Issues in Information Systems
Volume 25, Issue 3, pp. 327-341, 2024

DOI: https://doi.org/10.48009/3_iis_2024_125

How effective are large language models in detecting phishing emails?

Jing Hua, *Robert Morris University, jxhst641@mail.rmu.edu*
Ping Wang, *Robert Morris University, wangp@rmu.edu*
Peyton Lutchkus, *Robert Morris University, pglst259@mail.rmu.edu*

Abstract

Phishing emails exploit human vulnerabilities to illicitly obtain sensitive information, representing a critical research focus in cybersecurity. This study explores the effectiveness of artificial intelligence (AI), specifically Large Language Models (LLMs), in detecting phishing emails. Evaluating LLM-based models, including ChatGPT-4 and Gemini, the study directly tests them on a mixed dataset of sanitized phishing emails and non-phishing emails using a defined set of phishing indicators and measures. The findings contribute to understanding the practical application of AI in detecting phishing attempts, advancing discourse on AI's role in cybersecurity.

Keywords: phishing detection, phishing indicators, AI, LLMs, ChatGPT-4, Gemini

Introduction

The rapid advancement of artificial intelligence (AI) has driven transformative developments across various industries, including cybersecurity. Among these, the emergence of AI-powered text generation models, such as ChatGPT, has enabled the creation of highly convincing phishing emails, therefore presenting significant challenges to cybersecurity defenses (Karanjai, 2022; Sadasivan et al., 2023; Tang et al., 2024; Nov et al., 2023). Researchers like Kalla (2023) highlight that natural language models enhance phishing detection capabilities by analyzing the language used in emails. For instance, Jiang (2024) notes that models like GPT-3.5 and GPT-4 can effectively identify typical phishing indicators such as suspicious sender addresses, unusual links, and a sense of urgency. Furthermore, Heiding et al. (2024) provide quantitative evidence showing that LLMs can achieve up to 100% accuracy in phishing detection when fine-tuned with specific prompts. However, their approach also reveals a limitation: these prompts may not fully align with real-world detection scenarios, where email filters must decisively block or allow emails without the ambiguity of "certainty levels" (Heiding et al., 2024).

To address the challenges posed by advanced AI models in cybersecurity, this study aims to explore strategies to effectively detect phishing attempts. It specifically assesses the performance of two AI models, Gemini Advanced and ChatGPT-4, in distinguishing phishing emails from legitimate digital marketing communications within a diverse dataset. Through comparative analysis, this study evaluates these models' ability to accurately identify and differentiate phishing attempts, thereby contributing to enhanced cybersecurity measures.

Background

Phishing emails leverage social engineering to exploit human vulnerabilities, often exhibiting distinctive characteristics such as psychological manipulation patterns, unusual linguistic styles, and other deceptive tactics. These characteristics serve as indicators for distinguishing phishing attempts from legitimate communications. This section reviews recent research on key phishing indicators, as summarized in Table 1 below, which highlights advances and findings in the field over recent years.

Table 1: Summary of Research on Phishing Indicators

Authors & Year	Phishing Indicators and Tactics	Methodology	Key Findings/Conclusions
Jiang (2024)	Suspicious sender address, unusual links, grammar and spelling errors, the sense of urgency, and the generic salutation	Experiment	- Large Language Models (LLMs) like GPT-3.5 and GPT-4 are able to identify the common red flags or the listed indicators of phishing or scam emails - Results may vary with training data, fine tuning methods, and model versions
Zoltan & Weigand (2023)	Emotional manipulation; Suspicious media outlets; Visual appearance and wording; Suspicious origin/source; Lack of reliable/official media coverage; Suspicious images/videos; Exploiting prior beliefs and biases of target	"Media Detective" project study of a fake news database	- Taxonomy of 7 categories of recurring signs of fake news & phishing - A fake news post or phishing scam may involve multiple signs - Phishing scams are harder to detect due to artificial intelligence
Wang & Lutchkus (2023)	Psychological Principles of Influence in Phishing Emails: - Reciprocation - Consistency - Social Proof - Liking - Authority - Scarcity	Case studies	- Explores the in-depth psychological factors that target human vulnerabilities in phishing emails - Illustrates major principles of influence with case analyses of phishing emails from the Berkeley Phish Tank
Murtaza et al. (2022)	Social/group influence methods to alter target behavior or attitude. Persuasion tactics: - Similarity of interests - Distraction/manipulation - Exploiting curiosity - Authority/credibility (e.g. Use official logos or symbols to show authority and credibility for effective attacks)	Theoretical review and analysis of social engineering and phishing attacks and solutions	- Taxonomy of common types of phishing attacks - Taxonomy of influence methodologies used in social engineering and phishing - Exploiting human vulnerabilities is key to efficient social engineering attacks - Improving workforce security awareness is an effective approach to mitigating the human vulnerability factor in social engineering and phishing attacks

