

Phishing Analysis And Countermeasures

FOUCHÉ Stanislas stanislasfouche@gmail.com

PRIOU Antoine priou2002@gmail.com

Tutor : KHOUKHI Lyes

Abstract

This paper aims to test the current effectiveness of large language models (LLMs) for email phishing detection. Our hypothesis is that LLMs could replace traditional methods in this field. To test this hypothesis, we have developed an application that allows users to forward a suspicious email to a dedicated address, where an LLM automatically analyzes its content. In return, the system provides a probability score indicating whether the email is considered phishing or not. An experimental evaluation is conducted on a dataset to analyze the model's performance in terms of precision, recall, and robustness against evasion attempts. Finally, we discuss the limitations of the approach, including sensitivity to false positives and the need for continuous adaptation to new attack strategies.

1 Introduction

In 2023, 16.73 million cybersecurity incidents related to phishing attacks were recorded, causing estimated losses of several billion dollars, according to Statista Research Department [26]. These attacks exploit human vulnerabilities to steal sensitive information such as credentials, banking data, or confidential documents.

Phishing is now a global threat, affecting both high-profile individuals and entire networks, such as companies, with a broader but less targeted approach. Its consequences can be devastating: financial losses, privacy breaches, and even the destabilization of entire organizations. Given this growing danger, it is crucial to understand its mechanisms, its impacts, and, most importantly, the means to protect against it.

Over the years, various detection methods have emerged, ranging from traditional approaches to AI-based solutions. With the advent of large language models (LLMs), a fundamental question arises: can these models replace traditional phishing detection methods? This paper aims to analyze the current performance of an LLM for phishing detection by comparing it to existing methods.

2 State of the Art

2.1 Structure of Attacks

The traditional phishing approach involves sending an email, an SMS, or making a phone call to trick the victim into voluntarily providing sensitive information. However, this method has significantly evolved over time. It now relies on the principle of social engineering, which involves studying the target's profile to psychologically manipulate them.

Cybercriminals exploit various techniques to gain their victims' trust using different methods depending on the targeted individuals. Several types of phishing attacks can be distinguished:

- Email phishing: The most common technique, impersonating legitimate entities to trick the victim into disclosing sensitive information through redirection to fraudulent sites.
- Spear phishing: Uses personalized information to make the attack more credible.
- Whaling: Targets high-profile individuals or companies.
- Vishing: Phishing via phone calls.
- Pharming: DNS redirection to a fraudulent website.
- Clone phishing: Duplication of a legitimate email with malicious attachments.

This paper specifically focuses on analyzing and detecting email phishing. Although phishing shares similarities with spam, particularly due to its mass distribution, it differs in its malicious intent and often precise targeting. Unlike spam, which is generally promotional or advertising-related, phishing aims to deceive the victim into extracting sensitive information such as credentials, banking details, or personal data. While all phishing types can technically be classified as a form of spam, the inverse is not true: not all spam falls under phishing, as its intent is not always to directly harm the victim.

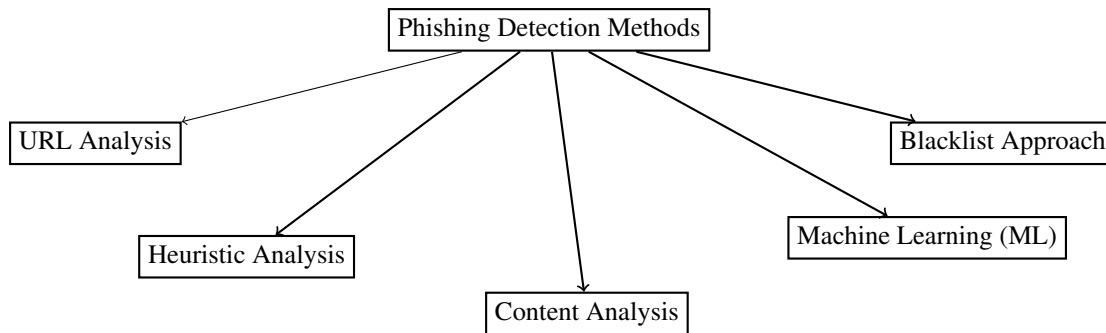


Figure 1: Existing Phishing Identification Methods

2.2 Detection Methods

Although there is some awareness around phishing, particularly through French government websites [19] highlighting key reflexes, it has proven to be an insufficient method as scams have evolved, especially in terms of techniques and social engineering. To protect against these attacks, software and experts rely on various tools and technologies to improve accuracy.

