Debate-Driven Multi-Agent LLMs for Phishing Email Detection

Ngoc Tuong Vy Nguyen
Department of Computer Science
Earlham College
Richmond, IN, USA
nnguyen24@earlham.edu

Felix D Childress
Department of Computer Science
Earlham College
Richmond, IN, USA
fdchild22@earlham.edu

Yunting Yin
Department of Computer Science
Earlham College
Richmond, IN, USA
yinyu@earlham.edu

***Abstract*—Phishing attacks remain a critical cybersecurity threat. Attackers constantly refine their methods, making phishing emails harder to detect. Traditional detection methods, including rule-based systems and supervised machine learning models, either rely on predefined patterns like blacklists, which can be bypassed with slight modifications, or require large datasets for training and still can generate false positives and false negatives. In this work, we propose a multi-agent large language model (LLM) prompting technique that simulates debates among agents to detect whether the content presented on an email is phishing. Our approach uses two LLM agents to present arguments for or against the classification task, with a judge agent adjudicating the final verdict based on the quality of reasoning provided. This debate mechanism enables the models to critically analyze contextual cue and deceptive patterns in text, which leads to improved classification accuracy. The proposed framework is evaluated on multiple phishing email datasets and demonstrate that mixed-agent configurations consistently outperform homogeneous configurations. Results also show that the debate structure itself is sufficient to yield accurate decisions without extra prompting strategies.**

***Index Terms*—phishing detection, large language models, multi-agent debate**

I. Introduction

Phishing attacks are one of the most prevalent and damaging cybersecurity threats. Attackers use feelings of fear and urgency to manipulate victims and deceive them into revealing sensitive credentials and financial information. According to industry reports, phishing remains a dominant vector for cybercrime, with a growing variety in attack strategies. Despite advancements in detection mechanisms, adversaries continuously evolve their tactics, using social engineering and obfuscation techniques to bypass automated filters. Traditional phishing detection methods primarily rely on rule-based filtering. While they are effective for detecting known phishing patterns, they often fail against novel ones. Moreover, machine learning models require extensive labeled datasets and can struggle with generalization when exposed to new forms of phishing. More recently, large language models (LLMs) have demonstrated exceptional text understanding capabilities, making them viable candidates for phishing detection. Unlike traditional classifiers, LLMs can analyze linguistic details, and even context and intent in emails given their extensive training on large datasets, allowing them to detect phishing emails intended for psychological manipulation.

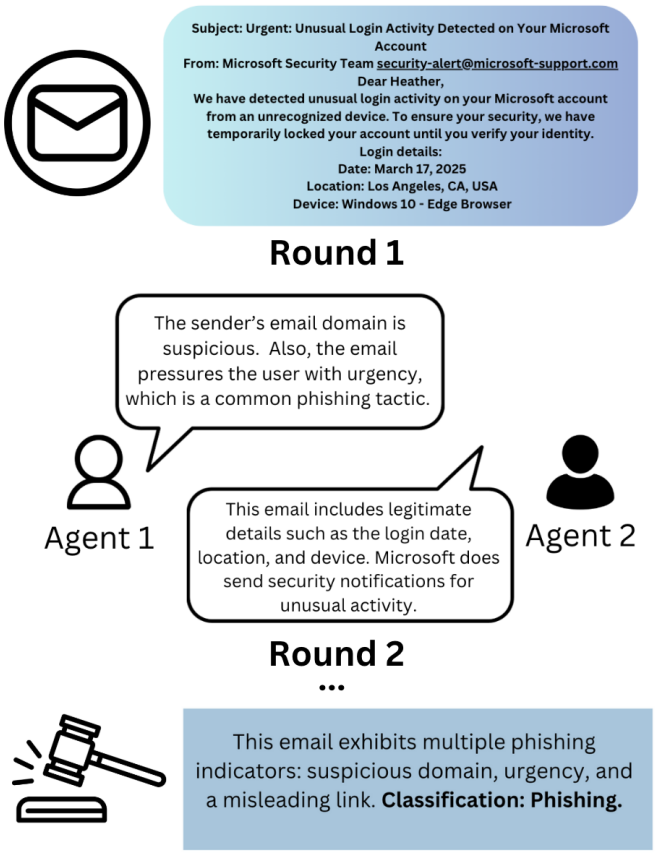

Fig. 1: An illustration of the proposed multi-agent debate framework, where two LLM agents engage in a structured debate over a potential phishing email, and a judge agent evaluates their arguments to produce a final classification.

