2025 12th International Conference on Future Internet of Things and Cloud (FiCloud) | 979-8-3315-5437-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/FiCloud66139.2025.00026

Utilising Generative AI To Improve Practical Spear Phishing Training

Charlie Green
University of Gloucestershire
Cheltenham, United Kingdom
charliegreen@connect.glos.ac.uk

Abu Alam
University of Gloucestershire
Cheltenham, United Kingdom
aalam@glos.ac.uk

***Abstract*—Despite current anti-phishing measures, businesses still fall for phishing attacks, this issue stems from modern phishing attacks leveraging artificial intelligence. To address this issue, this research proposes the use of an automated pipeline to generate spear phishing emails for use by internal security professionals in practical spear phishing exercises. The proposed pipeline is automated, capable of using real employee data, is fully customisable with modular components, accepts any input data with fine-tuneable outputs, and is portable, as it can be accessed through the web. The pipeline was well received by businesses with generally positive feedback on the effectiveness of the pipeline and potential use cases. The generated emails were of good quality and of 60 responses, on average, the participants rated the emails a score of 4.11/5 or "Very Accurate" and "Likely" to interact with the content of the emails, with an average score of 3.02/5. The pipeline took an average of 17.5 seconds per email. Using this number, it would take a business of 100 employees 29.17 minutes or a larger business with 1000 employees 4.86 hours. This level of speed and effectiveness showed that the pipeline could serve as an effective training tool for businesses.**

***Index Terms*—Phishing, Spear phishing, Machine learning, Generative AI, AI, Generative models, Large language model, LLM, Awareness training, Human factor, URL detection, Phishing detection, Pipeline, Business**

I. Introduction

A. Overview

As technology evolves, innovations such as generative artificial intelligence (AI) have drastically improved the realism of computer-generated language. This realism has been exploited by bad actors to improve phishing attacks. Using generative AI, these bad actors are creating highly accurate, very specific phishing emails that can fool employees, even with current phishing training. Barracuda, a reputable network security business, has gone as far as saying 'No business is immune' to these attacks (Barracuda, 2023). This type of phishing email has existed for a while, but the low barrier to entry that generative AI provides allows individuals with little subject knowledge to perform these attacks. Modern phishing filters, such as the Kamran, Sengupta, and Tavakkoli, 2022 model, cannot yet fully mitigate this issue with cyber security companies, including KnowBe4, stressing the importance of awareness training, highlighting that more consistent training provides better average results against phishing attacks (Grimes et al., 2023). As such, this research focuses on the use of a generative approach to improve practical spear phishing awareness for businesses.

B. Research Aim and Objectives

This research aims to enhance organisational resilience against spear phishing by investigating how generative AI can be used to simulate and counter such attacks effectively in training environments.

To achieve this, the study sets out to:

1) Analyse current strategies and tools used by businesses to mitigate spear phishing threats.
2) Design and implement a semi-automated pipeline that leverages generative AI to produce realistic spear phishing scenarios.
3) Evaluate the effectiveness, speed, and accuracy of the developed system in practical, business-oriented training exercises.

This research does not include the distribution of generated emails, as this is not required for the evaluation of the pipeline. Participant responses may have an inherent bias from prior awareness training. Similarly, all study participants are informed beforehand, which may lead to bias. All statistics for pipeline evaluation are limited by the system specifications.

II. Related Work

A. Phishing and Spear Phishing

Ayeni et al. defines phishing as a form of social engineering in which an attacker will deceive a user into disclosing sensitive information by impersonating a reputable entity; this type of attack is most commonly performed by email, but there are other vectors (Ayeni, Adebiyi, Okesola, and Igbekele, 2024). Common indicators of a phishing email include generic greetings without specifying a name, a problem that needs to be resolved, a sense of urgency, or a call to action where the user is required to do something to solve the issue. Over time, generic, high-spread, and low-yield emails have evolved into more sophisticated attacks such as Business Email Compromise (BEC) attacks. These attacks are highly targeted and very specific, with research into the business and employee resulting in low-spread and high-yield attacks despite current countermeasures. The 2024 OPSWAT report specifies that up to 80% of businesses have fallen victim to phishing within the year (OPSWAT, 2024). This highlights the threat that phishing still poses despite the long history of the threat and the importance of evolving countermeasures to match the evolution of the attack.

