, or the phishing category, a series of chi-square tests of independence were conducted. Detection outcome was treated as a binary variable, classified as either detected or not detected. Statistical testing was performed at a significance level of 0.05.

4.9.1 Detection Outcome × Language Model

A chi-square test of independence was conducted to examine the association between detection outcome and the large language model used to generate the phishing emails. Observed detection counts were aggregated across five language models: ChatGPT, Claude, Gemini, Meta AI, and Qwen 2.5.

The test yielded a chi-square statistic of χ^2 (df = 4) = 1.85 with a corresponding p-value of 0.76. This result indicates that the detection outcome was not statistically associated with the language model used to generate the phishing emails. Although observed detection rates varied across language models, these differences did not exceed what would be expected under the assumption of independence.

4.9.2 Detection Outcome × Email Filtering System

A chi-square test of independence was applied to assess whether detection outcome differed across the evaluated email filtering systems, namely Gmail, Yahoo Mail, Outlook, and Proton Mail. Detection outcomes were aggregated for each filtering system and analysed using a two-by-four contingency table.

The analysis produced a chi-square statistic of χ^2 (df = 3) = 52.17 with a p-value less than 0.001. This result indicates a statistically significant association between detection outcome and the email filtering system. Observed detection counts differed substantially across filtering platforms, contributing to the magnitude of the chi-square statistic.

4.9.3 Detection Outcome × Phishing Category

A further chi-square test of independence was conducted to evaluate whether detection outcomes varied across phishing categories, including business email compromise, credential theft, general phishing, payment or invoice fraud, and urgency-based account suspension attacks.

The test yielded a chi-square statistic of χ^2 (df = 4) = 0.73 with a corresponding p-value of 0.95. This result indicates that the detection outcome was not statistically associated with the phishing category in the evaluated dataset. Observed detection counts across phishing categories were closely aligned with expected values

under the null hypothesis of independence.

4.9.4 Summary of Statistical Test Results

Table 4.9 summarises the outcomes of the chi-square tests conducted across the three comparison dimensions. Collectively, the results indicate that while detection outcomes did not differ significantly across language models or phishing categories, statistically significant differences were observed across email filtering systems.

Table 4.9 Statistical Significance of Detection Differences

Comparison	Test Used	df	χ^2	p-value	Significant ($\alpha = 0.05$)
Detection Outcome × Language Model	Chi-square	4	1.85	0.76	No
Detection Outcome × Filter System	Chi-square	3	52.17	< 0.001	Yes
Detection Outcome × Phishing Category	Chi-square	4	0.73	0.95	No

4.10 Chapter Summary

This chapter presented the results of evaluating AI-generated phishing emails against multiple email filtering systems. Detection performance was quantified using detection rates, spam score distributions, rule trigger frequencies, and comparative analyses across filters, phishing categories, and language models. The results reveal systematic detection gaps and recurring false negative patterns, which are examined and contextualised in Chapter 5.

5. Discussion

5.1 Principal Findings and Alignment with Research Objectives

This study set out to empirically assess the detectability of AI-generated phishing emails by modern email filtering systems and to quantify the extent of detection failure in real-world defensive infrastructure. The results provide clear and convergent evidence that AI-generated phishing emails are frequently misclassified as legitimate, thereby validating the core research problem articulated in Chapter 1.

Across all evaluated systems, false negatives were pervasive, with detection rates ranging from complete failure (0%) to partial success (80%). These findings substantiate prior concerns in the literature that large language model (LLM)–generated phishing content represents a structural challenge to contemporary email filtering paradigms, rather than a marginal improvement over traditional phishing techniques (Hazell, 2023; Qi et al., 2024; Josten and Weis, 2024). Crucially, this study extends existing work by demonstrating that such failures are not confined to laboratory classifiers but are observable within widely deployed, production-grade filtering systems, thereby addressing a critical empirical gap.

5.2 Effectiveness of Modern Email Filters against AI-Generated Phishing (RQ1)

The overall detection performance reveals pronounced disparities across filtering platforms, with false negative rates ranging from 25% to 100%. Gmail exhibited the highest detection rate (0.8), while Proton Mail failed to detect any phishing emails. Yahoo Mail, Outlook, and Spam Assassin demonstrated detection rates below 0.3, underscoring a widespread inability to reliably identify AI-generated phishing content.

