Journal of Discrete Mathematical Sciences & Cryptography
ISSN 0972-0529 (Print), ISSN 2169-0065 (Online)
Vol. 28 (2025), No. 5-A, pp. 1621–1638
DOI : 10.47974/JDMSC-2162

Robust cybersecurity through discrete and large language models for effective phishing attack detection

Dinesh Goyal [@]
Anil Kumar [†]
Priya Mathur [§]
Department of Computer Science and Engineering
Poornima Institute of Engineering & Technology
Jaipur 302022
Rajasthan
India

Farhan Sheth [‡]
Amit Kumar Gupta *
Department of Computer Science & Engineering
Manipal University Jaipur
Jaipur 303007
Rajasthan
India

Abstract

This study explores the efficacy of four Large Language Models (LLMs)—BERT, DistilBERT, RoBERTa, and DeBERTa—in classifying URLs as either legitimate or phishing. The research methodology is structured into three phases: dataset processing, model fine-tuning, and performance evaluation. Each LLM is fine-tuned to distinguish between phishing and legitimate URLs. The models are evaluated using both a primary dataset with extensive features and an external dataset with minimal features to rigorously assess their robustness. The models consistently achieved high performance on the primary dataset, with AUC scores of 0.99, indicating near-perfect discrimination between phishing and legitimate URLs. DistilBERT excels in F1-score (99.992%), accuracy (99.991%), and precision (99.985%), showcasing its efficiency in real-world applications. BERT and DeBERTa also

[@] *E-mail:* dinesh.goyal@poornima.org
[†] *E-mail:* anilkumar@poornima.org
[§] *E-mail:* drpriyamathur21@gmail.com
[‡] *E-mail:* farhansheth.jb@gmail.com
* *E-mail:* dramitkumargupta1983@gmail.com; (Corresponding Author)
amit.gupta@jaipur.manipal.edu

demonstrate excellent results, while RoBERTa, though slightly lower in precision, remains competitive. The model's performance was also evaluated on an external test dataset containing 450,176 labeled URLs, the dataset helped access the model's performance under extremely constrained conditions. BERT and DeBERTa showed low F1-scores (1.14% and 2.78%, respectively) despite high precision, indicating poor recall. DistilBERT performs moderately better with a 42.44% F1-score, while RoBERTa achieves the highest F1-score of 68.03%, suggesting superior balance between precision and recall under severe feature constraints. Overall, while all models exhibit strong performance with rich features, their ability to maintain efficacy under limited feature conditions varies. This study underscores the importance of developing models with robust performance across diverse scenarios, highlighting RoBERTa's superior recall in feature-scarce environments and DistilBERT's overall efficiency.

Subject Classification: *68T07, 68P25.*

Keywords: *Phishing detection, URL security, Cybersecurity, Natural language processing, BERT, DistilBERT, RoBERTa, DeBERTa.*

1. Introduction

In Phishing attacks, a prevalent form of cyber threat, exploit deceptive techniques to deceive users into divulging sensitive information. These attacks often manifest as fraudulent emails, deceptive websites, or malicious links, aiming to compromise personal and organizational security. As phishing tactics become increasingly sophisticated, there is a growing need for advanced detection methods capable of distinguishing between legitimate and malicious URLs effectively. Traditional methods relying on heuristic or rule-based approaches are often inadequate in the face of evolving phishing strategies, necessitating the adoption of more advanced techniques. In recent years, deep learning approaches have emerged as a powerful tool for enhancing phishing detection capabilities. By leveraging vast amounts of data and complex feature representations, deep learning models can uncover subtle patterns and anomalies indicative of phishing activities. Among these models, Large Language Models (LLMs) such as BERT (Bidirectional Encoder Representations from Transformers), DistilBERT, RoBERTa (Robustly optimized BERT), and DeBERTa (Decoding-enhanced BERT with Disentangled Attention) have demonstrated remarkable performance in various natural language processing tasks. Their ability to process and interpret contextual information from textual data makes them promising candidates for URL classification tasks. BERT, introduced by [9], revolutionized language understanding by pre-training deep bidirectional representations, enabling the model to capture context from both directions [9]. DistilBERT, a more compact version of BERT, retains high performance while reducing computational requirements [10]. RoBERTa, an optimized variant of BERT, further refines the training process to achieve superior performance [11]. DeBERTa enhances BERT's architecture with disentangled attention mechanisms and an improved mask decoder, yielding