However, a major limitation of single-agent LLM approaches is confirmation bias, where the model tends to overfit its initial reasoning path and fails to explore alternative interpretations. Additionally, phishing detection often requires contextual analysis from different perspectives. For example, an email might appear legitimate at first glance, but deeper

analysis of inconsistencies or urgency cues may indicate possibility of phishing. To address these challenges, we propose a debate-driven multi-agent LLM framework for phishing email detection. Our approach employs multiple LLMs that engage in structured debates to determine whether a given message is phishing or ham (legitimate). Each agent assumes a specific stance: one arguing that the message is fraudulent and another arguing for its legitimacy. After two rounds of argument exchange, a judge agent evaluates the debate and issues a final classification decision based on the strength of the presented arguments.

Our results show that debater agent pairs with heterogeneous models outperform homogeneous ones in phishing email classification accuracy. Furthermore, we evaluate the impact of prompt engineering techniques including chain-of-thought and role prompting, and find that they do not significantly improve performance on top of our debate framework. These findings prove the value of multi-agent reasoning in phishing detection and suggest broader potential for debate-based frameworks in other NLP and cybersecurity applications.

This paper is organized as follows: In Section II, we review prior literature on phishing detection techniques and the use of multi-agent LLM systems. Section III describes the datasets used in our experiments and the preprocessing steps taken to ensure compatibility with LLM input constraints. Section IV describes our proposed multi-agent debate framework, including the design of prompt templates and the debate procedure. In Section V, we present experimental results comparing different agent configurations and prompting strategies across multiple datasets. Finally, we conclude our findings in SectionVI.

II. Related Works

In this section, we present a comprehensive literature review on phishing attacks, their impact, and the various approaches used for detection and mitigation. We survey how machine learning, deep learning, and LLM-based techniques have been used to combat phishing threats and analyze their effectiveness and limitations. Additionally, we explore the use of multi-agent debate systems in various computing tasks before proposing such a framework for phishing detection.

A. Phishing Email Attack Detection

Phishing email attacks are a form of cyberattack where attackers disguise themselves as trustworthy individuals or organizations to deceive recipients of their emails into taking harmful actions. These emails often attempt to steal sensitive information such as passwords, credit card details, or personal data [1]. They also attempt to spread malware and manipulate users into making financial transactions. Phishing emails typically include links that redirect users to malicious websites or attachments embedded with malware. Attackers often employ social engineering tactics, such as creating a sense of urgency or fear, or impersonating authority figures to trick victims into responding. Phishing emails can be difficult for humans to detect, especially when they look similar to safe emails [2]. Phishing email detection is a well-studied area in cybersecurity, and various techniques have been developed over time, ranging from traditional methods to machine learning and deep learning approaches.

Traditional detection methods include blacklist [3] and whitelist [4], [5] techniques that block emails from known malicious sources or only allow emails from trusted senders. However, phishing email attacks are evolving to bypass these traditional detection methods by closely mimicking legitimate communications and using sophisticated evasion techniques, such as dynamic URLs and AI-generated content [6]. As a result, cybersecurity defenses have transitioned from simple rule-based filters to advanced machine learning and deep learning models that derive patterns from email structure, language, and metadata to improve detection accuracy. Salahdine et al. [7] propose a machine learning-based phishing detection technique by extracting 10 key features from 4,000 phishing emails. Experimental results on classification task demonstrate that an artificial neural network achieves superior accuracy compared to other methods. Valecha et al. [8] investigate the effectiveness of persuasion cues, specifically gain and loss cues, in phishing email detection by developing three machine learning models that incorporate these cues. The results show that persuasion cues improve model performance, and that psychological tactics are useful in anti-phishing methods. Hamid and Abawajy [9] propose a technique to improve the accuracy of phishing email detection by using a hybrid feature selection method combining content-based and behavior-based features extracted from email headers. The proposed method has improved detection rates due to successful identification of attacker behaviors. Altwaijry et al. [10] explore one-dimensional CNN-based models with integrated recurrent layers for phishing email detection, and show that the 1D-CNNPD model with Bi-GRU achieves the highest accuracy and demonstrates the potential of deep learning techniques to reduce false positive rate.