The statistically significant association between detection outcome and filtering system ($\chi^2 = 52.17$, $p < 0.001$) confirms that defensive capability is contingent on filter architecture, training data regency, and adaptive learning capacity. This finding empirically supports the argument advanced by Qi et al. (2024) that legacy and static detection systems are fundamentally misaligned with AI-driven attack vectors.

Notably, even the strongest-performing system (Gmail) exhibited a non-trivial false negative rate, which is operationally significant given the elevated success rates of AI-generated phishing reported in prior studies (Dash and Sharma, 2023; Heiding et al., 2024). This suggests that relative robustness does not equate to adequate protection when facing high-impact, targeted phishing campaigns.

5.3 Rule-Based Detection Limitations and Feature-Space Evasion

The performance of Spam Assassin provides insight into the limitations of rule-based and heuristic-driven filtering approaches. With a detection rate of 0.28 and a false negative rate of 0.72, Spam Assassin exemplifies the fragility of systems reliant on predefined rules and historical indicators of maliciousness (Awad and ELseuofi, 2011; Anitha et al., 2021).

The spam score distribution shows that many AI-generated phishing emails failed to reach the detection threshold, indicating insufficient rule activation. This aligns with Josten and Weis (2024), who demonstrated that LLM-generated or LLM-modified spam systematically avoids lexical and grammatical cues traditionally exploited by Bayesian and rule-based detectors.

Further analysis of rule activation frequencies reveals that structural and sender-authentication rules were consistently triggered, while semantic phishing indicators were sparse and inconsistent across categories. This pattern confirms that AI-generated phishing neutralises the feature-space assumptions underpinning traditional detection by presenting linguistically "clean," semantically plausible content. As argued by Abuadbba et al. (2022) and Tang and Li (2025), such attacks evade detection not by introducing noise, but by removing it.

5.4 Comparative Performance across Filtering Systems

The comparative analysis highlights a meaningful distinction between provider-managed cloud filters and local rule-based engines. Gmail's comparatively higher detection rate suggests that large-scale platforms leveraging continuous telemetry, adaptive machine learning, and user feedback loops may possess partial resilience against AI-generated phishing.

However, the continued presence of false negatives even within advanced cloud-based systems indicates that current detection architectures remain reactive and pattern-dependent, rather than intent-aware. The poor performance of Yahoo Mail, Outlook, and Proton Mail further reinforces the argument that reliance on historically derived heuristics and static training corpora exposes systems to systematic evasion by generative models (Jeeva and Khan, 2023; Abdolrazzagh-Nezhad and Langarib, 2025).

5.5 Phishing Strategy and Detection Outcomes (RQ2)

Category-based analysis shows that general phishing emails achieved a substantially higher detection rate (0.8) compared to targeted attacks such as business email compromise, credential theft, payment fraud, and urgency-based scams (≤ 0.3). This observation is consistent with prior work suggesting that generalised phishing retains detectable regularities, while targeted attacks increasingly exploit contextual and psychological nuance (Ghazi-Tehrani and Pontell, 2021).

However, the absence of a statistically significant association between detection outcome and phishing category ($p = 0.95$) indicates that detection failure is systemic rather than strategy specific. This suggests that AI-generated phishing broadly undermines detection mechanisms, irrespective of the specific social engineering narrative employed.

5.6 Impact of Language Model Choice on Detectability (RQ3)

Detection rates varied modestly across language models, with Claude exhibiting the highest detectability and Meta AI and Qwen 2.5 the lowest. Nonetheless, statistical testing revealed no significant association between detection outcome and language model ($p = 0.76$).

This finding implies that detectability is driven by shared properties of modern LLMs, rather than model-specific artefacts. As argued by Eze and Shamir (2024), AI-generated phishing constitutes a stylistically distinct class characterised by linguistic fluency, semantic coherence, and professional tone features common across contemporary generative models. Consequently, defensive strategies targeting individual models or prompt artefacts are unlikely to yield durable protection.

5.7 Linguistic and Structural Factors Underpinning Detection Failure (RQ4)

Qualitative analysis of false negatives reveals consistent linguistic and structural characteristics, including grammatical correctness, conventional formatting, professional tone, and high semantic similarity to legitimate correspondence. These features align with findings by Du et al. (2024) and Olea et al. (2025), who demonstrated that AI-generated phishing is frequently perceived as benign by both humans and automated systems.