even greater accuracy and efficiency [9-12]. This paper aims to investigate the effectiveness of these LLMs in classifying URLs as either legitimate or phishing. By employing a comprehensive methodology that includes dataset processing, model fine-tuning, and rigorous performance evaluation, the study assesses the capabilities of BERT, DistilBERT, RoBERTa, and DeBERTa in distinguishing between malicious and non-malicious URLs. The evaluation is conducted using both a primary dataset with extensive features and an external dataset with minimal features, providing insights into the models' robustness under varying conditions. Previous research has demonstrated the potential of deep learning models in phishing detection. Yang, Zhao, and Zeng [1] employed multidimensional features and deep learning to enhance detection accuracy. Yi et al. [2] highlighted the adaptability of deep learning frameworks to phishing detection. Siddiq, Arifuzzaman, and Islam [3] further supported the efficacy of deep learning models in detecting phishing websites. However, the performance of these models under constrained feature scenarios remains underexplored. This study addresses this gap by evaluating the performance of LLMs in phishing detection, providing a comparative analysis of their effectiveness in different feature contexts. The findings contribute to the ongoing discourse on enhancing phishing detection mechanisms and offer practical insights for deploying advanced models in real-world applications.

2. Related Study

Recent Phishing detection has increasingly leveraged advanced machine learning techniques, with deep learning approaches gaining significant attention due to their effectiveness in identifying fraudulent activities. This literature review contextualizes recent advancements in phishing detection, particularly through deep learning models and Large Language Models (LLMs), in relation to the current study's analysis of URL classification using BERT, DistilBERT, RoBERTa, and DeBERTa. [1] explored phishing website detection by utilizing multidimensional features and deep learning models. Their approach highlights the efficacy of integrating various features to enhance detection accuracy. The study demonstrates that deep learning models, when combined with diverse features, can achieve robust phishing detection. This aligns with our findings that LLMs, with their ability to process extensive features, excel in distinguishing between phishing and legitimate URLs. [2] further investigated web phishing detection using a deep learning framework, emphasizing the adaptability of deep learning techniques to the evolving tactics of phishing attacks. Their work supports the notion that deep learning models can adapt to new phishing strategies, a feature corroborated by our results where models like BERT and its variants effectively classify URLs with high precision. [3] also contributed to the domain by applying deep learning techniques to phishing website detection. Their research reinforces the utility of deep learning models in improving detection capabilities, consistent with the high-performance metrics observed in our study's primary dataset. Recent advancements in the field have also explored the boundaries of model performance under constrained

conditions. [4] investigate machine learning's ability to detect LLM-generated text, including phishing emails. This study underscores the challenge of distinguishing between human and machine-generated content, relevant to our findings where LLMs, though highly effective with rich features, exhibit variable performance when faced with limited data. [5] provide a comparative study on prompt engineering versus fine-tuning for phishing detection with LLMs. Their work is pertinent to our analysis, as it highlights the effectiveness of fine-tuning models like BERT and its variants for specific tasks such as phishing detection. Our study's emphasis on fine-tuning aligns with their conclusions, demonstrating that tailored fine-tuning enhances model performance in classifying URLs. [6] present a comparative study on LLMs in large-scale organizational settings, focusing on spear phishing. Their research complements our findings by illustrating the practical application of LLMs in real-world scenarios, reinforcing the strength of models like DistilBERT and RoBERTa in handling sophisticated phishing attacks. In addition, the PhiUSIIL dataset introduced by [7] provides a benchmark for phishing URL detection, offering a valuable resource for evaluating model performance. Our study's use of this dataset for comparative analysis emphasizes the importance of diverse datasets in assessing model robustness. [8] conducted a comparative analysis of machine learning-based phishing detection using URL information, providing insights into the limitations and strengths of various models. This analysis is aligned with our findings, where the models' performance varied significantly between rich and minimal feature datasets. Lastly, the foundational works on BERT, DistilBERT, RoBERTa, and DeBERTa offer critical insights into the development and optimization of these models [9-14]. Our study builds on these advancements by applying these models to phishing detection, demonstrating their efficacy and the impact of model architecture on performance [15-18]. In conclusion, the literature highlights the significant advancements in phishing detection through deep learning and LLMs. Our study's findings contribute to this body of work by providing a detailed comparison of model performance across different datasets and feature sets, emphasizing the importance of model adaptability and robustness in real-world applications.