Authors & Year	Phishing Indicators and Tactics	Methodology	Key Findings/Conclusions
Wash (2020)	Cues triggering suspicion: Action link; Typo; Suggested by another person; Recipients in From line; Unusual subject line; Attachment; URL; Unusual information requested	Interviews and case studies	- Identified 3 stages of phishing email detection: 1) Sensemaking 2) Cognitive shift 3) Action to deal with the email - How experts become suspicious? 1) Noticing discrepancies 2) Cues triggering shift to suspicion - However, lack of expertise in each stage may lead to failure of detection
Williams & Polage (2019)	Message-specific factors that may impact trustworthiness and persuasiveness of phishing emails: 1) Loss and reward-based techniques; 2) Authentic design cues; 3) Reference to current events.	Survey	- A finding is that loss-based influence techniques and authentic design cues increase perceived trust and persuasiveness of emails - Psychological factors that may account for the finding need further study in the future
Kleitman et al. (2018)	Known Phishing Email Characteristics: - Ask for confidential information - Spelling/grammar errors - Pressure for response - Vague recipient - Suspicious email domain - Suspicious URL	Experimental	- Human-centered variables account for the majority of variance in phishing susceptibility - Identified major email features for phishing detection - Findings on characteristics of most successful phishing email detection - The major limitation is that there is no standard measure for phishing detection

The research efforts presented in Table 1 above employ various methodologies including experiments and surveys with some important progress and empirical findings. These studies also reveal important characteristics and tactics of phishing emails that could become indicators for detection of phishing. There are a wide range of indicators and tactics of phishing emails identified by these research efforts. Visual and linguistic cues, including imitation symbols, logos, errors of spelling and grammar, could be obvious telltale signs of phishing emails (Kleitman et al., 2018; Jiang, 2024; Murtaza et al., 2022; Wash, 2020; Williams & Polage, 2019; Zoltan & Weigand, 2023). Some common email elements that may trigger reader suspicion, including obscure and suspicious email domain, URLs, recipients, images, attachments, etc., could also be indicators for phishing detection (Kleitman et al., 2018; Jiang, 2024; Wash, 2020; Zoltan & Weigand, 2023). More subtle indicators of phishing include emotional manipulation and appeals to beliefs, trustworthiness, curiosity, urgency, and scarcity (Jiang, 2024; Murtaza et al., 2022; Wang & Lutchkus, 2023; Zoltan & Weigand, 2023). The least obvious and most challenging indicators of phishing emails may be the psychological tactics and principles of influence and persuasion used to manipulate victims' attitude and behavior (Murtaza et al., 2022; Wang & Lutchkus, 2023; Williams & Polage, 2019; Zoltan & Weigand, 2023). The psychological tactics and indicators require more in-depth cognitive and intellectual analysis for detection.

Given the wide range of possible tactics and indicators of phishing emails, having a core set of indicators of phishing emails would be conducive to the accuracy and efficiency of phishing detection research. In addition, the research on phishing email characteristics still has limitations to address. The psychological principles and mechanisms still need further research (Williams & Polage, 2019; Wang & Lutchkus, 2023). Increasing use of artificial intelligence appears to make it harder to detect phishing and scam emails (Zoltan & Weigand, 2023). Therefore, the impact of artificial intelligence (AI) on phishing detection is an increasingly important topic to research into.

The detection of phishing attempts through AI, ML, and LLM technologies forms the foundation of recent cybersecurity advancements. These technologies enhance defenses against phishing, improve training to recognize such threats, and identify sophisticated attacks. This section summarizes key findings from previous studies using AI models, machine learning algorithms, and Large Language Models like OpenAI's GPT, which are proficient in text classification and spam detection. The studies demonstrate these technologies' capabilities in mitigating the risks associated with phishing attacks effectively. Table 2 below summarizes research efforts and their contributions to phishing detection.

Table 2: Research on Detecting Phishing

Authors	Theory or Model	Method	Key Findings
Loh et al. (2024)	Generative AI models	Case Study	- AI can be used to help bolster cybersecurity defenses - AI models can make efficient phishing awareness training - Technology can be progressively trained to detect sophisticated phishing emails
Nguyen et al. (2024)	AI/ML algorithms	Case Study	- Using AI/ML algorithms to detect phishing in early stages has proven to be most effective in reducing negative effects caused
Trad & Chehab (2024)	Using LLMs for phishing detection	Experiment	- Fine tuning methods in LLMs have proven to be more accurate in phishing detection compared to prompt engineering
Uddin & Sarker (2024)	LLMs and AI	Experiment	- LLMs such as OpenAI's GPT can comprehend and produce text like a human - LLMs have become increasingly valuable in text classification and generation, especially in spam and phishing detection
Jiang	LLM detection	Experiment	- Effectiveness of LLMs can vary depending on nuances and the complexity of the text that is being analyzed - Building an effective LLM detector requires ongoing refinements and adaptations to counter new techniques
Heiding Schneier et al. (2024)	Detecting phishing emails using LLMs	Experiment	- Four LLMs were used to detect phishing, to compare the LLM results to human detection - LLMs demonstrated a strong ability to detect malicious emails, even in emails where phishing was not obvious - LLMs sometimes surpassed human detection, but were slightly less accurate than humans overall
Garg & Jayanthiladevi (2023)	Use of AI in cybersecurity	Case Study	- AI-powered systems offer robust capabilities to identify and mitigate risks in real-time