Adapted from Allodi et al. spear phishing, unlike regular phishing, is a highly targeted context-specific attack targeted at a specific entity and is generally characterised by a multi-stage process where the attacker collects information on the target which is then used in the creation of the spear phishing email (Allodi, Chotza, Panina, and Zannone, 2020). Normally including the victim's name or other personal information, the personalised nature of spear phishing emails makes them much more believable than regular phishing emails. Barracuda created a report in 2023 and discovered that half of the businesses contacted were victims of spear phishing, with an average cost of $5 million (Barracuda, 2023). BEC attacks have been highlighted by literature as an increasingly sophisticated attack, presenting challenges to businesses around the world due to their ability to pass through regular phishing filters. Wasserman et al. also discusses the danger of these attacks as the malicious content of these attacks are entirely text-based without need for links or attachments bypassing traditional malicious URL detection software (Wassermann, Meyer, Goutal, and Riquet, 2023). They also discuss the different types of text-based attacks depicting emotions, targets, impersonation, and manipulation principles from Cialdini's principles of persuasion; see Table I.

Attack Type	Principles	Emotions	Targets	Impersonated
CEO Fraud	Authority, Scarcity, Commitment, Liking	Urgency, Fear	Finance	CEO
W-2 Fraud	Authority, Scarcity, Commitment, Liking	Pressure, Trust	HR, Payroll	Execs, Consultant
Gift Card Scam	Authority, Scarcity, Reciprocity, Liking	Empathy, Obligation	Managers, Receptionists	Executive
Payroll Fraud	Scarcity, Commitment, Liking	Confusion, Trust	HR, Payroll	Employee
Lawyer Fraud	Authority, Scarcity, Liking	Fear, Urgency	Finance, Execs	Lawyer, Executive
Rapport-Building	Commitment, Liking	Trust, Familiarity	Employees	Executive

TABLE I: Characteristics of text-based attacks. Adapted from Wassermann et al., 2023

B. Awareness training

A large part of cyber security is the human factor, with awareness training being a key factor in training the human factor. The importance of the human factor in businesses cannot be overstated, with the 2022 LastPass breach being a prime example. Sugunaraj analysed the breach using a cyber-human approach, concluding that the primary source of infection was the compromise of employee devices (Sugunaraj, 2024). This emphasises the fact that any single employee can be a source of infection for a whole business. One method businesses use to mitigate this threat is cyber awareness training: taking many forms such as guest speakers, online classes, or practical activities. All methods aim to inform and educate employees about cyber threats and how to respond to them. Similarly, a healthy security culture is paramount in businesses, as Sun et al. discussed the importance of reducing blame and responsibility for falling victim to phishing attacks, since accountability only leads employees to cover up the breach (Sun et al., 2024). Similarly, Damiano discusses the reasons why employees hide these breaches, mainly due to doubts about self-efficacy and the repercussions of falling for an attack (Damiano, 2020). However, a positive culture would promote higher report rates and overall improved mental health. Another form of awareness training is practical training, in which a business will test employees with simulated phishing emails, mimicking a real attack but doing no actual harm. KnowBe4 released a report in 2023 that confirmed the value of practical phishing training, with employees who participate in awareness training performing better on simulated attacks (Grimes et al., 2023). Despite the importance of awareness training, Proofpoint found that only 53% of businesses trained all employees, which decreased 3% from the year before (Proofpoint, 2024a). This shows a flawed methodology as the whole business acts as a chain only as strong as the weakest employee.