3. Methods and Material

The research aims to investigate and leverage Large Language Models (LLMs) such as BERT, DistilBERT, RoBERTa, and DeBERTa for classifying URLs or Uniform Resource Locator as either legitimate or phishing. The methodology is divided into three main phases. The initial phase involves processing the dataset to extract and combine the most significant features into a single text input for the LLMs. The second phase focuses on fine-tuning the models for the classification task. The final phase evaluates the model's performance using an external dataset with a minimal set of features. The overview of the methodology is shown in Figure 1. To elucidate the procedural intricacies, Algorithm 1 encapsulates the algorithmic representation employed in the methodology.

Figure 1
Overview of the methodology

Algorithm 1. Methodology of the study
Input:
Dataset D
Output:
Classification Results C_r (Phishing or Legitimate URL)
Steps:
Step 1. Data Acquisition:

$$D_{input} \leftarrow \text{Collect dataset for analysis}$$
$$D_{input} = \{(x_1, y_1), (x_2, y_2), \dots\dots\dots (x_n, y_n)\}$$
$$where\ x_i\ is\ the\ feature\ vector\ and\ y_i \in \{phishing, legitimate\}$$

Step 2. Feature Selection (Preprocessing):

$$F_{selected} = \{f_1, f_2, \dots\dots\dots f_m\}\ where\ f_i \in D_{input}$$
$$F_{selected} \leftarrow Select\ critical\ features\ based\ on\ real - time\ availability\ and\ significance$$

Step 3. Feature Aggregation (Preprocessing):

$$T_{aggregated} = concat\ \{f_1, f_2, \dots\dots\dots f_m\}\ where\ f_i \in F_{selected}$$
$$T_{aggregated} \leftarrow Merge\ F_{selected}\ in\ to\ a\ single\ text\ format$$

Step 4. Data Splitting (Preprocessing):

$$\{D_{train}, D_{val}, D_{test}\} \leftarrow Split\ T_{aggregated}\ in\ to\ training, validation\ and\ testing\ stes$$
$$D_{train} = \{(x_1, y_1), (x_2, y_2), \dots\dots\dots (x_k, y_k)\}$$
$$D_{val} = \{(x_{k+1}, y_{k+1}), (x_{k+2}, y_{k+2}), \dots\dots\dots (x_p, y_p)\}$$
$$D_{train} = \{(x_{p+1}, y_{p+1}), (x_{p+2}, y_{p+2}), \dots\dots\dots (x_q, y_q)\}$$

Step 5. Model Training and Evaluation:

- BERT (Bidirectional Encoder Representations from Transformers):

$$H = BERT(x) = Transformer\ (E(x))$$
$$\hat{y} = Softmax\ (WH_{[CLS]} + b$$
$$L_{BERT} = CrossEntropyLoss\ (\hat{y}, y)$$

- DistilBERT (Distilled BERT):

$$H_{DistilBERT} = DistilBERT(x) = DistilledTransformer\ (E(x))$$
$$\hat{y}_{DistilBERT} = Softmax\ (W_{DistilBERT} H_{DistilBERT\,[CLS]} + b_{DistilBERT}$$
$$L_{DistilBERT} = CrossEntropyLoss\ (\hat{y}_{DistilBERT}, y)$$

- RoBERTa (Robustly Optimized BERT Pretraining Approach):

$$H_{RoBERTa} = RoBERTa(x) = Transformer\ (E_{RoBERTa}(x))$$
$$\hat{y}_{RoBERTa} = Softmax\ (W_{RoBERTa} H_{RoBERTa\,[CLS]} + b_{RoBERTa}$$
$$L_{RoBERTa} = CrossEntropyLoss\ (\hat{y}_{RoBERTa}, y)$$