Authors	Theory or Model	Method	Key Findings
			- AI algorithms can enhance security operations, reducing the likelihood of a successful cyber attack
Safi & Singh (2023)	Detection of phishing websites	Case Study	- Machine learning algorithms have been applied the most in phishing detection techniques
Hazell (2023)	Social engineering and LLMs	Experiment	- Social engineering is a popular attack method for carrying out phishing attacks - Many versatile LLMs have been built in recent years, with the ability to tackle sophisticated capabilities like detecting or creating phishing content
Basit et al. (2020)	Phishing attack detection techniques	Survey	- AI can detect spam, phishing, skewers phishing, and other attacks utilizing previous information from datasets - Machine learning, deep learning, scenario-based, and hybrid techniques have been deployed to detect phishing - Machine learning methods are the most frequently used and most effective to detect a phishing attack

AI, ML, and LLM technologies play a critical role in efforts to combat phishing. These technologies are employed extensively to improve phishing awareness, detect sophisticated emails, and enhance overall cybersecurity measures. Machine learning methods and LLMs like OpenAI's GPT are helpful for their proficiency in text classification and spam detection. These AI-powered systems provide robust capabilities for risk identification and mitigation in securing digital communications against phishing attempts.

Theoretical Model

General phishing indicators

Based on reviews of prior research on the characteristics and tactics of phishing emails, this study adopts a set of key phishing indicators for testing and detecting phishing emails from valid emails. Using the same set of indicators for research testing will contribute to the consistency of the metrics of the testing and to the reliability of the test results. Table 3 below lists and defines the adopted phishing indicators.

Table 3: Phishing Indicators Defined

Indicators	Definitions and Descriptions
Visual Cues	Visual cues like unusual or suspicious logos, symbols, subject lines, sender's address, recipients, URLs, etc. can be telltale signs of phishing scams (Jiang, 2024; Murtaza, Pak, & Siddiqi, 2022; Zoltan & Weigand, 2023).
Social Proof	This is a psychological tactic of phishing emails based on the influence principle of "Truths are Us" (Cialdini, 2007, p. 87). This indicator is often appeals to people's tendency to follow the suit or comply with the social majority or group behavior as the benchmark or reference (Murtaza et al.,

Indicators	Definitions and Descriptions
	2022; Wang & Lutchkus, 2023).
Appeal to Authority	Appeal to authority is a common phishing tactic and indicator based on the psychological influence principle of "Directed Deference" (Cialdini, 2007, p. 157). This principle is to appeal to human and social tendency to obey people or organizations in authoritative positions with implied penalty for disobedience (Wang & Lutchkus, 2023).
Scarcity and Urgency	This is a common phishing tactic and indicator based on the psychological influence principle of "The Rule of the Few" (Cialdini, 2007, p. 178). This persuasion tactic is to influence the decision-making abilities of the victim by appealing to one's feeling of more value to things and opportunities with limited availability, urgency, and a feeling of fear or panic for possible loss or missing out (Murtaza et al., 2022; Wang & Lutchkus, 2023).
Linguistic Cues	Spelling mistakes, punctuation errors, incorrect wording, misuse of capitalization and other formatting issues could indicate phishing scams that often originate from international sources (Jiang, 2024; Murtaza et al., 2022; Zoltan & Weigand, 2023).

The five categories of phishing indicators above capture major and comprehensive aspects of the common characteristics of phishing emails for detection. The visual cues are the most obvious signs that can be easily detected. The categories of Social Proof, Appeal to Authority, and Scarcity and Urgency are primarily psychological tactics targeting people's cognitive, social and emotional attributes and vulnerabilities in order to influence their decisions. The Linguistics Cues category is included in the model to help with phishing detection. Even though software tools may help to create stronger phishing emails with less linguistic errors, this category of indicators is important for this research study to compare the effectiveness in phishing detection by humans and artificial intelligence tools.

Psychological theories and specific phishing indicators

In this study, we utilize a phishing email dataset from recent private email inboxes, reflecting the current tactics of brand and clone phishing. This choice ensures that the AI models do not rely solely on previously encountered data or direct comparisons, thereby pushing the boundaries of their capabilities to process novel information in a manner similar to human cognitive processing. To refine our approach, we have adjusted the set of phishing indicators used. The indicators of "Social Proof" does not apply in this study; and "Appeal to Authority" in Table 3 were found to be inherently present within the brand and clone phishing emails. Therefore, this study focuses on visual cues, scarcity and urgency, and linguistic cues. This refinement ensures that the indicators are more aligned with the specific types of phishing emails analyzed, enhancing the accuracy and relevance of our findings. By strategically selecting these indicators, we aim to provide a detailed and nuanced analysis of how AI can detect sophisticated phishing attempts, further advancing our understanding of AI's capabilities in cybersecurity.

