E-ISSN : 3025-1311
https://ejournal.warunayama.org/kohesi

Kohesi: Jurnal Multidisiplin Saintek
Volume 8 No 5 Tahun 2025

PHISHING EMAIL DETECTION USING LARGE LANGUAGE MODELS (LLMS): A PERFORMANCE EVALUATION OF QWEN AND GEMINI

Andyana Muhandhatul Nabila[1], Moh Sulthan Arief Rahmatullah[2]
Department of Information Technology, Faculty of Intelligent Electrical and Informatics Technology, Sepuluh Nopember Institute of Technology
aa.andyana823@gmail.com[1], ramasedang@gmail.com[2]

Abstract

The increasing complexity of network infrastructure and the increasing sophistication of phishing attacks require advanced cybersecurity solutions. Artificial Intelligence for IT Operations (AIOps) integrates big data analytics, machine learning and automation to improve real-time detection and response to security threats. This study evaluates the zero-shot performance of Large Language Models (LLMs) - Gemini 2.5 Pro, Gemini 2.5 Flash, and Qwen 3 - in detecting phishing emails in an AIOps environment at Institut Teknologi Sepuluh Nopember (ITS). The findings show different strengths: Gemini 2.5 Pro achieved 99.8% accuracy in identifying legitimate emails, minimizing false positives and workflow disruption, while Gemini 2.5 Flash excelled in detecting phishing attempts with 89.1% accuracy, prioritizing threat prevention. Qwen 3 performed poorly, most likely due to its lack of alignment with the nuances of English-language phishing. Achieved without refinement, these results highlight LLM's out-of-the-box efficacy for cybersecurity, offering an accessible and high-performance tool for organizations with limited AI resources. This study underscores the potential of LLM in AIOps to improve automated security monitoring and incident response, advocating for a layered approach that combines smart technology, user training, and organizational policies to effectively combat evolving phishing threats.

Keywords: AIOps, Phishing Detection, Large Language Models, Cybersecurity, Zero-Shot Evaluation

Article History
Received: Juni 2025
Reviewed: Juni 2025
Published: Juni 2025

Plagirism Checker No 234
Prefix DOI : Prefix DOI : 10.8734/Kohesi.v1i2.365
Copyright : Author
Publish by : Kohesi

INTRODUCTION

As network infrastructure and cybersecurity systems become increasingly complex, the application of artificial intelligence (AI) in IT operations automation, known as Artificial Intelligence for IT Operations (AIOps), has become essential. AIOps integrates big data analytics, machine learning (ML), and automation to enhance efficiency in detecting, analyzing, and responding to security incidents in real-time (Chataut et al., 2024). One of the increasingly sophisticated and concerning cybersecurity threats is phishing attacks, which exploit human vulnerabilities and can result in significant financial and reputational losses (Carroll et al., 2022; Desai & R, 2024). Modern phishing techniques such as spear phishing and whaling are increasingly difficult to detect, necessitating more advanced prevention methods (Putra et al., 2024). While traditional defenses like email filters and two-factor authentication provide some protection, these methods have limitations (Desai & R, 2024). Large Language Models (LLMs) such as GPT-4 show great potential in text analysis for detecting phishing emails with superior contextual understanding and the ability to provide more transparent classification explanations (Andriu, 2023; Chataut et al., 2024). However, research on the application of LLMs in AIOps for network and cybersecurity, particularly at the Institute of Technology Sepuluh Nopember (ITS), remains limited. This study aims to evaluate the

performance of LLMs such as Qwen3-235B-A22B and Gemini (2.5 Flash and 2.5 Pro) in detecting phishing emails without additional training (zero-shot evaluation) in an AIOps environment, provide practical insights to enhance automated security monitoring and incident response at ITS, and contribute to the development of smarter and more adaptive cybersecurity systems (Putra et al., 2024).