- DeBERTa (Decoding-enhanced BERT with disentangled attention):

$$H_{\text{DeBERTa}} = \text{DeBERTa}(x) = DisentangledTransformer\ (E_{\text{DeBERTa}}(x))$$

$$\hat{y}_{\text{DeBERTa}} = Softmax\ (\text{W}_{\text{DeBERTa}}\text{H}_{\text{DeBERTa}[CLS]} + b_{\text{DeBERTa}}$$

$$L_{\text{DeBERTa}} = CrossEntropyLoss\ (\hat{y}_{\text{DeBERTa}}, y)$$

- General Model Training and Evaluation:

$$for\ each\ model\ M \in \{BERT,\ DistilBERT,\ RoBERTa,\ DeBERTa\}:$$

$$M_{trained} \leftarrow Train\ M\ on\ D_{train}$$

$$\hat{y}_{tset} = M_{trained}\ (D_{test})$$

$$P_{test} = Evalaute\ (\hat{y}_{tset}, y_{test})$$

$$Metrics\ used: AUC, F1 - score, Accuarcy, Precision, Sensitivity$$

Step 6. External Dataset Testing:

$$\hat{y}_{external} = M_{trained}\ (D_{external})$$

$$P_{external} = Evalaute\ (\hat{y}_{external}, y_{external})$$

$$where\ P_{external}\ are\ the\ true\ lables\ for\ the\ external\ dataset$$

Step 7. Result Analysis:

$$Analyze\ P_{test}\ and\ P_{external}\ results\ from\ different\ models$$

Step 8: Final Output:

$$C_r = Final\ classification\ result\ from\ the\ analysis$$

3.1 *Dataset Characteristics*

In this study, the primary dataset utilized is a publicly available resource known as the 'PhiUSIIL Phishing URL' dataset [7], which is referred to as primary or PhiUSIIL throughout the research. The dataset includes 54 features, excluding the URLs and their respective labels. These features offer distinct insights into the characteristics of the URLs and are listed in Table 1. The dataset consists of 134,850 legitimate URLs and 100,945 phishing URLs, totaling 235,795 URLs. Although the dataset is imbalanced, the substantial number of URLs helps mitigate the impact of this imbalance. The study also utilizes an additional external dataset for testing purposes. This dataset, referred to as the 'Phishing URL dataset' from Mendeley [19], is simply called the external dataset in this research. It is relatively straightforward, containing only the URLs and their classifications into phishing and legitimate categories. In total, the dataset comprises 345,738 legitimate URLs and 104,438 phishing URLs, making it a comprehensive dataset with a total of 450,176 URLs. This external dataset is employed to assess the models' performance on a large scale, with a limited set of features, thereby presenting a challenging evaluation scenario but providing the most rugged evaluation of models. The distribution of both datasets is illustrated in Figure 2. These features of the URLs provide a foundational framework for further analysis, facilitating a structured approach for identification task [8].

Table 1
Features in PhiUSIIL dataset

URL	URL Length	Domain	Domain Length	Is Domain IP
TLD	URL Similarity Index	Character Continuation Rate	TLD Legitimate Probability	URL Character Probability
TLD Length	No. Of Subdomain	Has Obfuscation	No. Of Obfuscated Character	Obfuscation Ratio
No. Of Letters In URL	Letter Ratio In URL	No. Of Digits In URL	Digit Ratio In URL	No. Of Equals In URL
No. Of QMark In URL	No. Of Ampersand In URL	No. Of Other Special Characters In URL	Special Characters Ratio In URL	Is HTTPS
Line Of Code	Largest Line Length	Has Title	Title	Domain Title Match Score
URL Title Match Score	Has Favicon	Robots	No. Of Popups	No. Of iFrame
Has External Form Submit	Has Social Net	Has Submit Button	Has Hidden Fields	Has Password Field
Bank	Pay	Crypto	Has Copyright Info	No. Of Image
No. Of CSS	No. Of JS	No. Of Self References	No. Of Empty References	No. Of External References
Is Responsive	No. Of URL Redirect	No. Of Self Redirect	Has Description	