Anti-Phishing efforts and AI's role

Effective anti-phishing measures such as Sender Policy Framework (SPF), Domain Keys Identified Mail (DKIM), Domain-based Message Authentication, Reporting, and Conformance (DMARC), are crucial for verifying the authenticity of email senders and detecting spoofed or forged emails (Staff, 2023). These protocols, which are standard across major email providers like Google and Yahoo, compare metadata and

email headers with known threats. Despite their effectiveness, these protocols struggle with identifying sophisticated phishing content associated with frequently changing phishing sites, often due to the low cost of domain name registrations (APWG, 2024).

Large Language Models in Phishing Detection

Large Language Models (LLMs), such as ChatGPT, are at the forefront of AI research and application, especially in natural language processing (NLP). Built on the transformer architecture, these models learn to predict and generate human-like text by training on vast datasets of diverse internet content (IBM, 2023). Unlike rule-based AI models, LLMs can parse and understand nuanced language and perform some tasks without dedicated prior training, saving time and resources (Ray, 2023), This flexibility allows LLMs to overcome the limitations of rule-based AI detection systems, which rely heavily on recognizing known phishing texts, domain names, and other typical content associated with phishing sites. LLMs' advanced capabilities enable it to detect subtle cues of phishing that may evade traditional detection systems. This includes understanding the context of an email conversation, assessing tone, and evaluating the intent behind the text, which makes it potentially capable of identifying novel phishing attempts.

This large language model, such as ChatGPT-4, demonstrates the ability to understand human text and respond with a realism that makes it hard to distinguish between human and machine. According to Fredrik Heiding et al. (2024), phishing emails that combine the capabilities of ChatGPT-4 with human-designed tactics achieve click-through rates between 43-81%, significantly surpassing the 18-69% rates of traditional methods. These findings align with research by Karanjai (2022) and Nov et al. (2023), indicating that ChatGPT can generate messages indistinguishable from those created by humans, effectively deceiving recipients and delivering services at a level comparable to human providers. Furthermore, in Heiding's study, Claude, a type of large language model (LLM) developed by Anthropic, demonstrated the ability to detect AI-generated phishing attempts with 100% accuracy when primed. However, its effectiveness dropped to 75% against typical real-world emails, highlighting a gap in handling everyday threats. This high success rate under primed conditions could imply that Claude's decision-making relies on comparing tasks with known data. These insights underscore the potential of LLMs to enhance cybersecurity, particularly in phishing detection, and reveal limitations that this study aims to address by employing real-world phishing examples without priming cues.

This study employs two advanced LLMs, ChatGPT-4 and Gemini Advanced, to test their effectiveness in a similarly uncontrolled, real-world setting. By using these sophisticated models, we aim to further understand the capabilities of LLMs in detecting varied and complex phishing attempts, specifically focusing on those that do not rely solely on previously encountered scenarios. This approach will provide a deeper insight into how these models perform under conditions that more closely mimic the challenges encountered in everyday cybersecurity operations.

Methodology

Measure effectiveness of identify phishing emails

To measure the effectiveness of AI models in identifying phishing emails, a quantitative methodology is employed, focusing on metrics such as accuracy, precision, recall, and F1 scores. The F1 score is selected as the primary metric because it balances precision and recall, making it particularly relevant for classification model evaluations (Yacouby & Axman, 2020). This approach provides a focused and quantifiable assessment of the AI models' performance in detecting phishing attempts, aligning with established research methodologies (Harikrishnan et al., 2018; Mardiansyah & Surya, 2024).

Accuracy measure
Accuracy measures the overall correctness of the model's predictions, calculated as the ratio of the sum of true positives and true negatives to the total number of predictions made. True Positives (TP) represent phishing emails correctly identified as phishing, while True Negatives (TN) are legitimate emails correctly identified as not phishing.
Accuracy = (True Positives + True Negatives) / Total Predictions

Precision measure
Precision measures the proportion of true positive predictions out of all positive predictions made by the model. It focuses on the accuracy of positive predictions. False Positives (FP) represent legitimate emails incorrectly classified as phishing.
Precision = True Positives / (True Positives + False Positives)

Recall measure
Recall, also known as sensitivity, measures the proportion of true positive predictions out of all actual positive instances in the dataset. It focuses on the model's ability to correctly identify positive instances.
Recall = True Positives / (True Positives + False Negatives)

F1 Score
The F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is particularly useful when dealing with imbalanced datasets because it considers both false positives and false negatives.
F1 score = (2 x Precision x Recall)/ (Precision + Recall)

