Social engineering and spam detection of AI-driven Phishing emails

Dr. V.Sridevi[1], Dr.SM Saravanakumar[2]
[1]*Assistant Professor, Department of Computer Science, PSG College of Arts & Science, Coimbatore*
[2]*Assistant Professor, Department of Computer Science, PSG College of Arts & Science, Coimbatore*

Abstract- Natural language processing has been transformed by the sophisticated design of advanced Language Modelswhich produces text that accurately appears like authentic communication including phishing emails. Phishing emails created by AI are becoming more common these days. This investigation aims to address this problem by examining AI driven emails and address how well Email services filter these harmful messages. The results showed that that many email services allowed more AI-driven phishing emails to circumvent their filters. The Generative AI social engineering conceptual model was incorporated to explore the complexity of Ai-driven social engineering attacks. In order to address these issues, logistic regression and XGBoost machine learning model were used to filter phishing emails based on factors the number of imperative verbs andpersonal pronouns. The Kaggle AI-generated phishing email dataset was used in this study.

Keywords: AI-driven phishing email, Textual and style analysis, Advanced Language Models (ALMs), Machine learning, cyber-attack

I. INTRODUCTION

Natural language processing has been transformed by Advanced Language Models (ALMs), which creates text that closely appears like human writing for a variety of uses. These developments have major advantages, like increasing user service (Chew, Lin, Chen, Fan, & Lee, 2024), automating content development (Kirova, Ku, Laracy, & Marlowe, 2024), and boosting technical services (Raman, Calyam, &Achuthan, 2024, V. Sridevi et al , 2024). But in addition to these advantages, ALMs pose significant hazards since they give hackers new avenues of attack. They pose a serious risk to people and organizations due to their misuse in creating incredibly convincing phishing emails (Chataut, Usman &Gyawali, 2024; Bernstein, Vishwanath, & Park, 2024; Naragam, Thota, Roy&Nilizadeh, 2023). Malicious actors may quickly create convincing phishing emails with suggestions alone, eliminating the requirement for explicit training examples. This makes it simpler to get past spam filters and take advantage of weaknesses in email-based security systems.

Phishing is a widely recognized form of social engineering attack where hackers pose as reputable organizations, including banks or government offices, in order to deceive victims into publishing financial or personal information (Drake, Oliver, & Koontz, 2004). Conventional phishing assaults have been successfully thwarted by traditional detection techniques, which depend on outside indicators such as dubious URLs, domain repute, or overt brand impersonation. Nevertheless, these cutting-edge techniques frequently rely on antiquated datasets, like SpamAssassin or Enron, which existed before ALM-driven attacks and so do not include instances of phishing content produced by AI (Alhogail&Alsabih, 2021; Maiello, Gallo, Ventre&Botta, 2021).

For detection systems, the increase of AI-driven phishing emails poses an increasing complication. Although recent research has investigated the use of ALMs to detect these threats (Heiding et al., 2024, Chataut et al., 2024;Sridevi et al., 2024), these methods frequently encounter these issues that are specific to deep learning models, making their solutions less transparent and trustworthy. Furthermore, some studies have looked into how well AI-driven phishing emails work to get rid of spam filters, but they usually only look at one email provider, providing little information(Bethany et al., 2024).

To address these problems, this paper aims the effective of spam filters and detecting AI-driven phishing emails using textual style and linguistic

features. Two well-known shallow machine learning classifiers were used to assess the efficacy of these features, and XGBoost achieved an astounding 98% accuracy. The number of imperative verbs and personal pronouns were important predictive factors. These findings show that urgent prompts and complex sentence structures are adequately used in AI-driven phishing emails to increase their persuasiveness.

II. RESEARCH CONTRIBUTIONS

Spam Filters Analysis: We test AI-generated phishing emails to empirically assess the spam screening filters given by Email providers. Our experiments highlight the advantages and disadvantages of these filters and show how simple it is for ALM-generated phishing emails to avoid conventional detection. Furthermore, we uncover notable false-positive rates with an equivalent amount of authentic AI-generated emails, highlighting the discrepancies in the effectiveness of the present spam filter.

Textual and Linguistic Features: This research emphasizes on basic textual style properties like imperative verb usage, personal pronouns and word patterns which offers an additional layer of protection.