Another shortcoming of current awareness training is the lack of practical training for spear phishing. Although there are practical exercises for regular phishing, the time, staff, and monetary costs of creating spear phishing emails for all staff are too high. For practical spear phishing training to be plausible, a cheaper, automated method must be created.

C. The Role Of Generative AI and LLM In Phishing

One of the major innovations in the creation of all types of media, including phishing emails, is generative AI. In the 2024 State of the Phish report, Proofpoint stated that generative AI promises to improve the quality of social engineering attacks and concluded that there is now a likely link between generative AI and the increase in BEC attacks (Proofpoint, 2024a). It is important to note that generative AI can be used in creating realistic phishing emails with little effort; the barrier to entry for this type of attack has decreased rapidly with the accessibility of chatbots and the myriad of other types of generative AI tools. Large language models (LLMs) are one of the most common and easy ways to access generative AI, some of the biggest names including ChatGPT, Deepseek and Gemini. These LLMs excel at generating synthetic text, ranging from summarising information, answering questions, or, in the context of phishing, creating phishing emails. These LLMs are fully capable of realistic Natural Language Processing (NLP) allowing them to understand human language and making them capable of semantic analysis. Proofpoint specifically named LLMs and ChatGPT as a tool that can be abused by bad actors to generate phishing emails, correlating the link between the launch of ChatGPT and the global increase in BEC attacks (Proofpoint, 2024b). These chatbots have become harder to tell apart from humans. A good example for this is the game "Human or Not?" in which a user is given a small amount of time to chat with an AI or a human and then guess which they were speaking to (AI21labs, 2023). The creators of this game released a paper discussing the accuracy of players with only 68% guessing correctly against a human and 60% against an AI, which they described as only slightly better than flipping a coin (Jannai, Meron, Lenz, Levine, and Shoham, 2023). This is a good simulation of phishing emails, conversing with someone often with limited time to respond, and gives plausibility to people believing they are talking with a real human when actually it is AI generated. It is important to compare the different LLM models, including the advantages and disadvantages of each to decide which is best for the generation of spear phishing emails; see Table II.

Model	**Local**	**Advantages**	**Disadvantages**
ChatGPT	No	Well-tested, multiple models, versatile input methods	No local version, API access requires payment
LLaMA	Yes	Open-source, strong community, high performance	High resource usage
DeepSeek	Yes	Multiple models, small parameter options	Response limitations, relatively new
Mistral	Yes	Scalable, strong performance	Limited community support
Gemini	Browser-only	User-friendly, Google integration	Only usable via browser
Grok	Yes	Intuitive interface	Very high system requirements

TABLE II: Comparison of LLM Models

In addition to traditional models, there are also jailbroken models, which have had their decoding probabilities modified to allow the generation of harmful content (Zhao et al., 2024). This harmful content is not always malicious, just outside the model's normal content filters. When a model is jailbroken, it is usually to bypass these filters, which can also allow for the generation of phishing emails.

D. Shortcomings of Automated Measures

The current approach to mitigating phishing attacks is focused heavily on automated measures, with much less focus on the human factor. These automated measures are comprised of machine learning and generative models that filter emails, mostly through malicious URL detection, such as the model by Islam et al. boasting accuracy rates of 99.85% (Islam et al., 2024). However, businesses still fall for phishing attacks. This is largely because most of these filters focus on malicious URL detection which fails against text-based attacks such as BEC attacks or other methods of bypassing detection such as replacing a URL with a QR code. Additionally, false positives can be catastrophic for businesses where urgent business emails may be flagged as malicious, resulting in missed deadlines or lost business opportunities. A systematic review from Thakur et al. revealed limitations in current deep learning detection methods, with a lack of privacy measures and email content analysis (Thakur, Ali, Obaidat, and Kamruzzaman, 2023). Specifically, these models handle user data and emails with little protections in place for improper access, potentially resulting in data breaches or other GDPR violations. Thakur also identified the possibility of multilingual attacks to bypass filters, where a model cannot read the email in another language, relying on the translation feature of gmail to translate back to the intended language. Other areas could be improved including the diversification of analysed areas for emails, focusing mostly on email structure while ignoring other factors such as the sender address. Databases exist of blacklisted email addresses such as Spamhaus or Cleantalk with sender scores which could be implemented into these systems to block bad actors.