In the following subsections, we will discuss the main detection methods.

A set of general detection methods is presented in Figure 1.

2.2.1 Blacklist-Based Detection

One of the most classic methods relies on the use of known databases containing email addresses, phone numbers, and malicious links. These records are based on previously identified phishing cases or other relevant sources.

Official databases can be found online and are regularly updated by security organizations and threat intelligence services. An example is Google Safe Browsing, which catalogs and updates its blacklists for web browsers and all services linked to Google [8]. Uribl is another site that lists domain names of websites hosting viruses, malware, and spyware [27]. Some lists are available on certain websites: Infoservice [15].

This method is very useful and simple to implement, preventing the massive spread of phishing. However, as stated in Esposito [11], this method quickly reaches its limits against new phishing cases (also called 0-Day Phishing¹), since attackers can bypass this mechanism by creating new email addresses, domain names, or spoofing phone numbers.

2.2.2 Heuristic Rule-Based Detection

Many anti-phishing techniques rely on heuristic rules. Heuristic rule-based analysis is a detection method that identifies suspicious behaviors in emails based on predefined rules. Unlike blacklists, this method achieves a higher detection rate for new attacks. It identifies abnormal patterns or characteristics typical of phishing

attacks to assign a heuristic score to each email.

This heuristic score is determined by several factors, such as:

- The presence of suspicious keywords associated with phishing attempts (e.g., "account update," "urgent," "password").
- The verification of links and domains, particularly identifying deceptive domains or hidden links.
- The analysis of email headers to detect anomalies in sender information.
- The presence of typos or grammatical errors, often found in fraudulent emails.
- The use of obfuscation techniques², such as replacing characters with visually similar ones.

Unlike the previous method, heuristic rule-based detection adapts to new phishing attacks by identifying fraud patterns even when they are not yet listed in known threat databases.

However, according to some studies ScholarWorks [23], this approach may also present certain limitations, particularly an increase in false positives in cases where overly strict rules are applied.

To enhance detection, this heuristic analysis approach is often combined with machine learning for better performance.

[20]

2.2.3 Machine Learning

Machine Learning (ML) has demonstrated superior accuracy and performance compared to other detection methods. [3]

These models can predict new or unknown attacks based on existing data.

For phishing detection, ML uses a dataset composed of emails, SMS, and web pages classified as phishing or legitimate, enabling the recognition of characteristic phishing attack patterns.

¹Refers to a phishing attack exploiting an unknown or recent vulnerability before an effective defense is implemented

²Refers to methods designed to make detection or analysis of information more difficult, often by modifying its format or structure.

Reference	Dataset Used (Safe/Phishing)	Proposed Method	Precision (%)
Fette et al. [13]	6950 / 860	PILFER (LIBSVM - SVM public)	99.00
Abu-Nimeh et al. [1]	1700 / 1700	6 Classifieurs - LR, CART, BART, SVM, RF, NNet	95.11
Chandrasekaran et al. [7]	100 / 100	SVM à classe unique	95.00
Rathod and Patterwar [21]	2500 / 2100	Classificateur bayésien (Naïve Bayes)	96.46
Rawal et al. [22]	414 / 1605	Random Forest et SVM	99.87
Hota et al. [14]	Jeu de données public	RRFST avec C4.5 et CART	99.27
Mbah et al. [18]	6951 / 2357	KNN et Arbre de décision (J48)	93.11
Emilin Shyni et al. [10]	5260 / 0	Multi-classifieur - SVM, Random Forest, LogitBoost	96.30
Smadi et al. [24]	5000 / 5000	Algorithme de classification J48	98.11
Sonowal [25]	1604 / 1824	Sélection de caractéristiques par recherche binaire	97.41
Li et al. [17]	Jeu de données public	SVM avec AdaBoost	97.61
Jameel and George [16]	3000 / 3000	Réseau de neurones Feed Forward	98.72
Aljofey et al. [2]	Jeu de données public	Réseau de neurones récurrent convolutionnel	95.02
Fang et al. [12]	Combinaison de divers jeux de données publics	Modèle THEMIS basé sur les CNN	99.848
Bagui et al. [5]	14 950 / 3416	Réseau de neurones convolutionnel	95.97

Table 1: Performance of Machine Learning Models [9]

- Text analysis: presence of suspicious keywords, alarming tone, grammar and spelling mistakes.
- URL analysis: length, presence of special characters, suspicious redirects, comparison with legitimate domains.
- Email header analysis: inconsistency between sender and server domain.
- Attachment analysis: suspicious extensions, presence of malicious scripts.