Machine Learning Method:We obtain a 98.2% classification accuracy by using these textual and style features on two interpretable shallow classifiers, with XGBoostgives the best comparing others. This method focuses on the textual style components—like urgent verbs and phrase density—that best differentiate authentic emails from phishing emails produced by artificial intelligence. Our approach provides a transparent substitute for black-box ALM-based detectors by emphasizing interpretability, which illuminates the stylistic characteristics that drive classification.

Publicly Available Dataset: AI-generated phishing dataset available on Kaggleis used in this study which provides ALM-based social engineeringthreats.

III. THE PURPOSE OFALMS IN PHISHING DETECTION

As phishing techniques change, Advanced language models, or ALMs, have taken center stage in phishing detection. The capability of models such as GPT-3 and GPT-4 to both create and identify phishing mails. For instance, Chataut et al. (2024) showed that how can recognize phishing emails, demonstrating its capacity to comprehend and produce intricate textual structures. Heiding et al. (2024) used sophisticated text generating capabilities to refine it for phishing detection. This investigation was expanded by Patel, Rehman, and Iqbal (2024), who assessed several ALMs on a variety of phishing datasets. These findings indicate the need for domain specific data for training.

Despite the fact that ALMs have used in phishing detection, a fundamental problem is that these models are proprietary, which restricts transparency into their operation. Furthermore, the examination of how spam filters manage phishing emails on commercial web-mail services, which are frequently the target of phishing assaults (APWG, 2024) is needed.

IV. ATTACK MODEL

This paper examines aattack model in which threat actors create sophisticated phishing emails that evade spam filters and trick message receivers by taking advantage of Advanced Language Models (ALMs). By creating particular cues that direct the prototype to generate reliable phishing emails, attackers use ALMs, like GPT-4. These emails imitate authentic communication styles while incorporating psychological strategies like urgency, authority, and interest that are frequently employed in phishing assaults. Fig. 1 illustrates the process of creating phishing emails with ALMs.

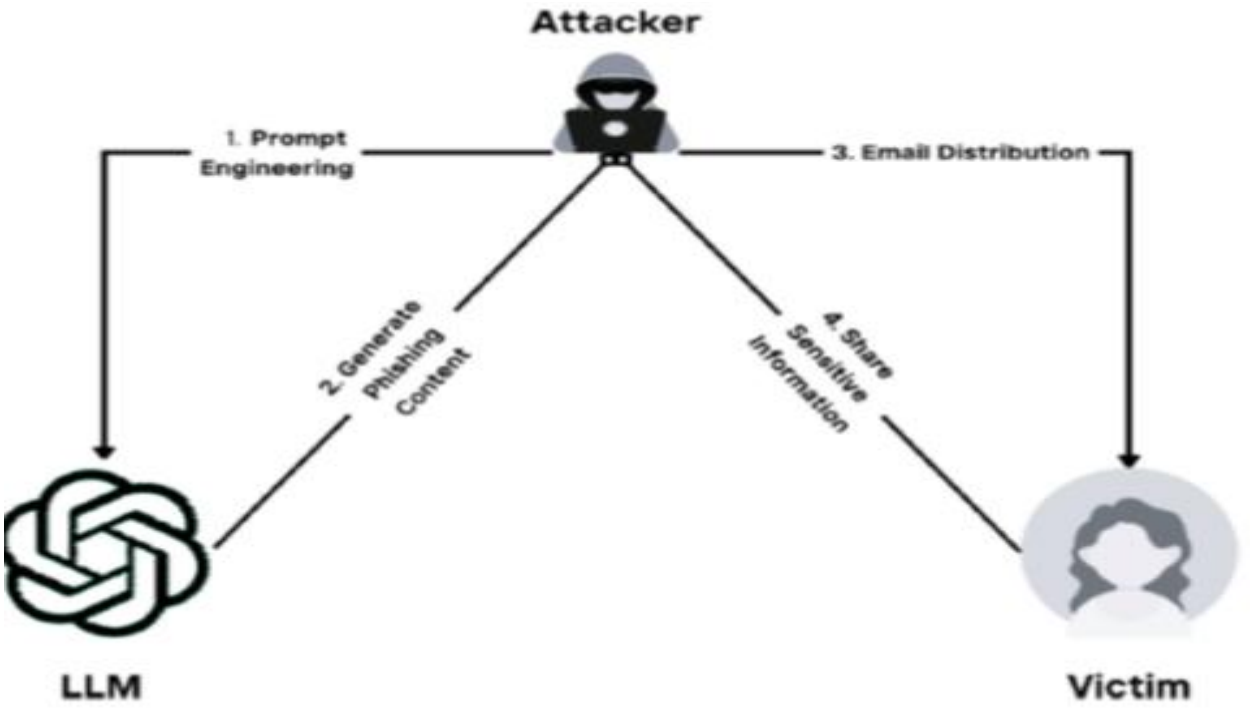

Fig 1. Process of Creating Email

The following steps are involved in the process:
1. Prompt Engineering: Attackers create prompts that are suited to particular themes (such job openings or financial notifications) and skill levels (basic, intermediate, and advanced). The ALM is guided by these suggestions to produce language that supports the targeted phishing goal.