It is evident that business still fall victim to spear phishing attacks, with current mitigations centred on automated measures and less on the human factor. The gap in the literature demonstrates the need for more accessible practical spear phishing training. For this to become accessible, the associated costs must be reduced, including time, money, and staff requirements for the generation of these emails. To fill this gap, this research proposes the use of an automated spear phishing email pipeline to generate emails quickly and automatically.

III. Methodology

The objectives of this implementation were to reduce the costs associated with practical spear phishing exercises such as time, staff, and money. To address these costs, the implementation had to be faster than existing methods, automated, and inexpensive. The most suitable solution that met these requirements was an automated pipeline. A pipeline is a system that takes information and performs as many different processes on the information in order and returns the fully processed data instead of performing each step separately, making a long process into one efficient process. The pipeline proposed in this research would take the details of a staff member, perform any necessary steps to process the details, and generate a spear phishing email. This would be accomplished by leveraging generative AI to flexibly accept any details and generate an email. Before implementing the pipeline, two key decisions were made. First, different types of generative AI models were compared to identify the most cost-effective option (see Table II). Based on this comparison, a large language model (LLM) was selected, as it offered the greatest flexibility, operated unsupervised, maintained manageable costs, and demonstrated higher speed compared to other generative AI models. The methodology followed to evaluate the implementation included evaluating both the pipeline itself and the generated emails. Evaluating the pipeline involved measuring how much it reduced the associated costs discussed previously; time, staff, and money. In addition, insights from both businesses and participants on the effectiveness of the pipeline as a training tool and its utility for a business. Similarly, generated emails are evaluated using participant responses, both qualitative and quantitative, to assess email quality and potential improvements. A list of all metrics is available in Table III. A feedback survey was created to gather qualitative and quantitative responses from participants with an initial pilot study conducted with a handful of participants, containing only five questions. Revisions were made with some questions being combined and additional questions being added, including text responses for better qualitative responses. This led to a refined survey with seven questions that gauged the accuracy and effectiveness of the emails produced by the pipeline. The survey had to be conducted in person, as a standard Google form would not allow integration with the pipeline easily, and using a public web server for this survey had the possibility of compromise, which infringes on GDPR and ethics. So, a local web server was designed for participants to use on a specific laptop in person.

IV. Design and Implementation

This section focusses on the design of the pipeline as described in the methodology. The pipeline is divided into 3 main modules; input, generation, and screening, with each major section being completely modular, using JSON for passing data. The whole solution should be created using Python with JavaScript for the web interface. Python was chosen because of well-documented libraries for interfacing with LLMs, locally hosted web servers, and connecting to SQL databases. The Web

Metric	Data	Description
Generation time	Quantitative	The time taken per character for the pipeline to generate an email.
Rate of failure	Quantitative	The percentage of times the model failed to generate an email for any reason.
Participant ratings	Quantitative	Ratings from 1-5 evaluating the generated email.
Participant feedback	Qualitative	Any comments participants gave regarding either the pipeline or the generated email.
Business insights	Qualitative	Comments from businesses regarding either the pipeline or the generated email.

TABLE III: Evaluation metrics

server will be built on FastAPI as it facilitates using Python scripts on input data and allows for easy passing of data between modules without storing the input details of participants.