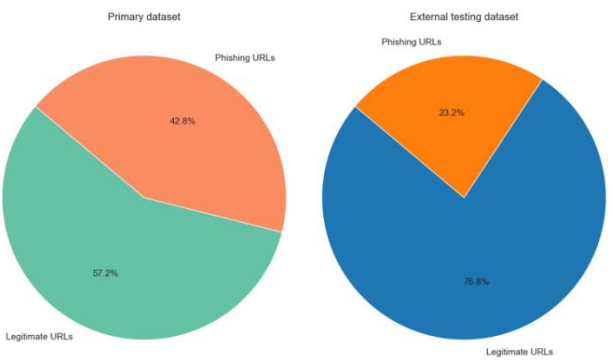

Figure 2
Distribution of URLs in the datasets

3.2 *Preprocessing*

Preprocessing is a vital component of Machine Learning, encompassing many processes such as cleaning, enhancement, and preparation of the dataset for model training. This multifaceted process involves several steps that address various issues within the dataset. Effective preprocessing enhances the dataset's

quality, contributing significantly to the models' reliability, robustness, and overall performance.

3.2.1 Feature Selection

The primary dataset, referred to as PhiUSIIL, comprises 54 features. However, in real-world scenarios, it may not always be possible to access all these features available in the dataset. Therefore, the most critical features were selected based on their real-time availability and significance in decision-making. A total of 12 features, excluding the URL and classification label, were identified as essential. These features are detailed in Table 2. The correlation matrix was employed to analyze the dependencies and complex relationships among the selected features. Figure 3 illustrates the correlation matrix, which excludes text-based features such as URL, TLD, Title, and Label. The matrix reveals strong correlations, particularly between 'Number of External References' and 'Number of Self References,' 'Has Copyright Info' and 'URL Similarity Index,' as well as 'Has Description' and 'URL Similarity Index.' This analysis provides valuable insights into the interdependencies among the selected features, offering a deeper understanding of their contributions to the model's performance.

Table 2
Features selected

URL	URL Length	TLD	Title	Is HTTPS
Is Responsive	Domain Length	URL Similarity Index	Has Description	No. Of URL Redirect
No. Of Self References	No. Of External References	Has Copyright Info	**Label**	

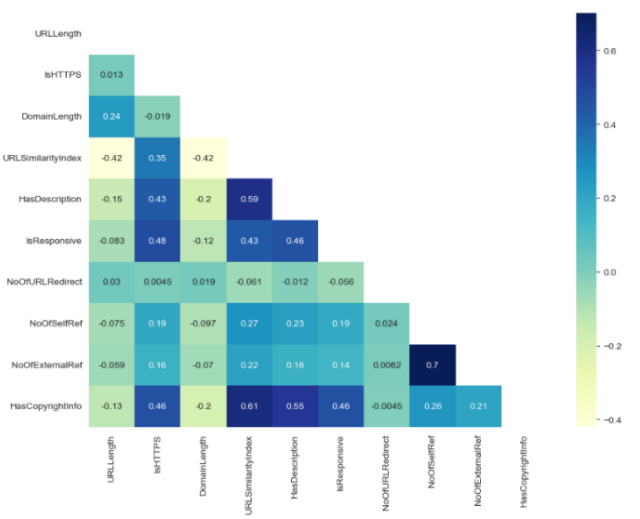

Figure 3
Correlation matrix of numerical features selected

3.2.2 Feature Aggregation

The Large Language Models (LLMs) employed in this study require a single text input, making it impractical to provide individual features separately. As a result, all the features were merged into a single text format. A sample of the resulting dataset is presented in Table 3. For the external test dataset, features that could be derived from the URL, such as the URL length and whether the URL is HTTPS protected, were similarly concatenated. Additionally, if the URL length exceeded a specified threshold, it was truncated to meet the length requirements of the LLMs. This approach ensured that the models could effectively process the input data while adhering to the necessary input constraints.