With the increasing popularity of LLMs, they are being increasingly used for developing phishing detection tools. Koide et al. [11] propose a phishing email detection system called ChatSpamDetector that uses GPT-4 to analyze email content and provide both classification and explanatory reasoning. Experimental results demonstrate that ChatSpamDetector outperforms traditional spam filters and baseline machine learning models by effectively identifying phishing tactics. Heiding et al. [12] compare phishing emails generated by GPT-4, V-Triad, and a combination of both, and conclude that LLM-generated content generally outperform generic phishing emails. Thy also argue that LLMs can effectively detect phishing intent and sometimes outperform human detection performance. Lee [13] uses hybrid feature selection and prompt engineering to investigate the effectiveness of LLMs in detecting phishing emails of various types, including spear phishing, traditional phishing, and AI-generated phishing. Experimental results show that Llama-3.1-70B achieves superior accuracy over other models while also provides interpretable reasoning.

B. Multi-agent Debate to Enhance LLM Reasoning

Recent research has showed that multi-agent debate frameworks are useful techniques to enhance the reasoning abilities of LLMs. Unlike single-agent prompting methods, multi agent setups create multiple LLM instances to critique and refine each other's responses, which often lead to more accurate and well-reasoned outputs. Du et al. [14] introduce a multi-agent debate approach where multiple LLM instances debate their responses over multiple rounds to improve reasoning, and their findings show that this method improves mathematical and strategic reasoning while reducing hallucinations. Liang et al. [15] identify the Degeneration-of-Thought problem in LLMs, where self-reflection fails to generate novel insights once the model becomes overconfident in its initial solution. They address the identified problem with a multi-agent debate framework with judge supervision, and concluded that the reasoning performance is improved on complex tasks like commonsense translation and arithmetic. Further investigating the use of multi-agent interactions, Estornell et al. [16] propose ACC-Collab, a learning framework that trains a two-agent team that contains an actor-agent and a critic-agent to facilitate collaboration between the agens and improve their problem-solving skills. The proposed framework outperforms existing multi-agent methods in various benchmarks. Wang et al. [17] make two observations on multi-agent discussions: they outperform single-agent setups when no demonstrations are provided, and in multi-LLM environments, stronger LLMs help weaker ones reason through interaction. Collectively, these studies demonstrate the effectiveness of multi-agent debate frameworks in improving decision-making and reasoning abilities of LLMs. They also demonstrate the potential of such frameworks in cybersecurity applications, such as phishing email detection studied in this work, where reasoning over text is critical.

III. Datasets

To evaluate the effectiveness of our proposed multi-agent debate framework, we use a diverse set of email datasets that capture different communication styles and phishing tactics. These datasets include both real-world and synthetically generated emails, spanning multiple time periods. Incorporation of multiple sources makes our experimental setup representative of the diverse scenarios typically encountered in phishing email detection. The datasets used in our study are as follows:

- **University of Twente Phishing Validation Emails Dataset [18]:** A dataset of 2,000 real and artificially generated example emails.
- **Eleven Curated Datasets of Phishing Email [19], [20]:** A collection of cleaned email corpus with phishing or ham labels, including CEAS-08, Ling-Spam, Enron, Nazario, Nazario_5, Nigerian_5, Nigerian_Fraud, SpamAssassin, TREC-05, TREC-06, and TREC-07.

To keep the number of input tokens within the practical limits of LLMs, we selected the University of Twente (UoT) dataset along with a representative subset of four curated datasets from the publicly available corpus collection. Specifically, Ling, Nazario_5, Nigerian_Fraud, and SpamAssasin are used for our experiments. The summary statistics of these datasets are presented in Table I. During exploratory data analysis, we observed that certain datasets contained emails with exceptionally long content, which could exceed token limits imposed by LLM architectures. To address this, we kept only the emails whose token lengths fall within the 75th percentile of their respective dataset distributions. After applying this filtering step, the final dataset used in our experiments comprises a total of 12,798 emails, of which 6,475 are ham (safe) and 6,323 are phishing.