2. Content Generation: Using the prompts as a guide, the ALM creates phishing emails that include urgent calls to action, spoof sender information, and directive language like "click" and "verify." The generated emails are more challenging for conventional detection algorithms to detect since they frequently lack obvious harmful indicators (such as dubious URLs).

3. Email Distribution: In order to reach a large audience, attackers disseminate phishing emails via popular webmail services (such as Gmail, Outlook, and Yahoo). They try to get around common spam detection heuristics by avoiding bulk sending and staggered dispatch schedules.

4. Victims' Reaction: The phishing emails' high degree of authenticity and contextual relevance trick recipients into doing things like divulging personal information or downloading malicious malware.
Attackers can get a lot of benefits by using Large Language Models (ALMs) to craft phishing scams. ALMs make it possible to create phishing content quickly and extensively, and their intuitive user interfaces make them accessible to a wide range of hackers.

V. METHODOLOGY

Fig 2 provides the overview of processes used to detect AI-driven phishing emials.

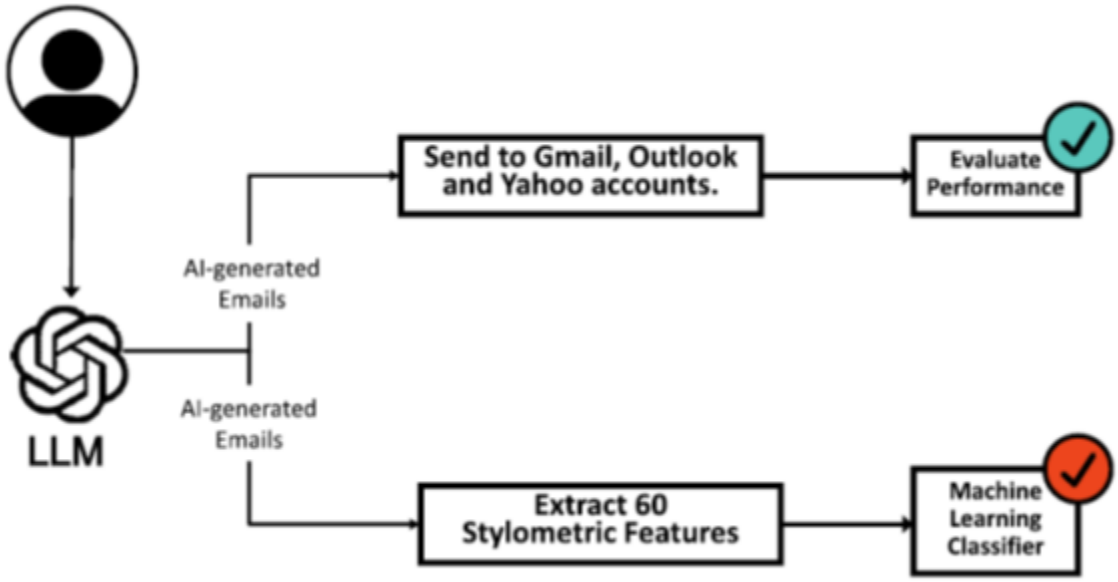

Fig 2. Overview of Process

5.1 AI-driven email generation

GPT-4o is used to create scam and fraudulent emails. The deceptive emails were written with a range of themes and varied degrees of skill (basic, intermediate, and advanced) to guarantee realism. These modifications mimic the intricacy and context frequently seen in actual communications. The created emails were then used to evaluate how reliable spam filters were across various webmail providers.