Table 3
Sample preview of the dataset after feature aggregation

Text	Label
URL is 'https://www.sfnmjournal.com', the length of URL is 26, the TLD is 'com', the title is 'home page: seminars in fetal and neonatal medicine', the URL	Legitimate
URL is 'http://www.teramill.com', the length of URL is 22, the TLD is 'com', the title is '0', the URL is not HTTPS protected, the length of domain is 16, the URL	Phishing
URL is 'https://www.bulgariaski.com', the length of URL is 26, the TLD is 'com', the title is 'bulgaria ski - bulgarian ski resorts and holiday b', the URL	Legitimate
URL is 'https://service-mitld.firebaseapp.com/', the length of URL is 37, the TLD is 'com', the title is '0', the URL is HTTPS protected, the length of domain is 29, the URL ……	Phishing

3.2.3 Dataset Splitting

The primary dataset was divided into three subsets: a training set comprising 70%, a validation set consisting of 15%, and a testing set also accounting for 15%. The training set was used to train the models, while the validation set was employed to validate the models during the training process. The testing set played a crucial role in evaluating the model's effectiveness on unseen data, providing a measure of its performance in real time scenarios.

3.3 *Large Language Models (LLM)*

Large Language Models (LLMs) such as BERT and its derivatives—DistilBERT, RoBERTa, and DeBERTa—have revolutionized the field of natural language processing by demonstrating unparalleled capabilities in understanding and generating human-like text. These models excel in capturing complex linguistic structures, intricate patterns, and subtle relationships within data. This makes them exceptionally well-suited for a wide range of applications, including the classification of URLs into phishing or legitimate. One of the key advantages of LLMs over traditional machine learning algorithms is their ability to generalize from large amounts of data and learn contextual cues. Traditional

algorithms often rely heavily on handcrafted features, and their performance can degrade significantly if the feature set is reduced. In contrast, LLMs, with their vast pretrained knowledge, can effectively interpret and analyze even limited inputs, such as raw URLs, without the need for extensive feature engineering. This is particularly beneficial in scenarios where feature extraction is challenging, data is unavailable or where data is unstructured [20].

In our study, the models were trained on URLs and their features, but we also harnessed the power of LLMs by testing them on a comprehensive external dataset, providing only the raw URLs without any additional features. The results underscored the models' superior ability to discern phishing attempts from legitimate URLs, even in the absence of explicit feature sets. This success can be attributed to the models' deep understanding of language semantics and syntax, allowing them to identify subtle indicators of malicious intent that might be overlooked by traditional methods [21].

Overall, LLMs represent a significant advancement in the field of machine learning, offering robust solutions for complex classification tasks. Their capacity to handle minimal and unstructured data, combined with their deep learning capabilities, makes them an invaluable tool in cybersecurity and beyond.

3.3.1 Models

The study utilizes four distinct Large Language Models (LLMs), specifically BERT and its variants DistilBERT, RoBERTa, and DeBERTa. BERT, or Bidirectional Encoder Representations from Transformers, is notable for its method of pre-training deep bidirectional representations on unlabeled text, enabling it to grasp context from both directions at each layer. This thorough approach results in a model that can be fine-tuned, with the addition of a single output layer, to achieve state-of-the-art performance across a wide range of tasks, including question answering and language inference. BERT's architecture is designed to be flexible, obviating the need for extensive task-specific architectural modifications, thus making it a highly versatile tool in the field of Natural Language Processing (NLP) [9]. DistilBERT, a more compact and efficient version of BERT, utilizes knowledge distillation during the pre-training phase, unlike previous approaches that focused on distilling task-specific models. This process reduces the size of the BERT model by 40%, while retaining 97% of its language understanding capabilities and increasing its speed by 60%. DistilBERT employs a triple loss function, integrating language modeling, distillation, and cosine-distance losses to effectively capture the inductive biases of larger models. Its reduced size and enhanced efficiency make DistilBERT particularly cost-effective for pre-training and suitable for on-device computations [10].