A. *Input*

The input module should be able to take JSON, CSV files, and form data from a web interface. The pipeline should also be accessible from both a command line and a Web interface. To accomplish this, the input module will have two scripts, input and web_ui, where input will handle the CLI and web_ui will handle web interface requests. Similarly, both scripts will cover all mentioned input types and send all inputs to the same generator and handle the returned response appropriately. User's would input details such as name, job title, or company name and that data would be sent to the LLM to be used in the generation of a personalised spear phishing email. A web interface is available for a user-friendly method of uploading a CSV or entering details into a form. After taking an input, each individual set of details is sent to the LLM using Ollama. This model is run locally to ensure that no employee data is stored on cloud servers. The first model tested was the official 8B parameter Llama3.1 model (Touvron et al., 2023). However, the content filters of this model prevented it from generating spear phishing emails. As such some initial tweaking with the system prompt allowed the model to occasionally generate an email, however, the model would still reject generation 92% of the time. In response, a modified version of the Llama3.1 model was used to ease content rejection (rolandroland, 2024).

B. *Generation*

During the methodology, it was decided to use an LLM for the pipeline; however, a specific LLM model must be chosen for use in the pipeline. A suitable model for this implementation must be locally hostable to ensure full control over sensitive data and it must be supported by other software such as Ollama. Similarly, the model must be feasible to run on a standard system due to limitations on system specifications. However, a business may choose to use a larger model on better equipment to yield faster results, but this comes with higher running costs. From the LLM comparison table, see Table II, the potential candidates that fit these requirements are Llama, Deepseek, and Mistral. Deepseek will not be used as it is a relatively new model with significantly less community resources. This leaves Llama and Mistral, of which Llama is the better choice given both models have high performance, but Llama has more community resources, compared to Mistral's slightly better resource consumption, which was deemed an acceptable compromise.

C. *Screening Web Interface*

The screening page would usually be a simple yes/no as a final human check but has been repurposed in this research for the evaluation page. After uploading details to the input module, the user is redirected to this page which dynamically displays all generated emails and a feedback form below each generated email. Feedback data is stored in a local database. In a business setting, after the email had been screened, the pipeline would automatically send it to the employee. However, this research does not cover the distribution code as it is not required for evaluation and this research does not facilitate the distribution of spear phishing emails.

V. EVALUATION

To evaluate the pipeline, both quantitative and qualitative metrics were used, ensuring that both technical performance and participant feedback were measured. As such, this section will be split into two parts, the technical evaluation and the participant study.

A. *Technical evaluation*

All technical evaluation tests were conducted using CSV files of varying sample sizes, all containing four fields: name, job title, company name, and company sector for simplicity. Each sample dataset was used 10 times to calculate average metrics and varying sample sizes were used to measure performance under different loads simulating different-sized companies. Initially, the generation time of emails was compared between sample sizes to measure the average speed for different sample sizes, but the generation time varied more than expected; see Figure 1.

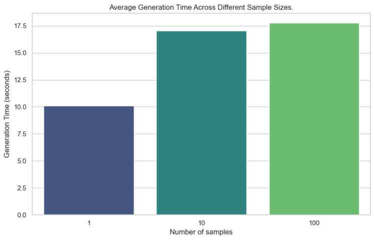

Fig. 1: Generation Time Across Sample Sizes

This variance led to the discovery of “cut” emails, where the email was incomplete and generated much faster as a result. This discovery meant that generation time was likely tied to character length. As such, another set of data was collected alongside the character length of each email and plotted against each other revealing a direct correlation between character length and generation time; see Figure 2.

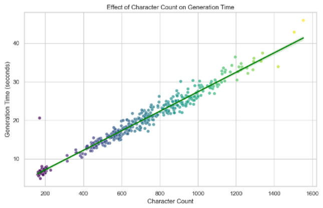

Fig. 2: Character Count Against Generation Time.

To mitigate future cut emails, measures were put in place to detect and regenerate cut emails up to three times to prevent hanging. A cut email is identified by a character length of less than 300, this number was determined from Figure 2 where a clear gap is visible between 200 and 400 characters, in which no emails are generated. To reflect the new speed metric, a new graph was created to replace the previous bar chart; see Figure 1.