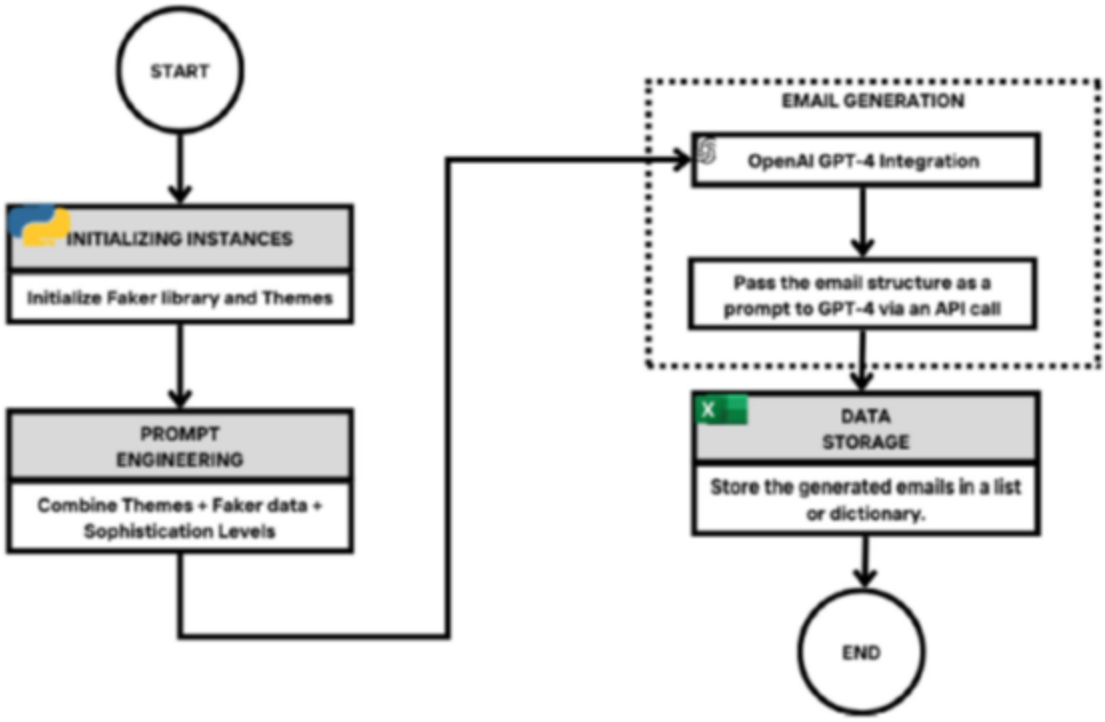

Fig. 3 Process of AI-driven email generation

Instance Initialization: It sets up the necessary elements, such as the Faker Library5, the API key, and

pre-made email templates. These components serve as the basis for creating realistic and varied content of emails.

Prompt Engineering: The library creates realistic databased on a randomly chosen email theme. The prompts in phishing emails are designed to look like authentic correspondence while including a call to prevention.

Content Generation: In this process, the script integrates the call-to-prevention phrases, the created phony data, the chosen theme, and the sophistication level. The call-to-prevention phrases in phishing emails were made to instill emergency and drive message receivers to act.In order to customize the composed emails to fit their intended purpose—whether it be genuine communication or dishonest phishing—these call-to-action words were crucial.

Export and Validation: To guarantee that the produced emails are methodically arranged and easily accessible for further examination, they are saved in a CSV file.

5.2 Textual style features extraction

The extraction of linguistic and stylometric features were done in this section. These characteristics, which include readability ratings and word category distributions, are essential for identifying phishing attempts produced by artificial intelligence.

Finding unique textual and structural features in phishing and authentic mails is the main goal of the stylometric feature extraction approach. By acting as trustworthy markers, these patterns allow strong machine learning techniques to recognize phishing threats based only on text style features. It is the key for advancing spam mail detection.

5.3 Machine learning driven solutions

Stylometric traits were used to formulate the job of identifying AI-generated phishing emails. Phishing or spam emails created by AI were regarded as negative samples, while authentic emails created by AI were seen as positive examples. In particular, machine learning models were trained and classified Email as either authentic or phishing. This process is depicted in Figure 4.