LITERATURE REVIEW

1. Artificial Intelligence for IT Operations

Artificial Intelligence for IT Operations (AIOps) is emerging as a crucial technology for enhancing IT operations, including cybersecurity (Laxmi Rijal et al., 2022). With its ability to automate data analysis and improve decision-making efficiency, AI has provided significant benefits in public administration (Talitha Salsabila et al., 2024). However, on the other hand, the advancement of AI also brings new challenges in cybersecurity. AI-powered cyber attacks, such as AI botnets, are increasingly threatening data security in Indonesia. These threats demand stricter protection measures, such as enhanced access control, data encryption, and more advanced detection and response systems (Anastasya Zalsabilla Hermawan et al., 2023). Although AI can be used to launch cyber attacks, this technology also plays a crucial role in strengthening cybersecurity by enabling early threat detection, in-depth attack analysis, and rapid response to emerging threats (Arthur Gregorius Pongoh et al., 2024). Various AI techniques, such as artificial neural networks and machine learning algorithms, have proven effective in detecting suspicious behavior patterns and identifying cyber threats before they harm information systems (Arthur Gregorius Pongoh et al., 2024). As cyber threats become increasingly complex, the integration of AI in cybersecurity is becoming more essential to build systems that are more adaptive and proactive in countering evolving attacks. This underscores that AI implementation in cybersecurity is not only a defensive tool but also a strategic solution in creating a safer and more reliable digital ecosystem.

2. Comparison of ML and LLM for Phishing

Recent research has extensively examined the effectiveness of Large Language Models (LLMs) in comparison to traditional Machine Learning (ML) approaches for phishing detection, shedding light on their respective advantages and limitations. Studies suggest that LLMs leveraging prompt engineering can achieve notable accuracy, with an F1-score of 92.74%. However, fine-tuned task-specific LLMs, which are trained on domain-specific phishing datasets, consistently outperform their general-purpose counterparts, reaching an F1-score of 97.29% (Fouad & Chehab, 2024). This improvement highlights the importance of domain adaptation, where models trained with phishing-specific knowledge exhibit superior performance in identifying fraudulent activities.

Beyond textual analysis, multimodal LLMs have demonstrated remarkable capabilities in phishing detection by incorporating multiple data sources, such as visual and textual cues. These models can effectively identify brand impersonation techniques used in phishing websites, surpassing the accuracy of state-of-the-art detection systems (Lee et al., 2024). The ability to analyze both textual content and visual elements, such as logos and design inconsistencies, allows multimodal models to detect sophisticated phishing attempts that may bypass conventional text-based filters. For phishing email detection, traditional ML approaches remain relevant, particularly lightweight models like logistic regression, which offer interpretable results alongside strong performance (Greco et al., 2024). These models are often favored in enterprise settings where transparency and explainability are crucial for security analysts. While deep learning models provide higher accuracy, the complexity of their decision-making process can make it difficult for security teams to interpret predictions, leading to challenges in trust and adoption.

Furthermore, comparative studies have revealed nuanced differences in the performance of various LLM architectures. For instance, the transformer-based DeBERTa V3

model has been found to slightly outperform GPT-4 in phishing detection, achieving a recall rate of 95.17% compared to GPT-4's 91.04% (Mahendru & Pandit, 2024). This finding underscores the significance of architecture-specific optimizations, where smaller, specialized models can sometimes rival or even exceed the performance of larger general-purpose models in specific security applications. Overall, these findings emphasize that both LLMs and traditional ML approaches offer distinct strengths in phishing detection. While LLMs, particularly those fine-tuned for cybersecurity tasks, provide superior accuracy and adaptability, traditional ML models remain valuable for their efficiency and interpretability. The future of phishing detection may lie in hybrid approaches that integrate the strengths of both methodologies, leveraging the scalability of LLMs while maintaining the transparency of traditional ML techniques.

3. Zero-Shot Learning for Cyber Security

Zero-shot learning (ZSL) is an emerging paradigm in machine learning designed to address the challenge of classifying instances from previously unseen classes. Unlike traditional supervised learning, which requires labeled training data for each class, ZSL enables models to generalize knowledge from known categories to novel ones based on shared semantic attributes. This capability is particularly valuable in cybersecurity, where new threats frequently emerge, often without sufficient labeled data for training. One notable application of ZSL in cybersecurity is network intrusion detection. Researchers have explored a Grassmannian approach, leveraging geometric representations to detect novel attack patterns without requiring labeled examples (Rivero Pérez et al., 2017). This method enhances the adaptability of intrusion detection systems by allowing them to identify previously unknown threats based on learned feature relationships.