According to the study by Dhruv Rathee [9], which consolidates several previous works (see Table 1), various machine learning methods have been used for phishing attack detection. These approaches include:

- SVM (Support Vector Machine)
- Logistic Regression
- Decision Trees
- Neural Networks
- Random Forests
- Bayesian Classifier
- k-Nearest Neighbors (k-NN)

Several of these models have demonstrated true positive rates exceeding 95%. However, despite these significant advancements, challenges remain, particularly in selecting relevant features and adapting to new attack techniques. These methods continue to evolve, aiming to further improve detection system accuracy and robustness. A scientific article [4] allows us to verify these statistics and claims once again.

Training requires a large volume of data, which can quickly become computationally expensive. Additionally, difficulties in interpreting model decisions may pose challenges. These disadvantages are explained in article [4].

The article concludes that, despite considerable improvements in phishing detection, feature selection and

adaptation to new attack techniques remain major challenges. However, these methods are still evolving towards even greater accuracy.

3 LLM

Large Language Models (LLMs) are increasingly being utilized due to their recent performance improvements, leveraging massive models and neural networks. The upcoming experiment aims to leverage the broad knowledge base of LLMs, utilizing their ability to efficiently understand text to analyze phishing emails. This includes detecting linguistic nuances, hidden intentions, sentence meaning, and manipulation techniques used in social engineering, making them well-suited for detecting both current and future sophisticated phishing attacks. Furthermore, they enable a single method to be used across all possible languages present in a potentially malicious email.

Compared to ML, LLMs can detect more complex phishing patterns that may not be explicitly present in training data, thanks to their linguistic understanding of content. This means they are, in many cases, capable of adapting to new and unknown phishing attack types.

3.1 Classification

The detection model's performance in our study is measured using classification metrics:

- **Precision:**

where represents the number of true positives (correctly identified phishing attempts), and the number of false positives (legitimate emails mistakenly identified as phishing).

4 Experimentation / Simulation

The design of this phishing detection service using LLM stems from our daily practice as developers. The frequent use of advanced models such as ChatGPT, LLaMA, or Claude in various development processes raised a central question: Could these models, initially designed for language generation and understanding

tasks, also be leveraged to replace complex tasks performed by specialized Machine Learning algorithms?

The objective of this experiment is to evaluate the feasibility of replacing a traditional machine learning model—often resource-intensive and time-consuming to train—with an LLM-based service. This approach could not only reduce training costs but also speed up processing while maintaining satisfactory phishing detection performance.

To achieve this, we developed an application allowing any user to verify the malicious nature of an email for free and within about ten seconds. This service uses the OpenAI API to automatically analyze email content and provide a probability score indicating whether it is a phishing attempt.

4.1 Spamurai

This service aims to facilitate our study of the current state of LLMs for phishing detection via email. It is also available for personal use.

The service process is as follows:

1. The user forwards a suspicious email to **Spamurai.analysis@gmail.com**
2. Extraction of email features (sender address, subject, content)
3. Processing and analysis by our servers via the OpenAI API to determine the email's threat level
4. Sending the query result to the user in response to their email

The application operates continuously to assist anyone uncertain about an email's authenticity.

4.2 Protocol

4.2.1 Dataset

To evaluate our approach's performance, we used the dataset [6] available on Kaggle.

This dataset includes 18,600 email samples distributed across three columns:

1. Email index in the dataset
2. Email body (*body*), containing the full email text
3. Email category, indicating whether it is a legitimate message (Safe Email) or a phishing attempt (Phishing Email)

Among which:

- 61% are considered safe emails (Safe Email*)
- 39% are identified as phishing attempts (Phishing Email)

4.2.2 Model

In this study, we chose to use OpenAI's 4o-mini model for phishing detection. Despite being smaller than more complex models, it offers a good trade-off between performance and cost at a rate of \$0.15 per million tokens.

4.2.3 Prompt

The prompt plays a central role in the effectiveness of our method. It serves as the instruction given to the model to guide its reasoning and ensure a relevant analysis.

For this study, we designed a prompt based on essential criteria for phishing detection:

- **Sender address:** Verification of suspicious domain names, unusual characters, or brand imitations.
- **Email subject:** Identification of urgency tactics or emotional manipulation.
- **Message content:** Detection of grammatical errors, requests for sensitive information, and suspicious links.
- **Psychological manipulation techniques:** Analysis of messages using authority figures or threats.
- **Urgency:** Presence of pressure to obtain a quick response.