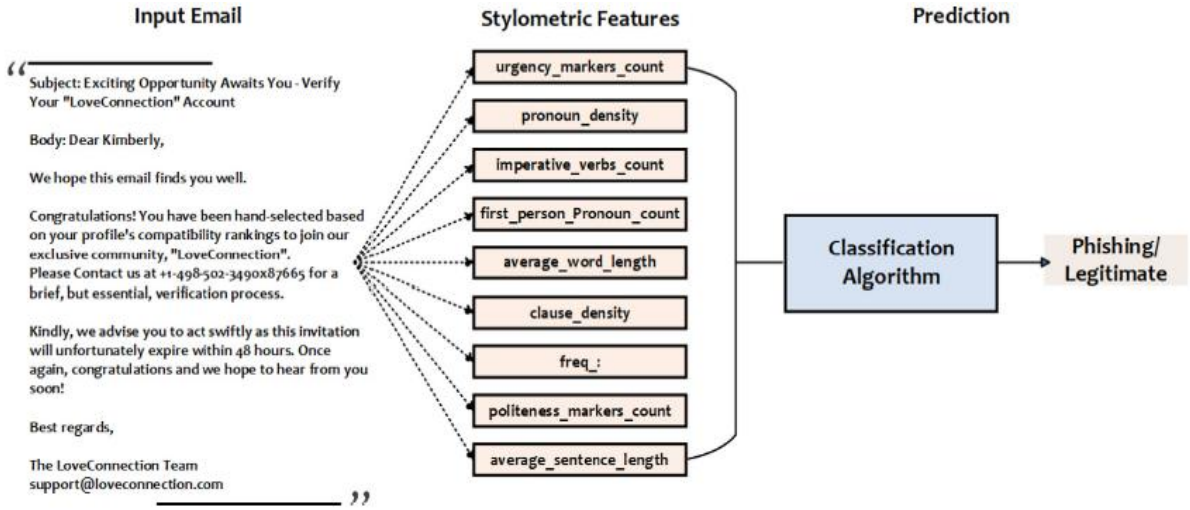

Fig. 4. Machine learning implementations

The machine learning techniques Logistic Regression and XGBoost were employed in this study. These models are used by the suggested phishing detection algorithm to categorize emails according to stylometric characteristics. As described, it adheres to a systematic procedure that comprises feature extraction, classifier training, evaluation, and data preprocessing. This methodology guarantees a methodical way to applying machine learning for categorization and identifying significant patterns in email text.

5.4 Experimental design and performance metrics

All experiments were implemented on the Google Colab platform. To guarantee accurate findings, the datasetwhich contains textual style featureswas partitioned into training and test sets. This division enables the models should use 20% of the data for

evaluation and 80% of the data for training. The StandardScaler, which adjusts the characteristics to have unit variance and zero mean, was used to normalize the features. The efficiency of the models was analyed using F1-score, recall, accuracy and precision. Area Under the Curve (AUC) was also assessed to examine the model's capacity to differentiate between two classes authentic phishing email or not. Better model performance is indicated by a higher AUC value.

5.4 Result Analysis

According to the analysis, phishing emails produced by AI closely resemble the structural patterns of phishing emails written by humans. Like their human-written counterparts, they often use urgency and emotional appeals, authority cues, and URLs. The performance metrics of two machine learning techniques are presented in Table 1.

Table 1. Efficiency of machine Learning Techniques

Techniques	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-Score (%)
Logistic regression	94	95	94	94	97
XGBoost	98	98	97	98	99

The confusion matrix values for techniques are summarized in Table 2. The following are reported: False Negatives (phishing emails misclassified as authenticate), True Negatives (authenticate emails correctly identified), False Positives (authenticate emails incorrectly classified as phishing), and True Positives (phishing emails correctly identified).

Table 2. Confusion Matrix

Techniques	True Negative	False Positive	False Negative	True Positive
Logistic Regression	22	2	1	23
XGBoost	24	1	0	23

According to Table 6's classification report, the XGBoost model classified the great majority of the test set's emails correctly, achieving an accuracy of 98.2%. This high accuracy shows how well the model can differentiate between phishing and authentic emails.

VI. CONCLUSION

One important question is how well existing email systems can identify these threats as AI-driven phishing emails become more complex. This study examined how well email servicesperformed in filtering AI-driven phishing emails that were created using AI. This study used textual style and linguistic features from machine learning techniques, Logistic Regressionand XGBoost—to address these issues. WithAUC score of 99% and 98.2% accuracy, XGBoost was the most successful. The model's successful application of stylometric elements, which are essential for detecting phishing intent, such as imperative verb usage and personal pronouns, is responsible for this impressive performance.

REFERENCE

[1] Chew, C.-J., Lin, Y.-C., Chen, Y.-C., Fan, Y.-Y., & Lee, J.-S. (2024). Preserving manipulated and synthetic deepfake detection through face texture naturalness. *Journal of Information Security and Applications*, *83*, Article 103798. http://dx.doi.org/10.1016/j.jisa.2024.103798.