For web-based anomaly detection, a ZSL method utilizing convolutional neural networks (ZSL-CNN) has demonstrated promising effectiveness. This approach achieved a 99.29% true positive rate in detecting malicious web requests, highlighting its potential for strengthening web security frameworks (Yilmazer Demirel & Sandikkaya, 2023). The success of ZSL-CNN underscores the importance of deep learning techniques in improving the accuracy and reliability of anomaly detection in online environments. ZSL techniques generally follow a two-stage process: attribute learning and inference. Attribute learning involves mapping input features to a semantic space, while inference enables classification by associating these features with unseen categories (Rivero Pérez et al., 2017). This methodological flexibility has led to ZSL being applied across various domains, with researchers exploring different semantic representations and inference strategies to optimize performance (Wang et al., 2019). As cybersecurity threats continue to evolve, organizations must adopt proactive security measures. Understanding current cyber risks, implementing robust security strategies, and leveraging advanced technologies, including ZSL-based models, can significantly enhance threat detection and response capabilities (Laksana & Mulyani, 2024). By integrating ZSL into cybersecurity frameworks, organizations can improve their resilience against novel attacks while reducing reliance on extensive labeled datasets.

RESEARCH METHOD

1. Research Gap and Novelty

Previous studies in phishing email detection predominantly relied on conventional machine learning algorithms such as logistic regression, decision trees, and support vector machines. These approaches often required manual feature extraction and demonstrated limitations in interpreting the nuanced linguistic context inherent in email communications. More recent advancements introduced deep learning methods, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), capable of identifying more intricate data patterns yet still lacking comprehensive contextual language understanding.

Recent developments have started exploring Large Language Models (LLMs) such as GPT-4 for phishing email detection. These models exhibit superior capabilities in comprehensive textual analysis and provide explanatory context for their classifications, enhancing transparency and interpretability.

This research uniquely contributes to the existing body of literature by specifically evaluating and comparing the performance of newer specialized LLMs, namely Qwen version Qwen3-235B-A22B and Gemini models (2.5 Flash and 2.5 Pro) an area yet to be extensively investigated. Additionally, unlike many previous studies, this evaluation intentionally avoids further training or fine-tuning, reflecting realistic industry scenarios where immediate implementation and utility of existing LLMs are paramount. Thus, this study addresses both academic and practical gaps, providing valuable insights for deploying advanced LLM-based phishing detection systems effectively in real-world applications.

2. Research Design

This research employs an experimental comparative approach to evaluate and analyze the performance of Large Language Models (LLMs) in detecting phishing emails. Specifically, the study compares the capabilities of Qwen version Qwen3-235B-A22B and Gemini models (2.5 Flash and 2.5 Pro) in identifying phishing attempts directly, without additional model training.

3. Data Collection

The dataset utilized in this research is publicly available and comprises labeled email samples categorized clearly into phishing and non-phishing (legitimate) emails. The dataset contains characteristics representative of real-world email scenarios, such as emails mimicking financial institutions, social media platforms, online services, and personal communications.

The dataset is preprocessed by:

1. Cleaning and removing irrelevant metadata (e.g., email headers, timestamps).
2. Ensuring balance between phishing and legitimate emails to avoid biases.
3. Confirming dataset randomization to reflect real-world scenarios accurately.

4. Research Tools and Platforms

These models will be accessed directly via their respective API services without fine-tuning or additional training, using default configurations provided by the service providers.

5. Experimental Procedure

The systematic steps followed are:

1. Input Preparation: Cleaning and standardizing email content for direct input.
2. Model Evaluation : Submitting email content to each LLM (Gemini and Qwen) via APIs.
3. Performance Metrics Evaluation : Metrics include Accuracy, Precision, Recall, F1-score, and ROC-AUC.

RESULTS AND DISCUSSION

1. Data Collection

https://www.kaggle.com/datasets/subhajournal/phishingemails

The dataset utilized in this research is sourced from Kaggle, a public platform for data science projects, and is specifically titled "Phishing Email Detection." This dataset is publicly available and comprises a collection of emails explicitly labeled as either "Phishing Email" or "Safe Email." It contains two primary features: "Email Text," which includes the body of the email, and "Email Type," which serves as the label for classification. The dataset is designed to be representative of real-world scenarios, encompassing various deceptive tactics commonly found in phishing attacks. sample email at table 1

Table 1 Sample Email

No	Email	Label
1	re: equistar deal tickets - are you still available to assist robert with entering the new deal...	Safe Email
2	Hello I am your hot lil horny toy. I am the one you dream about, I am a very open minded pe...	Phishing Email

"Phishing_Email.csv" file has a size of 52.03 MB and is licensed under the GNU Lesser General Public License 3.0. The data is structured with a significant volume of email samples, with approximately 39% of the emails categorized as "Phishing Email" and the remaining 61% as "Safe Email." This distribution provides a substantial number of instances for both classes, which is crucial for training and evaluating the performance of machine learning models in detecting phishing attempts through text analytics.