Dataset	Size	Email Length		Num of Ham	Num of Phishing
		Avg	75%		
UoT	2000	86.71	95.00	1000	1000
Ling	2859	3222.32	4014.50	2401	458
Nazario_5	3065	3545.33	1630.00	1500	1565
Nigerian_Fraud	3332	2644.38	3211.75	0	3332
SpamAssasin	5808	2406.91	2028.00	4091	1718

TABLE I: Summary statistics of selected phishing and ham email datasets, including size, email length distribution, and label distribution.

IV. Methods

A. Multi-Agent Debate Framework

We propose a multi-agent debate framework for phishing email detection, composed of three components: two debater agents, a pre-defined and scripted debate procedure, and a judge agent. The debater agents consist of two LLM-based instances, which may be instantiated from the same or different models. The first agent is prompted to argue that the given email is a phishing attempt, while the second agent is prompted to respond to the first agent's output by countering those claims and arguing for the email's legitimacy. The two agents then engage in another round to make sure that the arguments are well-formulated while maintaining computational efficiency.

The debate procedure is pre-defined and scripted to generate template prompts for each email in the dataset:

1) Round One:
   - Carefully analyze the following email and argue why it is likely to be a phishing attempt **(Agent 1)**
   - Carefully analyze the following email and argue why it is likely to be legitimate and not a phishing attempt **(Agent 2)**
2) Round Two:
   - Given your opponent's rebuttal, reinforce your position that the following email is a phishing attempt **(Agent 1)**
   - Given your opponent's rebuttal, reinforce your position that the following email is not a phishing attempt **(Agent 2)**

Arguments made by the two agents are logged for subsequent judge evaluation. Following the two-round debate, a

TABLE II: Phishing email detection accuracy and F1 scores on five benchmark datasets, comparing different agent configurations and the effects of prompt engineering techniques.

Agent Configuration			UoT		Ling		Nazario_5		Nigerian_Fraud		SpamAssassin	
Agent 1	**Agent 2**	**Judge**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
GPT-4	GPT-4	GPT-4	98.12%	0.98	98.76%	0.98	98.03%	0.98	98.54%	/	98.40%	0.98
LLaMA-2	LLaMA-2	LLaMA-2	98.01%	0.98	98.32%	0.98	98.22%	0.98	98.17%	/	98.09%	0.98
GPT-4	LLaMA-2	GPT-4	**98.91%**	**0.98**	**99.43%**	**0.99**	**99.02%**	**0.99**	**99.27%**	/	98.73%	0.98
LLaMA-2	GPT-4	GPT-4	98.36%	0.98	99.02%	0.99	98.71%	0.98	98.85%	/	**99.12%**	**0.99**
GPT-4–LLaMA-2–GPT-4-CoT			98.65%	0.98	99.12%	0.99	98.77%	0.98	99.00%	/	98.63%	0.98
GPT-4–LLaMA-2–GPT-4-Role			98.65%	0.98	98.95%	0.98	98.52%	0.98	98.74%	/	98.90%	0.99
GPT-4–LLaMA-2–GPT-4-CoT-Role			98.38%	0.98	98.77%	0.98	98.69%	0.98	98.59%	/	98.45%	0.98

third LLM instance serving as the judge agent is fed the four arguments and is prompted to evaluate the strength and coherence of the arguments. The judge is then prompted to provide a final binary classification verdict, phishing or legitimate, which is logged alongside its reasoning for performance assessment.

B. Synergy with Prompt Engineering Techniques

We also incorporate two prompting engineering techniques in our multi-agent debate framework to evaluate their effectiveness on phishing email detection.

1) Chain-of-Thought Prompting: Chain-of-Thought (CoT) prompting is a technique that encourages language models to generate intermediate reasoning steps before arriving at a final answer. In order to guide the agents to articulate the rationale behind their judgment, the following prompts are appended to the end of the basic template shown in Section IV-A:

- **Agent 1:** Break down your reasoning **step-by-step** using these guiding questions: 1. Is the language designed to invoke urgency, fear, or greed? 2. Are there misleading links or unusual requests? 3. Does the email resemble common phishing patterns?
- **Agent 2:** Break down your reasoning **step-by-step** using these guiding questions: 1. Is the tone and language professional and consistent? 2. Are the links safe and are the requests expected? 4. Does the context match what a legitimate sender would send?