The model returns only a score between 0 and 100, indicating the probability that the email is a phishing attempt, ensuring each prompt retrieves a clear result without ambiguity.

Although more complex prompts could provide better performance, they are often less reliable at scale due to increased sensitivity to false positives. Therefore, we opted for a simplified yet robust approach to ensure balanced detection.

4.2.4 Detection Threshold

The API returns a probability score between 0 and 100% indicating the likelihood of phishing. To optimize detection accuracy, an experiment determined an appropriate threshold. As illustrated in Figure 2, the highest precision rate was observed when the phishing classification threshold was set at 75. This threshold optimally balances correct phishing detection and misclassification errors.

5 Results

We evaluated the performance of our approach on a sample of 51 emails taken from the dataset, consisting of 32 non-phishing emails and 19 phishing emails.

Although this sample is relatively small compared to those used in traditional machine learning-based approaches, it is important to highlight that due to the general and robust nature of our model, such a sample remains adequate.

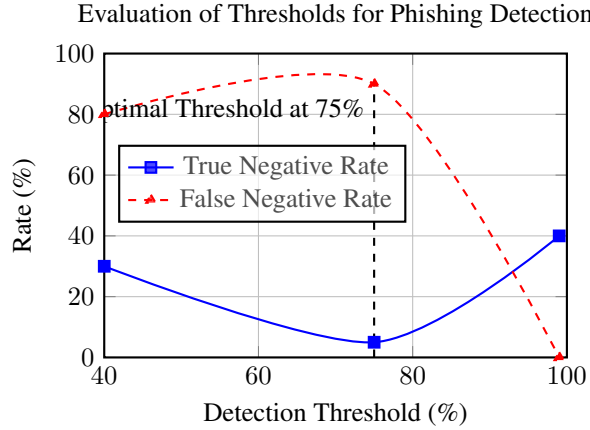


Figure 2: Impact of Different Thresholds on Detection Rates. An optimal compromise is achieved at a 75% threshold, providing a final accuracy of 90.19%.

Since the model is designed to handle a wide range of cases without requiring specific training cases, the results obtained do not heavily depend on the sample size. Moreover, due to the practical constraints of using an email detection service, testing on a larger set is challenging.

	Predictions: Safe	Predictions: Phishing	Total
Safe	27	5	32
Phishing	0	19	19

5.1 Analysis of Results

5.1.1 False Positives

For phishing emails, we observed very few false positives throughout our experiment (only when the threshold was > 95).

For non-phishing emails, a limitation in the LLM-based approach was identified due to model censorship. Some messages were misclassified as phishing due to internal factors related to model censorship, where the reasoning of the model was hindered by restrictions imposed during training.

5.1.2 True Positives

Regarding phishing emails, our model showed an excellent detection rate for characteristic phishing factors. The elements commonly used by cybercriminals to manipulate victims were effectively identified, leading to the correct classification of most malicious emails.

For legitimate emails, apart from the previously mentioned issues (related to false positives), the model demonstrated high accuracy in recognizing safe emails, with very few cases of misclassification. These results suggest that the model has a robust capability to distinguish legitimate emails from phishing attempts in most cases.

6 Limitations

Although the results are promising, some limitations must be considered. For example, the small size of the tested dataset (51 emails) may not fully reflect the diversity of phishing cases encountered in real-world environments. To address this, a service facilitating a smoother transition from dataset testing to email submission via another self-developed tool would be needed.

Additionally, using more expensive and efficient models could ensure greater accuracy in our research by leveraging multimodal models capable of processing email attachments (currently not analyzed). Finally, we observed that the usual bias of model overlays prevents reasoning on emails containing offensive, defamatory, or sexually explicit content, leading to their instant classification as phishing content.

7 Conclusion

In this study, we experimented with the effectiveness of LLMs for phishing detection. By comparing the results to traditional methods, we achieved superior performance, particularly in terms of adaptability to new attacks, thanks to their flexibility and semantic understanding of emails.

However, when compared with machine learning-based detection studies, the achieved accuracy was, on average, lower. This discrepancy may be due to the imprecision of the chosen model. Despite these results, LLMs are expected to improve in accuracy over time, as they hold great potential to surpass the current limitations of other methods, whether in phishing detection or other highly complex domains.