2. Data Preprocessing

Prior to evaluation by the Large Language Models (LLMs), the dataset undergoes a comprehensive preprocessing pipeline to ensure data quality and suitability for the experimental procedure. This process is crucial for cleaning the data, handling inconsistencies, and preparing it in a format optimized for the models. The following steps are systematically applied:

1. Initial Data Loading and Inspection

   The dataset, stored in a CSV file named "Phishing.csv," is loaded into a pandas DataFrame. An initial inspection reveals the structure of the data, including the columns "Email Text" and "Email Type." A new feature, length, is created to store the character length of each email text to provide initial descriptive statistics.
2. Handling Missing and Duplicate Data

   The dataset is checked for missing values. It was found that the "Email Text" column contained 16 null entries. These rows with missing email text are removed to ensure that every entry has content to be analyzed. Subsequently, any duplicate rows are identified and removed from the dataset to prevent redundancy and potential bias in the evaluation. After these cleaning steps, the dataset comprises 18,634 unique email entries.
3. Label Encoding

   The categorical labels in the "Email Type" column, originally "Phishing Email" and "Safe Email," are converted into a numerical format. The LabelEncoder from scikit-learn is utilized for this transformation. The labels are encoded such that 1 represents a phishing email and 0 represents a safe email.
4. Data Subsetting and Balancing

   To create a manageable and balanced dataset for the zero-shot evaluation, a subset of the data is created. First, any rows where the 'Email Text' is simply the word 'empty' are filtered out. Then, 1,000 sample emails are randomly selected from the "Safe Email" category (labeled as 0). Similarly, 1,000 sample emails are randomly selected from the "Phishing Email" category (labeled as 1).
5. Final Dataset Creation and Randomization

   The two balanced samples (1,000 safe and 1,000 phishing emails) are concatenated to form a new DataFrame. This resulting DataFrame, containing a total of 2,000 emails, is then shuffled randomly to ensure that the order of emails does not influence the evaluation process. The final, preprocessed, and balanced dataset is saved

to a new CSV file named new_dataframe.csv. This file serves as the direct input for the experimental evaluation of the LLMs.

3. Result Prediction

After the dataset was processed and prepared according to the methodology previously described, a zero-shot evaluation was conducted on three different Large Language Models (LLMs): Gemini 2.5 Flash, Gemini 2.5 Pro, and Qwen 3. Each model's ability to classify 2,000 emails from the balanced dataset, consisting of 1,000 safe emails and 1,000 phishing emails, was tested.

The objective of this stage was to measure the raw accuracy of each model without fine-tuning, reflecting a practical usage scenario where models are directly applied to classification tasks. The prediction results from each model are summarized in the table and visualized in the graph below at figure 1 and detail at table 2.