Dataset for testing

Due to concerns about the real-time accessibility of Gemini and the premium version of ChatGPT-4, which now have the capability to browse the internet ("ChatGPT users…", 2023), there is a risk that these AI models may inadvertently identify phishing emails from search results and shared information. As a precautionary measure, the study deliberately avoids using public phishing email databases. Instead, all phishing emails were collected between February and April 2024 using individual email accounts to minimize the exposure of publicly accessible data to the AIs. Additionally, only commercial emails will be chosen as legitimate emails to reduce privacy risks and copyright issues. While this approach may not eliminate the risk, it represents the best strategy available to mitigate it. Ideally, conducting continuous testing of the AIs on a daily basis over a specified period and comparing the means of the F1 scores for these two models would provide more accurate data. However, due to constraints on data collection quantity, accessing a large number of phishing emails in real-time is not feasible. This limitation is acknowledged as an inherent challenge in conducting comprehensive AI testing in a dynamic phishing landscape.

For the assessment, ChatGPT-4 and Gemini Advanced were used to evaluate a total of 23 brand and cloning phishing emails collected during the study period, along with five legitimate emails. To simulate real-world scenarios and account for human factors and public literacy levels, the emails were saved in PDF format, displaying only essential headers such as sender, receiver, timestamps, and subject lines. This format mimics the real-world situations in which typical users interact with emails.

Prompting AI models
Both ChatGPT-4 and Gemini Advanced were prompted to provide feedback using the following structured format to ensure a detailed and specific evaluation:

"Please help to identify if the email is phishing email or legitimate email and reply as the format 'I read the email titled "filename". I believe it is a phishing (or legitimate) email. The subject is [describe]. The content is [describe]. The rationale behind my judgment is [one sentence].'"

Steps for Testing AI Models

1. Initial Email Set: A total of three emails were provided to both ChatGPT-4 and Gemini Advanced simultaneously for evaluation using the same prompt. This initial set tests the AI models' basic ability to discern phishing attempts.
2. Upload File Method: The three phishing emails were initially presented using an upload file method to maintain the format and information integrity of the emails. This method tests the AI's ability to evaluate content as it would appear in a typical email environment.
3. Testing with Limited Information: Following the initial assessment, the rest of emails was then provided without header information using a screenshot method. This step challenges the AI to detect phishing based solely on content, similar to how individuals with lower digital literacy might overlook technical cues like email headers.
4. Contextual Reassessment: If the initial assessments by the AI models are incorrect, the emails are resubmitted with headers included. This step was conducted to observe whether the inclusion of additional contextual information would help the AI models correct their initial judgments. This situation occurred twice:
    - ChatGPT-4: Made false negative errors twice, initially failing to identify phishing emails and correctly once with header information.
    - Gemini Advanced: Incorrectly identified a phishing email as legitimate once.

Results and Discussion

Detection with header information

Both AI models, ChatGPT-4 and Gemini Advanced, demonstrated strong capabilities in identifying suspicious header information, such as misaligned sender email addresses or URLs unrelated to the claimed companies. In tests involving five emails, three phishing and two legitimate, all with header information, both models successfully detected phishing emails by visual cues. However, both models failed to recognize one sophisticated phishing email that contained embedded malicious links. These links were disguised using subdomains or alterations in the URL to mimic legitimate company addresses. The deceptive use of such URLs presents a significant challenge, as it exploits the trust in recognized brand names to mislead recipients, including both human users and AI systems tasked with detecting such threats.

Detection without header information

When tested with 20 emails lacking header information (including 5 legitimate), the AI models continued to perform effectively, as summarized in Table 4. ChatGPT-4, however, showed slightly lower recall than Gemini Advanced, indicating a potential gap in identifying some phishing emails without header cues.

Table 4: Performance of AI Models Without Header Information

Assessment	ChatGPT-4	Gemini Advanced
Accuracy	90%	90%
Precision	100%	100%
Recall	87.5%	93%
F1	93%	96%

Despite the absence of headers, both models maintained high accuracy and precision. However, the recall and F1 scores highlight slight differences in their ability to detect phishing attempts, with Gemini Advanced slightly outperforming ChatGPT-4.

Accuracy (90% for both models)
Accuracy is a general measure of how often the model is correct. Both AI models correctly identified 90% of the emails, whether they were phishing or legitimate.

Precision (100% for both models)
Precision measures the accuracy of the positive predictions (phishing emails in this context), indicating no false positives. Both models had a precision of 100%, meaning that every email they identified as phishing was indeed a phishing email.

Recall (87.5% for ChatGPT-4, 93% for Gemini Advanced)Recall, also known as sensitivity, measures the ability of the models to correctly identify all phishing emails. ChatGPT-4 identified 87.5% of the actual phishing emails, while Gemini Advanced identified 93%. This suggests that Gemini Advanced is slightly better at detecting phishing emails and is less likely to miss them compared to ChatGPT-4.