B. Participant Study

For the participant study, a laptop with lower specs was used, which resulted in some inconsistencies compared to the output from the workstation, such as an increase in cut responses even after three regenerations. Of the 60 responses from participants, the pipeline had a fail rate of 8.6%, see Figure 4, likely caused by the inconsistencies. Additional checks or stronger hardware could be used to reduce this statistic. For the study, several questions with a Likert scale were asked, with the results converted to a 1-5 scale. From the question “How accurate is the email to the details provided?” the average rating was 4.11 or “Very Accurate” with lower rating having a common reason of the sender and recipient details being swapped, with the email being written as if it were coming from the participant instead of going to them. An improvement in the system prompt may mitigate this issue. Similarly, the question “How likely are you to interact with this email?” Received an average rating of 3.02, or “likely”. This lower score was likely caused by confirmation bias, due to the informed nature of the study with all participants understanding that the email was a phishing email beforehand.

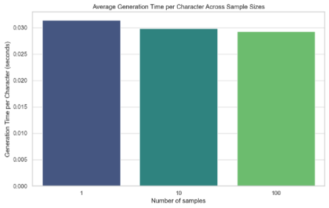

Fig. 3: Generation Time per Character Across Sample Sizes.

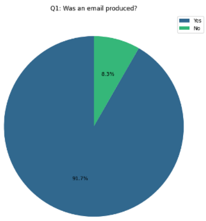

Fig. 4: Question 1, failure rate Pie Chart

Participants were also asked to identify the persuasive techniques used by the pipeline; see Figure 5. The most common techniques were scarcity, authority, and liking, with other techniques used half as frequently. This shows a bias to these techniques that could be reduced through fine tuning.

In addition to collecting responses, as a supplementary evaluation, 4 anonymous local businesses were contacted and given an in-person demonstration of the pipeline.

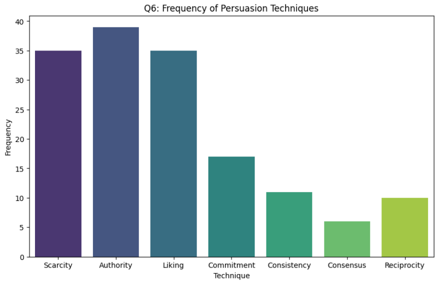

Fig. 5: Frequency of Persuasion Techniques

After which, they were asked for any business insight on the pipeline. The general sentiment was positive, with businesses seeing use for training remote workers or being offered as a service by cyber insurance. Additionally, use in unscheduled training exercises was discussed with value in the speed and effectiveness of unsupervised generation being a key sentiment. One business commented on the utility for smaller businesses due to the low setup and running costs with low minimum specifications. Another business discussed how it would be a good follow up activity after training days as a "top-up" activity. However, one major concern was discussed; generated emails could contain links and contact details that are not managed by the business such as "example.com" which would need to be replaced by managed details as part of the exercise to measure click rate.

C. *Limitations and Future Work*

Some limitations were discovered during the implementation and evaluation that should be addressed in future work. Firstly, the current pipeline is sequential, only generating one email at a time. Addressing this through the use of Ollama's concurrency features could rapidly speed up large-scale generation for businesses. Similarly, hardware limitations on both the laptop and the workstation used affected the speed and quality of the generated emails, with better hardware allowing for larger models and faster generation. If done again, real-world tests with uninformed employees would give more realistic data over informed participants. Similarly, only four businesses were contacted; as such, a larger number of businesses would allow greater insights and grouping of business sectors to analyse trends in the data. In future work, better models with better performance could be investigated. Similarly, the system prompt could be fine-tuned to reduce failure rate, and additional steps could be added to the pipeline to further refine emails including the identification and substitution of dummy details such as links and attachments with managed ones for accurate data collection from campaigns. With employee permission, automatic data exploration could be implemented to gather even more information, including recent social media posts to further tailor emails. Finally, incorporating this pipeline into existing systems such as cyber insurance or a security tools suite could see a real business impact in awareness training, improving cyber awareness in businesses.