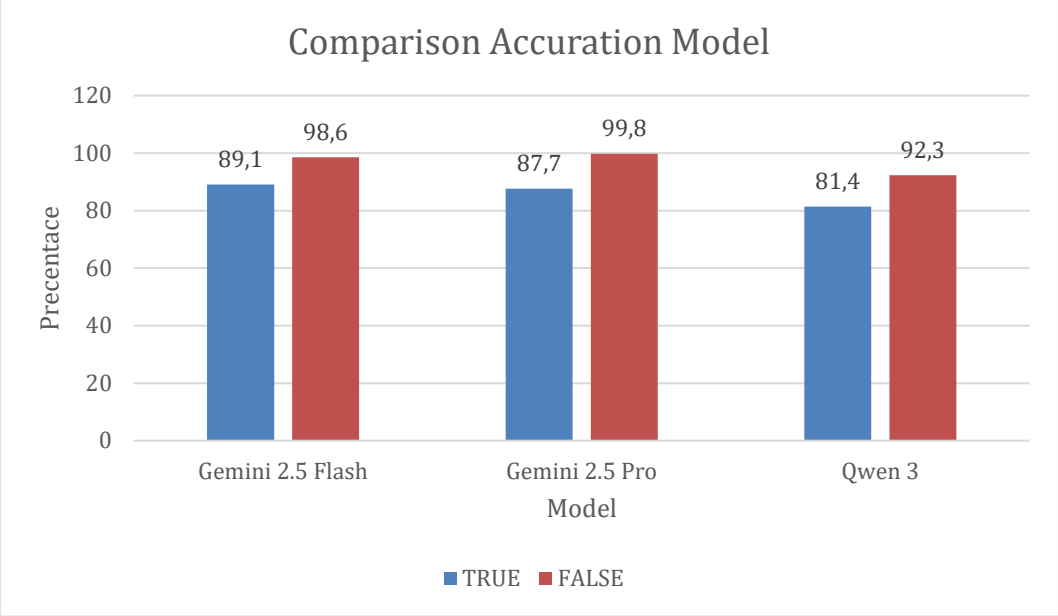

Figure 1 Comparison Accuration Model

Table 2 Data Calculation

Model	Scenario Type	Total Data	True Prediction	Accuration (%)
Gemini 2.5 Flash	Detection Safe (False)	1000	986	98,6
Gemini 2.5 Flash	Detection Phishing (True)	1000	891	89,1
Gemini 2.5 Pro	Detection Safe (False)	1000	998	99,8
Gemini 2.5 Pro	Detection Phishing (True)	1000	877	87,7
Qwen 3	Detection Safe (False)	1000	923	92,3
Qwen 3	Detection Phishing (True)	1000	814	81,4

4. Discussion

The results from the zero-shot evaluation offer a detailed and multi-faceted view of the current capabilities of advanced Large Language Models in the critical domain of phishing detection. The findings not only highlight significant performance differences between the models but also reveal a nuanced trade-off between security effectiveness and operational usability.

a. **The Paramount Importance of Minimizing False Positives**

The most remarkable result from this study is the exceptional performance of **Gemini 2.5 Pro** in identifying legitimate emails, achieving a staggering **99.8%** accuracy. In a practical context, this means that out of 1,000 safe emails, the model only misclassified two. This level of specificity is not merely a statistical achievement; it is a critical feature for real-world deployment. In any organization, the cost of false positives is substantial. When legitimate emails—such as client inquiries, internal directives, or system notifications—are incorrectly quarantined, it can lead to severe disruptions in workflow, loss of business opportunities, and a significant drain on IT resources who must manually review and release these emails. Furthermore, a high false positive rate erodes user trust in the security system, leading to "alert fatigue" where users begin to ignore warnings, paradoxically increasing the risk of a successful phishing attack. Gemini 2.5 Pro's ability to minimize this friction makes it an outstanding candidate for environments that prioritize seamless operations and user trust.

In comparison, Gemini 2.5 Flash, with an accuracy of 98.6%, would misclassify 14 safe emails out of 1,000, while Qwen 3, at 92.3%, would misclassify 77. While these numbers may seem small, at the scale of a large enterprise processing millions of emails daily, the difference becomes a matter of thousands of unnecessary support tickets, highlighting the tangible impact of Gemini 2.5 Pro's superior specificity.

b. **The Security vs. Usability Trade-off: Detecting the Actual Threat**

While Gemini 2.5 Pro excelled at recognizing safe correspondence, **Gemini 2.5 Flash** demonstrated a slight edge in correctly identifying malicious emails, with the highest phishing detection accuracy of **89.1%**. This metric, often referred to as sensitivity or True Positive Rate, is the cornerstone of a security system's primary function: to stop threats. Gemini 2.5 Flash correctly identified 891 out of 1,000 phishing attempts, while Gemini 2.5 Pro identified 877. This difference of 14 missed threats could be the difference between a minor incident and a major security breach.