F1 Score (93% for ChatGPT-4, 96% for Gemini Advanced)
The F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics. ChatGPT-4 achieved an F1 score of 93%, while Gemini Advanced achieved 96%. This indicates that Gemini Advanced has a slightly better overall performance in balancing precision and recall, making it more reliable in detecting phishing emails while minimizing both false positives and false negatives.

Analysis of AI assessment reasoning

Both AI models provided their reasoning for assessment. For legitimate emails, ChatGPT-4 based its assessments on the email's professional layout, detailed content, genuine requests, and helpful advice. In contrast, Gemini Advanced focused on legitimate email addresses, official information, and reasonable promotions. Gemini Advanced more frequently compared emails to common business practices and noted spelling and grammar mistakes. ChatGPT-4 did not identify any grammar mistakes but indicated other phishing tactics such as offering unrealistic high-value prizes, using urgent tones, and requesting immediate responses, emphasizing suspicion on offers that seem "too good to be true." Tables 5 and 6 illustrate the reasoning used by ChatGPT-4 and Gemini Advanced, respectively, in their assessments on phishing emails.

Table 5: ChatGPT-4 Assessment Indicators

Indicators (Rational)	count
Seek for sensitive information	7
Visual cue	6
Urgent and fear	3
Too good to be true	2

Table 6: Gemini Advanced Assessment Indicators

Indicators (Rational)	Count
Visual cue (others)	6
Visual cue (compare to legitimate practice)	6
Linguistic cues	4
Too good to be true	2

Table 5 shows that ChatGPT-4 frequently uses criteria such as the purpose of emails, specifically identifying attempts to solicit sensitive information from recipients, alongside visual cues, urgency, and suspicion of offers that are too good to be true. Table 6 indicates that Gemini Advanced places emphasis on visual cues, particularly by comparing them to legitimate practices, and also notes linguistic cues like spelling and grammar mistakes. This comparison highlights different strengths in their assessment approaches: ChatGPT-4 focuses more on the content's plausibility, while Gemini Advanced utilizes a broader range of cues, including linguistic accuracy.

Both ChatGPT-4 and Gemini Advanced demonstrated strong capabilities in detecting phishing emails using the defined indicators, with slight differences in their approaches and performance metrics. With header information, both models achieved perfect scores across all metrics. Without header information, they maintained high performance, though Gemini Advanced showed a slightly higher recall rate. Recall rate, or sensitivity, measures the ability of the model to correctly identify all relevant instances, indicating that Gemini Advanced may miss fewer phishing emails than ChatGPT-4. The detailed rationales for their assessments align with the defined phishing indicators: ChatGPT-4 focused on content plausibility and typical phishing tactics, while Gemini Advanced emphasized visual and linguistic cues and comparisons to legitimate practices. However, both models failed to detect a highly sophisticated phishing email that embedded invisible malicious links, highlighting a limitation in handling advanced phishing techniques.

Detailed Analysis of Specific Cases

Successful Detection by Gemini Advanced

One notable instance where Gemini Advanced succeeded while ChatGPT-4 initially failed involved an email allegedly from ZippyLoan. Without header information, ChatGPT-4 assessed the email as legitimate, citing its structured layout and professional design typical of financial institutions. In contrast, Gemini recognized the scam, noting the email's offer of a large sum of money with a quick application process, a common phishing tactic. Upon a second review with header information, ChatGPT-4 adjusted its judgment, identifying the email as phishing based on an unusual greeting, aligning its assessment with Gemini Advanced.