2) Role Prompting: Role prompting is a prompting technique to instruct a language model to take on a specific role or persona to influence its tone and reasoning. Instead of simply asking the model to perform a task, it is first told who it is in the context of the task. This technique leads to more coherent and contextually appropriate responses, and also prevents agents from echoing each other by further anchoring them in distinct roles. To assign roles for the agents, the following prompts are appended to the front of the basic template:

- **Agent 1:** You are a senior cybersecurity analyst at a large tech company. Your job is to review suspicious emails reported by employees and determine that they are phishing attempts.
- **Agent 2:** You are an email forensics expert working for an IT compliance team. Your job is to validate that a flagged email is legitimate and not a phishing attempt.

V. Experiments

To evaluate the effectiveness of our proposed multi-agent debate framework, we conduct experiments using different combinations of two LLMs, GPT-4 and LLaMA-2, as debater agents and judges. The different agent-agent-judge configurations allow us to analyze the impact of different model pairings on debate performance and classification accuracy.

For each email in the filtered dataset described in Section III, the debate procedure was executed using the prompt templates detailed in Section IV-A. The arguments from both agents were collected, and the judge agent was prompted to evaluate the debate and produce a binary classification label, along with a brief justification. The predicted labels were then compared to the ground-truth labels to compute classification accuracy across the five datasets.

Table II presents phishing email detection accuracy for each agent-agent-judge setup, evaluated on the five selected datasets: UoT, Ling, Nazario_5, Nigerian_Fraud, and SpamAssassin. We observe that both the fully GPT-4 and fully LLaMA-2 configurations consistently underperform relative to mixed-model setups. Notably, the mixed configuration GPT-4–LLaMA-2–GPT-4 achieves the highest accuracy on four out of the five datasets, indicating that heterogeneous agents can complement each other's reasoning capabilities. This finding supports the observations made by Wang et al. [17], which state that multi-agent systems often outperform single-agent setups, and that collaboration between agents can improve their task performance.

To further investigate the factors contributing to model performance within the debate framework, we conducted experiments evaluating the impact of chain-of-thought prompting, role prompting, and the combination of both, with the best-performing agent configuration (GPT-4–LLaMA-2–GPT-4) as the baseline. As shown in Table II, neither CoT prompting, role prompting, nor their combination outperformed the baseline configuration without these enhancements. While CoT prompting encourages step-by-step reasoning and role prompting assigns each agent a distinct persona, the differences in reasoning did not translate into measurable gains in classification accuracy or F1 score across the datasets.

One possible explanation is that the debate framework itself already results in sufficiently structured reasoning, particularly with the multi-round interaction procedure, as the exchange of arguments and counterarguments naturally prompts each

agent to support its position with evidence. Furthermore, the predefined prompt template already assigns opposing positions to the two agents, as one must advocate that the email is phishing while the other defending its legitimacy. As a result, explicit role prompting may offer limited added value.

These results show the effectiveness of multi-agent debate in phishing email detection and highlight the importance of selecting a mixture of capable agents for both argumentation and judgment. In particular, our findings suggest that using mixed models is critical for accurate final decisions, and that they have a greater impact than added prompting complexity.

VI. Conclusion

In this work, we present a multi-agent large language model debate framework for phishing email detection. Unlike traditional rule-based or single-model classifiers, our approach simulates a structured argument between two LLM agents followed by a third judge agent that issues the final classification verdict. Our experiments on five benchmark phishing datasets demonstrate that mixed-model agent configurations consistently outperform homogeneous setups. These results support the hypothesis that heterogeneous agents can complement each other's reasoning abilities, leading to more accurate classification outcomes. Additionally, we explored the impact of two prompting strategies, chain-of-thought and role prompting, but found that they did not significantly improve performance over the best baseline. This suggests that the structured debate mechanism itself already elicits rich reasoning without additional prompting engineering efforts.