This discrepancy illustrates a classic security trade-off. Gemini 2.5 Flash appears to be tuned with a higher sensitivity to potential threats, making it more effective at catching phishing emails, but at the cost of being slightly more likely to misinterpret an unconventional but legitimate email as malicious. Conversely, Gemini 2.5 Pro seems to be tuned for higher precision, making it more conservative in its judgments and thus less likely to raise a false alarm. The choice between these models, therefore, is not about which is "better" overall, but which aligns with an organization's specific risk appetite. A financial institution or government agency might prefer the higher threat detection rate of Flash, accepting the occasional false positive as a necessary price for enhanced security. A fast-moving commercial enterprise, however, might opt for Pro to ensure communication flows with minimal interruption.

c. **Interpreting Model Performance and Potential Hypotheses**

The performance of **Qwen 3**, which was consistently lower than both Gemini models, suggests that its pre-training data or architectural nuances may be less aligned with the specific linguistic markers and deceptive strategies common in the English-language phishing emails present in this dataset. It serves as a capable, yet more generalized, baseline against which the specialized performance of the Gemini models can be appreciated.

The differing strengths of the Gemini models themselves may hint at subtle variations in their training. One could hypothesize that Gemini 2.5 Pro's training data contained a vast corpus of formal, professional, and standard communication, honing its ability to recognize "normalcy" with high fidelity. In contrast, Gemini 2.5 Flash may have been exposed to a wider variety of informal, urgent, and syntactically diverse texts, making it more adept at spotting the unusual patterns—such as manufactured

urgency, grammatical errors, and suspicious calls-to-action—that are hallmarks of phishing, even if it sometimes over-corrects.

d. Implications for Zero-Shot Implementation

It is crucial to re-emphasize that these results were achieved in a **zero-shot** setting, without any fine-tuning on the specific dataset. This is a pivotal finding for practical application, as it demonstrates the immense out-of-the-box value of these models. The ability to deploy a highly effective phishing detection system without the need for extensive data collection, labeling, and model training democratizes access to state-of-the-art security, allowing even organizations with limited AI/ML resources to significantly bolster their defenses. The study validates that modern LLMs can serve as powerful, ready-to-use tools for cybersecurity.

CONCLUSION

1. Conclusion

This research conducted a zero-shot performance evaluation of three Large Language Models—Gemini 2.5 Pro, Gemini 2.5 Flash, and Qwen 3—on the task of phishing email detection. The findings reveal significant distinctions in their capabilities, highlighting a crucial trade-off between minimizing false positives and maximizing threat detection. Gemini 2.5 Pro demonstrated exceptional performance in correctly identifying legitimate emails, achieving an accuracy of 99.8%. This high level of specificity is paramount for enterprise environments, as it minimizes workflow disruptions and reduces the burden on IT support stemming from incorrectly quarantined emails.

Conversely, Gemini 2.5 Flash exhibited a superior ability to detect actual phishing attempts, with a detection accuracy of 89.1%, the highest among the evaluated models. This indicates a higher sensitivity to malicious content, which is critical for organizations prioritizing threat prevention above all else. The Qwen 3 model performed consistently below both Gemini models, suggesting its general pre-training may be less aligned with the linguistic nuances of English-language phishing emails.

Crucially, these results were achieved without any model fine-tuning, demonstrating the powerful "out-of-the-box" utility of modern LLMs for cybersecurity applications. The study validates that these advanced models can be deployed as effective, ready-to-use tools, democratizing access to state-of-the-art security for organizations with limited AI/ML resources.

2. Suggestions

Based on the evaluation, the selection of an LLM for phishing detection should be guided by an organization's specific risk appetite and operational priorities.

1. For organizations prioritizing operational continuity and user trust, such as commercial enterprises where uninterrupted communication is vital, Gemini 2.5 Pro is the recommended model. Its near-perfect accuracy in identifying legitimate emails ensures minimal disruption and prevents alert fatigue, fostering a more reliable security ecosystem.
2. For organizations with a lower risk tolerance for security breaches, such as financial institutions or government agencies, Gemini 2.5 Flash is the preferred choice. Its higher phishing detection rate provides a more aggressive defense against incoming threats, accepting a slightly higher rate of false positives as a necessary cost for enhanced security.
3. It is recommended that organizations implement a pilot program to test their chosen model in their specific email environment. This allows for an assessment of model performance against real-world data before full-scale deployment.
4. For future research, it would be beneficial to explore hybrid approaches that could potentially combine the high specificity of Gemini 2.5 Pro with the high sensitivity of Gemini 2.5 Flash. Furthermore, fine-tuning these models on domain-specific datasets

could yield even greater performance, tailored to the unique communication patterns of an individual organization.