VI. Conclusion

This study found that the novel pipeline is a feasible approach well accepted by businesses due to the low associated costs and the ability to connect remote workers to awareness training exercises. The generated emails were of good quality, with an average participant rating of 4.11 or "very accurate" and "likely" to interact with the contents of the emails or an average score of 3.02. This study contributed to general cybersecurity awareness for businesses and future research, the pipeline is available on GitHub for research purposes (Charlie Green, 2025).

Acknowledgments

Thank you to all businesses that contributed their time and insights to the study and to all participants who contributed to the evaluation of the pipeline.

References

AI21labs (July 2023). *Human or Not? // A Social Turing Game*. en. Available at: https://www.humanornot.ai/ (Accessed: Feb. 8, 2025).

Allodi, L., Chotza, T., Panina, E., and Zannone, N. (Mar. 2020). 'The Need for New Antiphishing Measures Against Spear-Phishing Attacks'. *IEEE Security & Privacy*, 18.(2). Conference Name: IEEE Security & Privacy, pp. 23–34. ISSN: 1558-4046. doi: 10.1109/msec.2019.2940952. Available at: https://ieeexplore.ieee.org/document/8852647/?arnumber=8852647 (Accessed: Dec. 11, 2024).

Ayeni, R. K., Adebiyi, A. A., Okesola, J. O., and Igbekele, E. (Apr. 2024). 'Phishing Attacks and Detection Techniques: A Systematic Review'. *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*, pp. 1–17. doi: 10.1109/seb4sdg60871.2024.10630203. Available at: https://ieeexplore.ieee.org/document/10630203/?arnumber=10630203 (Accessed: Feb. 27, 2025).

Barracuda (May 2023). *2023 spear-phishing trends*. Tech. rep. Barracuda. Available at: https://assets.barracuda.com/assets/docs/dms/2023-spear-phishing-trends.pdf (Accessed: Dec. 11, 2024).

Charlie Green (2025). *KazukiKoto/Spear_Phishing_Pipeline: An automatic spear phishing generation pipeline intended for use in businesses for training purposes*. en. Available at: https://github.com/KazukiKoto/Spear_Phishing_Pipeline (Accessed: Apr. 24, 2025).

Damiano, A. (Jan. 2020). 'Why do users not report spear phishing emails?' en. *Telematics and Informatics*, Available at: https://www.academia.edu/81896914/Why_do_users_not_report_spear_phishing_emails (Accessed: Feb. 8, 2025).

Grimes et al. (2023). *Data Confirms Value of Security Awareness Training and Simulated Phishing*. Tech. rep. KnowBe4. Available at: https://www.knowbe4.com/hubfs/Data-Confirms-Value-of-SAT-WP_EN-us.pdf?hsLang=en-us (Accessed: Dec. 10, 2024).

Islam, M. R., Islam, M. M., Afrin, M. S., Antara, A., Tabassum, N., and Amin, A. (Apr. 2024). *PhishGuard: A Convolutional Neural Network Based Model for Detecting Phishing URLs with Explainability Analysis*. arXiv:2404.17960 [cs]. doi: 10.48550/arxiv.2404.17960. Available at: http://arxiv.org/abs/2404.17960 (Accessed: Feb. 16, 2025).

Jannai, D., Meron, A., Lenz, B., Levine, Y., and Shoham, Y. (May 2023). *Human or Not? A Gamified Approach to the Turing Test*. arXiv:2305.20010 [cs]. doi: 10.48550/arxiv.2305.20010. Available at: http://arxiv.org/abs/2305.20010 (Accessed: Feb. 8, 2025).