Unlike earlier adversarial text attacks that relied on perceptible perturbations (Boucher et al., 2021), AI-generated phishing leverages linguistic normalcy as an evasion mechanism, forcing detection systems to infer malicious intent rather than identify surface-level anomalies. This represents a fundamental shift in the attacker–defender dynamic.

5.8 Implications for Cybersecurity Practice and Policy (RQ5)

The findings have significant implications for cybersecurity practice. Organisations relying primarily on automated email filtering may be systematically exposed to AI-generated phishing threats, particularly in high-value, targeted contexts. The results underscore the continued importance of defence-in-depth strategies, including user awareness training and contextual verification processes.

The failure of email filters observed in this study is not merely a model weakness but an architectural issue, where detection systems designed for historical human-generated threats are structurally unprepared for AI-generated behaviour, echoing broader concerns about legacy institutional systems facing AI integration challenges (Obeta et al., 2026).

At a broader level, the study reinforces calls for a paradigm shift in phishing defence, moving beyond surface-level textual analysis toward intent-aware, context-sensitive detection frameworks (Ferrara, 2024; Abuadbba et al., 2022). Addressing the dual-use nature of LLMs, as discussed by Koplin (2023), will require coordinated technical, organisational, and policy-level responses.

6. Conclusion

This study makes a substantive contribution by providing real-world empirical evidence of the limitations of modern email filters when confronted with AI-generated phishing emails. By evaluating production-grade filtering systems rather than laboratory classifiers, the research bridges a critical gap between academic theory and operational cybersecurity practice. The results demonstrate that AI-generated phishing emails systematically evade detection by modern email filtering systems, irrespective of phishing strategy or language model used. Detection performance varies significantly across platforms, revealing structural weaknesses in current defences. These findings confirm the urgent need for adaptive, AI-aware detection strategies and establish a robust empirical foundation for future research and defensive innovation.

Reference

Abdolrazzagh-Nezhad, M. and Langarib, N. (2025) 'Phishing detection techniques: A review', Data Science: *Journal of Computing and Applied Informatics*, 9(1), pp. 32–46. doi: https://doi.org/10.32734/jocai.v9.i1-19904.

Abuadbba, A., Wang, S., Almashor, M., Ahmed, M.E., Gaire, R., Camtepe, S. and Nepal, S. (2022) 'Towards web phishing detection: Limitations and mitigation', *arXiv preprint*, arXiv:2204.00985. https://doi.org/10.48550/arXiv.2204.00985.

Alahmed, Y., Abadla, R. and Al Ansari, M.J. (2024) 'Exploring the potential implications of AI-generated content in social engineering attacks', in *Proceedings of the 2024 International Conference on Multimedia Computing, Networking and Applications (MCNA)*, Valencia, Spain, pp. 64–73. https://doi.org/10.1109/MCNA63144.2024.10703950

Aldam, D. (2025) 'AI-powered phishing: The current landscape and future projections', XRDS: Crossroads, *The ACM Magazine for Students*, 31(4), pp. 18–21. https://doi.org/10.1145/3744693

Alsuwit, M.H., Haq, M.A. and Aleisa, M.A. (2024) 'Advancing email spam classification using machine learning and deep learning techniques', *Engineering, Technology & Applied Science Research*, 14(4), pp. 14994–15001. https://doi.org/10.48084/etasr.7631

Anitha, P., Rao, C.G. and Babu, D.S. (2021) 'Email spam filtering using machine learning based XGBoost classifier method', *Turkish Journal of Computer and Mathematics Education*, 12(11), pp. 2182–2190.

Awad, W.A. and ELseuofi, S. (2011) 'Machine learning methods for spam e-mail classification', *International Journal of Computer Science & Information Technology (IJCSIT)*, 3(1), pp. 173–184. doi: 10.5121/ijcsit.2011.3112.

Boucher, N., Shumailov, I., Anderson, R. and Papernot, N. (2022) 'Bad characters: Imperceptible NLP attacks', in *Proceedings of the 2022 IEEE Symposium on Security and Privacy (SP)*. https://doi.org/10.1109/

SP46214.2022.9833641

Chien, A. and Khethavath, P. (2023) 'Email feature classification and analysis of phishing email detection using machine learning techniques', in *Proceedings of the 2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pp. 1–8. https://api.semanticscholar.org/CorpusID:268931485

Dash, B. and Sharma, P. (2023) 'Are ChatGPT and deepfake algorithms endangering the cybersecurity industry? A review', *International Journal of Engineering and Applied Sciences*, 10(1), pp. 1–5. doi:10.31873/IJEAS.10.1.01.