REFERENCES

Andriu, A.-V. (2023). Adaptive phishing detection: Harnessing the power of artificial intelligence for enhanced email security. Romanian Cyber Security Journal, 5(1). https://doi.org/10.54851/v5i1y202301

Carroll, F., Adejobi, J. A., & Montasari, R. (2022). How good are we at detecting a phishing attack? Investigating the evolving phishing attack email and why it continues to successfully deceive society. SN Computer Science, 3, 170. https://doi.org/10.1007/s42979-022-01069-1

Chataut, R., Gyawali, P.K., & Usman, Y. (2024). Can AI Keep You Safe? A Study of Large Language Models for Phishing Detection. 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC), 0548-0554.

Demirel, D.Y., & Sandikkaya, M.T. (2023). Web Based Anomaly Detection Using Zero-Shot Learning With CNN. IEEE Access, 11, 91511-91525.

Desai, V., & Kavitha, R. (2024). Unveiling the depths of phishing: Understanding tactics, impacts, and countermeasures. International Journal of Innovative Research in Science, Engineering and Technology, 13(5). https://doi.org/10.15680/IJIRSET.2024.1305331

Greco, F., Desolda, G., Esposito, A., & Carelli, A. (2024). David versus Goliath: Can Machine Learning Detect LLM-Generated Text? A Case Study in the Detection of Phishing Emails. Italian Conference on Cybersecurity.

Hermawan, A. Z., & Pongoh, A. G. (2023). AI-powered Cyber Attacks: The rise of AI botnets in Indonesia. Pacific Journal of Information Systems Engineering, 1(2), 78-90. https://doi.org/10.47134/pjise.v1i2.2401

Laksana, T. G., & Mulyani, S. (2024). Pengetahuan dasar identifikasi dini deteksi serangan kejahatan siber untuk mencegah pembobolan data perusahaan. Jurnal Ilmiah Multidisiplin, 3(01), 109-122. https://doi.org/10.56127/jukim.v3i01.1143

Lee, J., Kim, S., & Park, H. (2024). Multimodal large language models for phishing webpage detection and identification. arXiv preprint arXiv:2408.05941. https://doi.org/10.48550/arXiv.2408.05941

Mahendru, A., & Pandit, A. (2024). Comparative analysis of transformer-based models for phishing detection. In Proceedings of the 2024 IEEE International Conference on Big Data and Artificial Intelligence (BDAI) (pp. 123–130). https://doi.org/10.1109/BDAI62182.2024.10692765

Putra, F. P. E., Ubaidi, Zulfikri, A., Arifin, G., & Ilhamsyah, R. M. (2024). Analysis of phishing attack trends, impacts, and prevention methods: Literature study. Brilliance: Research of Artificial Intelligence, 4(1). https://doi.org/10.47709/brilliance.v4i1.4357

Pongoh, A. G., & Rijal, L. (2024). Strengthening Cybersecurity with AI: Early threat detection and rapid response. Cybersecurity Journal, 7(1), 112-125. https://doi.org/10.14421/csecurity.2024.7.1.4486

Rijal, L., Salsabila, T., & Hermawan, A. Z. (2022). Artificial Intelligence for IT Operations (AIOps): Enhancing IT operations and cybersecurity. Journal of Information Technology, 15(2), 123-135. https://doi.org/10.1007/978-3-030-80821-1_2

Rivero, J., Ribeiro, B., Chen, N., & Leite, F. S. (2017). A Grassmannian approach to zero-shot learning for network intrusion detection. Proceedings of the International Conference on Neural Information Processing (ICONIP). https://doi.org/10.48550/arXiv.1709.07984

Salsabila, T., & Hermawan, A. Z. (2024). AI in Public Administration: Benefits and cybersecurity challenges. Sitasi: Jurnal Ilmiah, 3(1), 45-58. https://doi.org/10.33005/sitasi.v3i1.363

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), 367–384. https://doi.org/10.3390/make6010018

Wang, W., Zheng, V.W., Yu, H., & Miao, C. (2019). A Survey of Zero-Shot Learning. ACM Transactions on Intelligent Systems and Technology (TIST), 10, 1 - 37.