References

- [1] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair. 2007. Comparison of machine learning techniques for phishing detection. In *APWG eCrime Researchers Summit*, Pittsburgh, USA.
- [2] A. Aljofey, Q. Jiang, Q. Qu, M. Huang, and J.-P. Niyigena. 2020. An effective phishing detection model based on character level convolutional neural network from url. *Electronics*, 9(9):1514.
- [3] Eman Abdelfattah Ammar Odeh, Ismail Keshta. 2020. [Machine learning techniques for detection of website phishing: A review for promises and challenges](#). *IEEE Xplore*.
- [4] Anu Vazhayil, Harikrishnan NB, Vinayakumar R, Soman KP. 2020. [Phishing email detection using classical machine learning techniques](#). *Amrita School of Engineering*.
- [5] S. Bagui, D. Nandi, and R. J. White. 2021. Machine learning and deep learning for phishing e-mail classification using one-hot encoding. *Journal of Computer Science*, 17(7):610–623.
- [6] Subhadeep Chakraborty. 2022. [Phishing emails dataset](#). Accessed: January 23, 2025.
- [7] M. Chandrasekaran, K. Narayanan, and S. Upadhyaya. 2006. Phishing e-mail detection based on structural properties. In *First Annual Symposium on Information Assurance: Intrusion Detection and Prevention*, pages 2–8, New York.
- [8] Wikipedia Contributors. 2025. [Google safe browsing](#).
- [9] Suman Mann Dhruv Rathee. 2022. [Detection of e-mail phishing attacks – using machine learning and deep learning](#). *International Journal of Computer Applications*. Accessed: January 23, 2025.
- [10] C. Emilin Shyni, S. Sarju, and S. Swamynathan. 2016. A multi-classifier based prediction model for phishing emails detection using topic modelling, named entity recognition and image processing. *Circuits and Systems*, 7:2507–2520.
- [11] Andrea et d’autres auteurs Esposito. 2021. [An effective detection approach for phishing websites using url and html features](#). *PubMed Central*.
- [12] Y. Fang, C. Zhang, C. Huang, L. Liu, and Y. Yang. 2019. Phishing e-mail detection using improved rcnn model with multilevel vectors and attention mechanism. *IEEE Access*, 7:56329–56340.
- [13] I. Fette, N. Sadeh, and A. Tomasic. 2006. Learning to detect phishing e-mails. Technical report, Institute of Software Research International, School of Computer Science, Carnegie Mellon University.
- [14] H. Hota, A.K. Shrivastava, and Rahul Hota. 2018. An ensemble model for detecting phishing attack with proposed remove-replace feature selection technique. *Computer Science*, 132:900–907.
- [15] ABE Infoservice. 2025. [Listes noires des autorités](#).
- [16] Noor M. Jameel and Loay George. 2013. Detection of phishing e-mails using feed forward neural network. *International Journal of Computer Applications*.
- [17] Y. Li, Z. Yang, X. Chen, H. Yuan, and W. Liu. 2019. A stacking model using url and html features for phishing webpage detection. *Future Generation Computer Systems*, 94:27–39.
- [18] K. F. Mbah, A. H. Lashkari, and A. A. Ghorbani. 2022. A phishing e-mail detection approach using machine learning techniques. *World Academy of Science, Engineering and Technology, Computer and Information Engineering*, 3:2333.
- [19] Ministère de l’Économie, des Finances et de la Souveraineté industrielle et numérique. [Phishing, hameçonnage et filoutage](#).
- [20] Hong Qin, Ramprasath Jayaprakash, and Saurav et d’autres auteurs Mallik. 2021. [Heuristic machine learning approaches for identifying phishing threats across web and email platforms](#). *PubMed Central*.
- [21] Sunil B. Rathod and Tareek M. Pattewar. 2015. Content based spam detection in e-mail using bayesian classifier. In *IEEE ICCSP Conference*.
- [22] Srishti Rawal, Bhuvan Rawal, Aakhila Shaheen, and Shubham Malik. 2017. Phishing detection in e-mails using machine learning. *International Journal of Applied Information Systems*, 12:21–24.
- [23] California State University ScholarWorks. 2025. [Research paper on phishing detection techniques](#). *California State University ScholarWorks*.
- [24] Sami Smadi, Nauman Aslam, Li Zhang, Rafe Alasem, and M.A. Hossain. 2015. Detection of phishing e-mails using data mining algorithms. In *9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*.
- [25] G. Sonowal. 2020. Phishing e-mail detection based on binary search feature selection. *SN Computer Science*, 1.
- [26] Statista Research Department. 2024. [Number of cyberattacks worldwide per year](#).
- [27] URIBL. 2025. [Uribl - universal resource identifier blacklist](#).