Kamran, S. A., Sengupta, S., and Tavakkoli, A. (Nov. 2022). *Semi-supervised Conditional GAN for Simultaneous Generation and Detection of Phishing URLs: A Game theoretic Perspective*. arXiv:2108.01852 [cs]. doi: 10.48550/arxiv.2108.01852. Available at: http://arxiv.org/abs/2108.01852 (Accessed: Feb. 8, 2025).

OPSWAT (May 2024). *2024 Report: Email Security Threats Against Critical Infrastructure Organisations*. Tech. rep. OPSWAT. Available at: https://info.opswat.com/hubfs/FY23-FY24%20-%20Email%20Assets/2024-05-03%20Osterman%20Research%2C%202024%20Report%20on%20Email%20Security%2C%20OPSWAT%20(May%202024).pdf (Accessed: Dec. 11, 2024).

Proofpoint (2024a). *2024 State of the Phish*. Tech. rep. Proofpoint. Available at: https://www.proofpoint.com/sites/default/files/threat-reports/pfpt-us-tr-state-of-the-phish-2024.pdf (Accessed: Dec. 11, 2024).

Proofpoint (May 2024b). *GenAI Is Powering the Latest Surge in Modern Email Threats | Proofpoint UK*. en-gb. Available at: https://www.proofpoint.com/uk/blog/email-and-cloud-threats/genai-powering-latest-surge-modern-email-threats (Accessed: Feb. 8, 2025).

rolandroland (2024). *rolandroland/llama3.1-uncensored*. Available at: https://ollama.com/rolandroland/llama3.1-uncensored (Accessed: Apr. 22, 2025).

Sugunaraj, N. (May 2024). *Human Factors in the LastPass Breach*. arXiv:2405.01795 [cs]. doi: 10.48550/arxiv.2405.01795. Available at: http://arxiv.org/abs/2405.01795 (Accessed: Feb. 8, 2025).

Sun, Z. et al. (Sept. 2024). 'From Victims to Defenders: An Exploration of the Phishing Attack Reporting Ecosystem'. *Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses*. RAID '24. New York, NY, USA: Association for Computing Machinery, pp. 49–64. ISBN: 979-8-4007-0959-3. doi: 10.1145/3678890.3678926. Available at: https://dl.acm.org/doi/10.1145/3678890.3678926 (Accessed: Feb. 8, 2025).

Thakur, K., Ali, M. L., Obaidat, M. A., and Kamruzzaman, A. (Jan. 2023). 'A Systematic Review on Deep-Learning-Based Phishing Email Detection'. en. *Electronics*, 12.(21). Number: 21 Publisher: Multidisciplinary Digital Publishing Institute, p. 4545. ISSN: 2079-9292. doi: 10.3390/electronics12214545. Available at: https://www.mdpi.com/2079-9292/12/21/4545 (Accessed: Feb. 16, 2025).

Touvron, H. et al. (Feb. 2023). *LLaMA: Open and Efficient Foundation Language Models*. arXiv:2302.13971 [cs]. doi: 10.48550/arxiv.2302.13971. Available at: http://arxiv.org/abs/2302.13971 (Accessed: Apr. 23, 2025).

Wassermann, S., Meyer, M., Goutal, S., and Riquet, D. (Sept. 2023). *Targeted Attacks: Redefining Spear Phishing and Business Email Compromise*. arXiv:2309.14166 [cs]. doi: 10.48550/arxiv.2309.14166. Available at: http://arxiv.org/abs/2309.14166 (Accessed: Feb. 27, 2025).

Zhao, X. et al. (Feb. 2024). *Weak-to-Strong Jailbreaking on Large Language Models*. arXiv:2401.17256 [cs]. doi: 10.48550/arxiv.2401.17256. Available at: http://arxiv.org/abs/2401.17256 (Accessed: Feb. 8, 2025).