Dionysiou, A. and Athanasopoulos, E. (2021) 'Unicode evil: Evading NLP systems using visual similarities of text characters', in *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security (AISec '21)*, pp. 1–12. https://doi.org/10.1145/3474369.348687

Du, W., Ju, T., Ren, G., Li, G. and Liu, G. (2024) 'Backdoor NLP models via AI-generated text', in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italy. ELRA and ICCL, pp. 2067–2079. https://aclanthology.org/2024.lrec-main.186/

Eze, C.S. and Shamir, L. (2024) 'Analysis and prevention of AI-based phishing email attacks', *Electronics*, 13(10), p. 1839. https://doi.org/10.3390/electronics13101839

Ferrara, E. (2024) 'GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models', *Journal of Computational Social Science*, 7(1), pp. 549–569. https://doi.org/10.1007/s42001-024-00250-1

Ghazi-Tehrani, A.K. and Pontell, H.N. (2021) 'Phishing evolves: Analyzing the enduring cybercrime', *Victims & Offenders*, 16(3), pp. 316–342. https://doi.org/10.1080/15564886.2020.1829224

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T. and Fritz, M. (2023) 'Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection', in *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*. https://doi.org/10.48550/arXiv.2302.12173

Hazell, J. (2023) 'Spear phishing with large language models', *arXiv preprint*, arXiv:2305.06972. https://doi.org/10.48550/arXiv.2305.06972

Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J. and Park, P.S. (2023) 'Devising and detecting phishing: Large language models vs. smaller human models', *arXiv preprint*, arXiv:2308.12287. https://doi.org/10.48550/arXiv.2308.12287

Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J. and Park, P.S. (2024) 'Devising and detecting phishing emails using large language models', *IEEE Access*, 12, pp. 42131–42146. https://doi.org/10.1109/ACCESS.2024.3375882

Humphreys, D., Koay, A., Desmond, D. and Mealy, E. (2024) 'AI hype as a cyber security risk: The moral responsibility of implementing generative AI in business', *AI and Ethics*, 4(3), pp. 791–804. https://doi.org/10.1007/s43681-024-00443-4

Jaiswal, R., Marshal, R., Rao, V.V. and Singh, K.P. (2024) 'AI phishing detection framework for businesses with limited resources', in *Proceedings of the 2024 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*. https://doi.org/10.1109/3ict64318.2024.10824248

Jeeva, L. and Khan, I.S. (2023) 'Enhancing email spam filter's accuracy using machine learning', *International Journal for Multidisciplinary Research*, 5(4). https://doi.org/10.55524/ijircst.2023.11.4.2

Josten, M. and Weis, T. (2025) 'Investigating the effectiveness of Bayesian spam filters in detecting LLM-modified spam mails', in Goel, S., Uzun, E., Xie, M. and Sarkar, S. (eds) *Digital Forensics and Cyber Crime*. ICDF2C 2024. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 613. Cham: Springer, pp. 285–295. https://doi.org/10.1007/978-3-031-89363-6_16

Koplin, J.J. (2023) 'Dual-use implications of AI text generation', *Ethics and Information Technology*, 25(2), p. 32. https://doi.org/10.1007/s10676-023-09703-z

Kulal, D.H., Arannonu, C.P., Anwar, A., Rastogi, N. and Niyaz, Q. (2025) 'Robust ML-based detection of conventional, LLM-generated, and adversarial phishing emails using advanced text preprocessing', *arXiv preprint*, arXiv:2510.11915. https://doi.org/10.48550/arXiv.2510.11915

Kumar, N. and Patel, N.M. (2025) 'Social engineering attack in the era of generative AI', *International Journal for Research in Applied Science and Engineering Technology*, 13(1), pp. 1737–1747. doi:10.22214/ijraset.2025.66688.

Kumar, S., Menezes, A., Giri, S. and Kotikela, S. (2024) 'What the Phish! Effects of AI on phishing attacks and defense', in *Proceedings of the International Conference on AI Research*. Academic Conferences and Publishing Limited. https://doi.org/10.34190/icair.4.1.3224

Xu, W. and Parhi, K.K. (2025) 'A survey of attacks on large language models', *arXiv preprint*, arXiv:2505.12567. https://arxiv.org/abs/2505.12567.