References

[1] Z. Alkhalil, C, Hewage, L. Nawaf, and I, Khan, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy," Front. Comput. Sci. 3:563060, 2021. Available: doi: 10.3389/fcomp.2021.563060

[2] K. Singh, P. Aggarwal, P. Rajivan, and C. Gonzalez, "What makes phishing emails hard for humans to detect?" Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 64(1), 431-435, 2021. Available: https://doi.org/10.1177/1071181320641097

[3] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, "An Empirical Analysis of Phishing Blacklists," International Conference on Email and Anti-Spam, 2009.

[4] B. Amine, E. Aimeur, and M. A. Chikh, "A personalized whitelist approach for phishing webpage detection," Proceedings - 2012 7th International Conference on Availability, Reliability and Security, ARES 2012. 249-254.

[5] N. A. Azeez, S. Misra, I. A. Margaret, L. Fernandez-Sanz, S. M. Abdulhamid, "Adopting automated whitelist approach for detecting phishing attacks," Computers & Security, Volume 108, 2021.

[6] A. G. Desetty, V. Dutt, and S. R. Pulyala, "Phishing attacks: evolving techniques, emerging trends, and countermeasure strategies, " International Journal for Innovative Engineering and Management Research, 09. 985-991, 2020.

[7] F. Salahdine, Z. El Mrabet and N. Kaabouch, "Phishing attacks detection a machine learning-based approach," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 2021, pp. 0250-0255.

[8] R. Valecha, P. Mandaokar and H. R. Rao, "Phishing email detection using persuasion cues," in IEEE Transactions on Dependable and Secure Computing, vol. 19, no. 2, pp. 747-756, 1 March-April 2022. Available: doi: 10.1109/TDSC.2021.3118931

[9] I. R. A. Hamid and J. Abawajy, "Hybrid feature selection for phishing email detection," Algorithms and Architectures for Parallel Processing, ICA3PP, 2011. Available: https://doi.org/10.1007/978-3-642-24669-2_26

[10] N. Altwaijry, I. Al-Turaiki, R. Alotaibi, F. Alakeel, "Advancing phishing email detection: a comparative study of deep learning models," Sensors (Basel). 2024 Mar 24;24(7):2077.

[11] T. Koide, N. Fukushi, H. Nakano and D. Chiba, "ChatSpamDetector: leveraging large language models for effective phishing email detection," 2024. Available: 10.48550/ARXIV.2402.18093.

[12] F. Heiding, B. Schneier, A. Vishwanath, J. Bernstein and P. S. Park, "Devising and Detecting Phishing Emails Using Large Language Models," in IEEE Access, vol. 12, pp. 42131-42146, 2024, doi: 10.1109/ACCESS.2024.3375882.

[13] C. Lee, "Enhancing Phishing Email Identification with Large Language Models," 2025. Available: 10.48550/arXiv.2502.04759.

[14] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," In Proceedings of the 41st International Conference on Machine Learning (ICML'24), Vol. 235. JMLR.org, Article 467, 11733–11763, 2024.

[15] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S. Shi, and Z. Tu, "Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate," In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics, 2024.

[16] A. Estornell, J. Ton, Y. Yao, and Y. Liu, "ACC-Debate: An Actor-Critic Approach to Multi-Agent Debate," 2024. Available: 10.48550/arXiv.2411.00053.

[17] Q. Wang, Z. Wang, Y. Su, H. Tong, and Y. Song, "Rethinking the Bounds of LLM Reasoning: Are Multi-Agent Discussions the Key?" In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6106–6131, Bangkok, Thailand. Association for Computational Linguistics, 2024.

[18] R. Miltchev, D. Rangelov, G. Evgeni, Phishing validation emails dataset. Zenodo, 2024. Available: 10.5281/zenodo.13474745

[19] A. I. Champa, M. F. Rabbi, and M. F. Zibran, "Why phishing emails escape detection: A closer look at the failure points," in 12th International Symposium on Digital Forensics and Security (ISDFS), 2024, pp. 1–6.

[20] A. I. Champa, M. F. Rabbi, and M. F. Zibran, "Curated datasets and feature analysis for phishing email detection with machine learning," in 3rd IEEE International Conference on Computing and Machine Intelligence (ICMI), 2024, pp. 1–7.