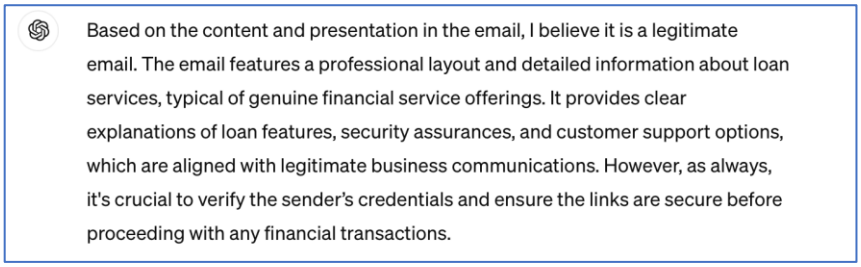

Figure 1: Initial Response from ChatGPT-4 on the ZippyLoan Email

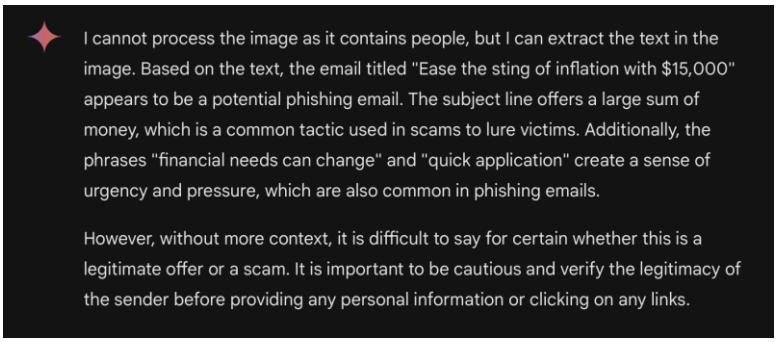

Figure 2: Initial Response from Gemini Advanced on the ZippyLoan Email

Detection Failures in Advanced Phishing Techniques

Both AI models struggled with a highly convincing cloning phishing email, which both failed to identify as malicious. ChatGPT-4 classified the email as legitimate, noting the absence of typical phishing red flags such as urgent personal information requests or poor grammar. Similarly, Gemini Advanced validated the email as coming from a legitimate Danaher source without requesting sensitive information. Unfortunately, both models missed the embedded malicious links, highlighting a critical area for improvement in detecting sophisticated phishing techniques.

Conclusion

Study overview and key findings

Our study originated by using specific prompts that required AI models to provide explicit assessments rather than probabilities of phishing likelihoods, which are typical in existing research. This approach led to more decisive responses from the AI models, enhancing their utility for real-world applications. Unlike previous studies where AI might provide hesitant responses (e.g., 80% likelihood), or refuse to provide definitive answers, our study prompted AI to make judgments based on content analysis. For example, Gemini Advanced, while occasionally refusing to provide assessments citing its inability to "read" humans, still offered evaluations based on content clues. The phishing emails in our study specifically targeted popular brands. Unlike humans, who might be influenced by the reputation or perceived credibility of these brands, both AI models assessed the emails purely based on detectable indicators. They relied predominantly on visual cues and focused on the intent of the emails, such as requests for personal information and the presence of urgency tones. This objective approach allows AI models to evaluate potential threats based on content and structure without the biases and emotional responses that might affect human recipients.

Performance and limitations

Both AI models, ChatGPT-4 and Gemini Advanced, demonstrated high precision in accurately identifying phishing emails without generating false positives, which is crucial for avoiding unnecessary alarms. However, the slightly lower recall rate for ChatGPT-4 suggests it might miss more phishing emails than Gemini Advanced. Notably, both models struggled with highly sophisticated phishing emails that included invisible malicious links, indicating a limitation in handling advanced phishing techniques. Despite these challenges, Gemini Advanced marginally outperformed ChatGPT-4 in terms of recall and F1 score, suggesting it may be slightly more reliable in real-world scenarios.

The study acknowledges several limitations that may impact the generalizability and robustness of the findings. Particularly, the constraints in the volume of email datasets used for testing have limited our ability to capture a broader range of phishing attempts in real-time. These limitations suggest the need for further research with enhanced access to larger and more varied email datasets and more frequent testing to more accurately assess the performance of AI models in phishing detection.

While our findings demonstrate the potential of LLMs in detecting phishing emails, it is important to consider broader challenges inherent in AI applications that could influence theses outcomes. In addition to the primary findings, we recognize that underlying issue in AI development, such as model biases and occasional incorrect responses (hallucinations), could potentially impact the performance of LLMs in phishing detection tasks. Although not the central focus of the study, future research could explore robust training techniques to minimize biases and enhance the reproducibility of detection results, thereby increasing the reliability of LLMs in practical cybersecurity applications.

Broader implications and future direction

Phishing detection is a cat-and-mouse game, where technological solutions must continually evolve to keep pace with sophisticated phishing strategies that frequently bypass conventional detection methods (Basit et al., 2020). While technological defenses like firewall filters are prevalent, new phishing strategies continually emerge, often bypassing conventional detection methods. The role of Large Language Models (LLMs) as potential content filters could be particularly beneficial, especially for protecting populations with limited digital literacy, by simplifying the detection process and providing more understandable and actionable warnings. However, deploying LLMs involves challenges such as substantial computational demands, privacy concerns, and potential misuse in crafting sophisticated phishing attacks. Future research should continue to evaluate LLMs against a diverse range of phishing emails and compare their performance to human detection, to better understand the strengths and limitations of LLMs in the dynamic landscape of email security.

References

APWG. (February 13, 2024). *Phishing Activity Trends Reports:4th quarter 2023*. Apwg.org. https://apwg.org/trendsreports/

Basit, A., Zafar, M., Liu, X., Javed, A. R., Jalil, Z., & Kifayat, K. (2020). A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommunication Systems*, *76*(1). https://doi.org/10.1007/s11235-020-00733-2

Cialdini, R. B. (2007). *Influence: The psychology of persuasion*. HarperCollins e-books.