[2] Kirova, V. D., Ku, C. S., Laracy, J. R., & Marlowe, T. J. (2024). Software engineering education must adapt and evolve for anllm environment. In *Proceedings of the 55th ACM technical symposium on computer science education v. 1* (pp. 666–672). http://dx.doi.org/10.1145/3626252.3630927.

[3] V. Sridevi, D. Bhuva, A. Bhuva, M. K. Sharma, S. Gupta and M. Soni, (2024). Deep Neural Networks Implementation on IoT Devices: A Practical Guide. *2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, Chennai, India,doi:10.1109/ACCAI61061.2024.10602133.

[4] Raman, R., Calyam, P., &Achuthan, K. (2024). ChatGPT or bard: Who is a better Certified Ethical Hacker? *Computers & Security*, *140*, Article 103804. http://dx.doi.org/10.1016/j.cose.2024.103804.

[5] Chataut, R., Gyawali, P. K., & Usman, Y. (2024). Can ai keep you safe? a study of large language models for phishing detection. In *2024 IEEE 14th annual computing and communication workshop and conference* (pp. 0548–0554). IEEE, http://dx.doi. org/10.1109/ccwc60891.2024.10427626.

[6] Vishwanath, A., Herath, T., Chen, R., Wang, J., & Rao, H. R. (2011). Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems*, *51*(3), 576–586. http://dx.doi.org/10.1016/j.dss.2011.03.002.

[7] Roy, S. S., Thota, P., Naragam, K. V., &Nilizadeh, S. (2023). From Chatbots to PhishBots?–Preventing Phishing scams created using ChatGPT, Google Bard and Claude. http://dx.doi.org/10.48550/arxiv.2310.19181, arXiv preprint arXiv:2310.19181.

[8] Drake, C. E., Oliver, J. J., & Koontz, E. J. (2004). Anatomy of a phishing email. In *CEAS*.

[9] Alhogail, A., &Alsabih, A. (2021). Applying machine learning and natural language processing to detect phishing email. *Computers & Security*, *110*, Article 102414. http://dx.doi.org/10.1016/j.cose.2021.102414.

[10] Gallo, L., Gentile, D., Ruggiero, S., Botta, A., &Ventre, G. (2024). The human factor in phishing: Collecting and analyzing user behavior when reading emails. *Computers & Security*, *139*, Article 103671. http://dx.doi.org/10.1016/j.cose.2023.103671.

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

[12] Sridevi, V., Samath, J.A. (2024). A combined deep CNN-lasso regression feature fusion and classification of MLO and CC view mammogram image. *Int J Syst Assur EngManag* **15**, 553–56.https://doi.org/10.1007/s13198-023-01871-x

[13] Bethany, M., Galiopoulos, A., Bethany, E., Karkevandi, M. B., Vishwamitra, N., &Najafirad, P. (2024). Large language model lateral spear phishing: A comparative study in large-scale organizational settings. http://dx.doi.org/10.48550/arxiv.2401. 09727, arXiv preprint arXiv:2401.09727.

[14] Patel, H., Rehman, U., & Iqbal, F. (2024). Evaluating the efficacy of large language models in identifying phishing attempts. In *2024 16th international conference on human system interaction* (pp. 1–7). IEEE, http://dx.doi.org/10.1109/hsi61632.2024.10613528.

[15] Anandhi, Sridevi et al. (2024). Innovative Pedagogies for 6G Security Educating the Next Generation. DOI: 10.4018/979-8-3693-7421-4.ch001

[16] APWG (2024). *Phishing activity trends report, 1st quarter 2024*: *Technical report*, Anti-Phishing Working Group.

[17] Yamin MM, Ullah M, Ullah H, Katt B (2021) Weaponized AI for cyber attacks. J Inform Secur Appl 57:102722. https://doi.org/10.1016/j.jisa.2020.102722

[18] Zscaler (2024) ZscalerThreatLabz 2024 Phishing Report

[19] Mouton F, Malan MM, Leenen L, Venter HS (2014) Social engineering attack framework. In: 2014 information security for South Africa - proceedings of the ISSA 2014 conference. Institute of Electrical and Electronics Engineers Inc

[20] Kaur R, Gabrijelčič D, Klobučar T (2023) Artificial intelligence for cybersecurity: literature review and future research directions. Inform Fusion. https://doi.org/10.1016/j.inffus.2023.101804

[21] Desolda G, Ferro LS, Marrella A et al (2022) Human factors in phishing attacks: a systematic literature review. ACM ComputSurv. https://doi.org/10.1145/3469886