Liesnaia, Y. and Malakhov, S. (2023) 'The analysis of development, typical objectives and mechanisms of phishing attacks', *Computer Science and Cybersecurity*, (1), pp. 6–27. https://orcid.org/0000-0001-8826-1616

Murti, Y. S., & Naveen, P. (2023). Machine learning algorithms for phishing email detection. *Journal of Logistics, Informatics and Service Science*, 10(2), 249-261. doi:10.33168/JLISS.2023.0217.

Obeta, S., Chigere, A., Ibanga, I., Itua, R., Oraegbunam, L., Nwanakwaugwu, A.C., Ozioma, G.N. and Anumaka, C. (2026) 'From hierarchies to loops: Rethinking public sector structures for responsible AI integration', *Ethics*, 1(1), pp. 1–10.

Olea, C., Christensen, A., Fazio, L., Cutting, L., Lieb, M., Phelan, J. and Tucker, H. (2025) 'Evaluating phishing email efficacy', in *Proceedings of the 2025 Computers and People Research Conference (SIGMIS-CPR '25)*. Article 7, pp. 1–8. https://doi.org/10.1145/3716489.372843

Osamor, J., Ashawa, M., Shahrabi, A., Phillip, A. and Iwend, C. (2025) 'The evolution of phishing and future directions: A review', in *Proceedings of the International Conference on Cyber Warfare and Security*, 20(1), pp. 361–368. https://doi.org/10.34190/iccws.20.1.3366

Ozioma, G., Obeta, S., Ibanga, D.I., Oraegbunam, L., Itua, R., Amaefule, A., Ozioma, E.O. and Anumaka, C. (2025) 'AI as cognitive ecology: Revealing the invisible cognitive, cultural, and epistemic costs of generative models', *Cultural, Cognitive, and Epistemic Costs of Generative Models*, 16 January.

Pang, Y., Meng, W., Liao, X. and Wang, T. (2025) 'Paladin: Defending LLM-enabled phishing emails with a new trigger-tag paradigm', *arXiv preprint*, arXiv:2509.07287. https://doi.org/10.48550/arXiv.2509.07287

Pawar, K. and Patil, M. (2015) 'Pattern classification under attack on spam filtering', in *Proceedings of the IEEE*

International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN 2015).

Qi, Q., Luo, Y., Xu, Y., Guo, W. and Fang, Y. (2024) 'SpearBot: Leveraging large language models in a generative-critique framework for spear-phishing email generation', *arXiv preprint*, arXiv:2412.11109. https://doi.org/10.48550/arXiv.2412.11109

Tang, Q. and Li, X. (2025) 'An investigation of large language models and their vulnerabilities in spam detection', *arXiv preprint*, arXiv:2504.09776. https://doi.org/10.48550/arXiv.2504.09776

Vadisetty, R. and Polamarasetti, A. (2024) 'Generative AI for cyber threat simulation and defense', in *Proceedings of the 12th International Conference on Control, Mechatronics and Automation (ICCMA 2024)*. London, United Kingdom, pp. 272–279. https://doi.org/10.1109/ICCMA63715.2024.10843938

Valencia, L.J. (2024) 'Artificial intelligence as the new hacker: Developing agents for offensive security', *arXiv preprint*, arXiv:2406.07561. https://doi.org/10.48550/arXiv.2406.07561

Vishwanath, A. (2022) *The weakest link: How to diagnose, detect, and defend users from phishing*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/14653.001.0001

Vu, M.T., Tran, Q.A., Jiang, F. and Tran, V.Q. (2015) 'Multilingual rules for spam detection', Journal of Machine to Machine Communications, 1(2), pp. 107–122. doi: 10.13052/jmmc2246-137x.122.

Yoo, S., Yang, Y., Lin, F. and Moon, I.-C. (2009) 'Mining social networks for personalized email prioritization', in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09). New York: ACM*, pp. 967–976. https://doi.org/10.1145/1557019.1557124

Zhang, W.E., Sheng, Q.Z., Alhazmi, A. and Li, C. (2020) 'Adversarial attacks on deep-learning models in natural language processing: A survey', *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3), pp. 1–41.