ChatGPT users can now browse internet, OpenAI says. (2023, September 27). *Reuters*. https://www.reuters.com/technology/openai-says-chatgpt-can-now-browse-internet-2023-09-27/

Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J., & Park, P. S. (2024). Devising and detecting phishing emails using large language models. *IEEE Access*, 1–1. https://doi.org/10.1109/access.2024.3375882

Garg, R., & Jayanthiladevi. (2023). Preventing cyber attacks using artificial intelligence. I-Manager's Journal on Software Engineering, 18(2), 1-9. doi:https://doi.org/10.26634/jse.18.2.20367

Harikrishnan, N. B., Vinayakumar, R., & Soman, K. P. (2018, March). A machine learning approach towards phishing email detection. In *Proceedings of the anti-phishing pilot at ACM international workshop on security and privacy analytics (IWSPA AP)* (Vol. 2013, pp. 455-468).

Hazell, J. (2023) Spear Phishing with Large Language Models. ArXiv.org. https://doi.org/10.48550/arXiv.2305.06972

IBM. (2023). *What are large language models?* www.ibm.com. https://www.ibm.com/topics/large-language-models

Jiang, L. (2024). Detecting scams using Large Language Models. ArXiv.org. https://doi.org/10.48550/arXiv.2402.03147

Kalla, D., & Smith, N. (2023). Study and analysis of Chat GPT and its impact on different fields of study. *International Journal of Innovative Science and Research Technology*, 8(3).

Karanjai, R. (2022). Targeted phishing campaigns using large scale language models. arXiv preprint arXiv:2301.00665.

Kleitman, S., Law, M. K. H., & Kay, J. (2018). It's the deceiver and the receiver: Individual differences in phishing susceptibility and false positives with item profiling. *PLoS One, 13*(10), 1-29. https://doi.org/10.1371/journal.pone.0205089

Koide, T., Fukushi, N., Nakano, H., & Chiba, D. (2024). ChatSpamDetector: Leveraging large language models for effective phishing email detection. *ArXiv (Cornell University).* https://doi.org/10.48550/arxiv.2402.18093

Loh, P. K. K., Lee, A. Z. Y., & Balachandran, V. (2024). Towards a hybrid security framework for phishing awareness education and defense. Future Internet, 16(3), 86. doi:https://doi.org/10.3390/fi16030086

Mardiansyah, K., & Surya, W. (2024). Comparative Analysis of ChatGPT-4 and Google Gemini for Spam Detection on the SpamAssassin Public Mail Corpus.

Murtaza, A. S., Pak, W., & Siddiqi, M. A. (2022). A study on the psychology of social engineering-based cyberattacks and existing countermeasures. *Applied Sciences, 12*(12), 6042. https://doi.org/10.3390/app12126042

Nguyen, V., Wu, T., Yuan, X., Grobler, M., Nepal, S., & Rudolph, C. (2024, February 26). *A pioneering study and an innovative information theory-based approach to enhance the transparency in phishing detection*. ArXiv.org. https://doi.org/10.48550/arXiv.2402.17092

Nov, O., Singh, N., & Mann, D. (2023). Putting ChatGPT's Medical Advice to the (Turing) Test: Survey Study. *JMIR Medical Education, 9*(1), e46939.

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*(1), 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected?. *arXiv preprint arXiv:2303.11156*.

Safi, A., & Singh, S., A systematic literature review on phishing website detection techniques, *Journal of King Saud University - Computer and Information Sciences*, Jan. 2023, doi: https://doi.org/10.1016/j.jksuci.2023.01.004.

Staff, T. (2023). TestOut CyberDefense Pro.

Tang, R., Chuang, Y.-N., & Hu, X. (2024). The Science of Detecting LLM-Generated Texts. Communications of the ACM, 67(04). https://doi.org/10.1145/3624725

Trad, F., & Chehab, A. (2024). Prompt engineering or fine-tuning? A case study on phishing detection with large language models. *Machine Learning and Knowledge Extraction, 6*(1), 367. doi:https://doi.org/10.3390/make6010018

Uddin, M. A., & Sarker, I. H. (2024). An explainable transformer-based model for phishing email detection: A large language model approach. *ArXiv (Cornell University).* https://doi.org/10.48550/arxiv.2402.13871.

Vishwanath, A. (2022). The weakest link: How to diagnose, detect, and defend users from phishing. MIT Press.

Wang, P., & Lutchkus, P. (2023). Psychological tactics of phishing emails. Issues in Information Systems, 24(2).

Wash, R. (2020). How experts detect phishing scam emails. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW, Article 160 (October 2020), 28 pages. https://doi.org/10.1145/3415231

Williams, E., & Polage, D. (2018). How persuasive is phishing email? The role of authentic design, influence and current events in email judgements. *Behavior & Information Technology*, *38*(2), 184-197. https://doi.org/10.1080/0144929X.2018.1519599

Yacouby, R., & Axman, D. (2020, November). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In Proceedings of the first workshop on evaluation and comparison of NLP systems (pp. 79-91).

Zoltan, K. A., & Weigand, L. (2023). From Provoking Emotions to fake Images: The recurring signs of fake news and phishing scams spreading on social media in Hungary, Romania and Slovakia. *Proceedings of the 22nd European Conference on Cyber Warfare and Security, ECCWS 2023.* 726-732.