RoBERTa, an optimized variant of BERT, enhances the original model's training methodology, achieving and surpassing the performance of subsequent models. RoBERTa sets new benchmarks on datasets like GLUE, RACE, and SQuAD, underscoring the significance of often overlooked design choices in

model development [11]. DeBERTa (Decoding-enhanced BERT with Disentangled Attention) introduces two key innovations that build upon the foundations of BERT and RoBERTa. The first is the disentangled attention mechanism, where each word is represented by two separate vectors: one for content and another for position. The attention weights are calculated using disentangled matrices that consider both these vectors. The second innovation is an improved mask decoder, which replaces the conventional softmax output layer for predicting masked tokens during pre-training. These advancements significantly enhance the model's efficiency during pre-training and its performance on downstream tasks. Remarkably, DeBERTa, trained on only half the data used for RoBERTa-Large, consistently outperforms RoBERTa-Large across various benchmarks, demonstrating its superior effectiveness [12].

3.3.2 Model Optimization

For the model configurations, the official tokenizers corresponding to each respective model were employed, with the maximum sequence length set to 256 tokens. Padding and truncation were enabled to ensure consistent tokenized inputs throughout the training process. The models were initialized using their official pre-trained weights from the Hugging Face library [13]. A batch size of 32 was used for all models except for DeBERTa, for which the batch size was set to 16. The training process spanned 2 epochs, utilizing the AdamW optimizer, also known as ADAM weight decay, a variant of the ADAM optimizer frequently used in training transformers [14]. ADAM is an algorithm designed for first-order gradient-based optimization of stochastic objective functions, with adaptive estimates of lower-order moments. It combines the benefits of AdaGrad and RMSProp, making it effective in both online and non-stationary settings. One of ADAM's key features is that its parameter updates are invariant to gradient rescaling, with step sizes bounded by a stepsize hyperparameter. Moreover, ADAM does not require a stationary objective and handles sparse gradients proficiently [15]. The AdamW optimizer enhances this approach by decoupling weight decay from gradient updates, improving regularization and training dynamics. This separation helps in better regularizing the models and fine-tuning training dynamics [14]. The learning rate and weight decay are critical hyperparameters that control the step size during training and play a significant role in optimizing the models. For optimal training performance, these hyperparameters were carefully selected according to the official recommendations. This meticulous tuning ensures that the models are trained efficiently and effectively, leveraging the full potential of the optimization algorithm.

3.3.3 Fine Tuning

To enable the models to accurately distinguish between phishing and legitimate URLs, they must be trained on a relevant dataset. This process of adapting a pre-trained model to a specific task is known as fine-tuning. The primary goal is to equip the Large Language Models (LLMs) with the ability to

recognize and understand the intricate details and patterns of URLs, which the original pre-trained model might not be capable of discerning. The fine-tuning process begins by feeding the dataset into a tokenizer, which generates input IDs, token type IDs, attention masks, and associated labels. These tokenized inputs are then passed to the models, allowing them to make predictions. After making predictions, the loss is computed by comparing the actual labels with the predicted ones. During the backward pass, gradients of the loss are calculated for the model's parameters (weights and biases). These gradients are then used to update the model parameters in an effort to minimize the loss, thereby enhancing the model's performance. This iterative process is repeated over multiple epochs, gradually refining the model's accuracy and robustness. Through this continuous adjustment, the model learns to identify patterns and nuances in the data, enabling it to generalize well to unseen data. As a result, the fine-tuned model becomes proficient at applying its acquired knowledge to new, previously unencountered URLs, improving its effectiveness in real-world scenarios.

3.3.4 Model Evaluation

To thoroughly evaluate the performance of the LLMs on both the test dataset and the external test dataset, a range of metrics were calculated based on the models' outputs. The models underwent extensive and detailed testing using a comprehensive set of evaluation metrics to ensure a robust assessment. The key metrics computed include AUC (Area Under the ROC Curve), Accuracy, Sensitivity, Specificity, Precision, and F1-score. These metrics provide a multifaceted view of the models' effectiveness, capturing various aspects of their predictive capabilities. The formulas for these metrics are as follows:

$$Accuracy = \frac{count\ of\ accurate\ predictions}{overall\ count\ of\ samples} \quad \#(1)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad \#(2)$$

$$Specificity = \frac{TN}{TN + FP} \quad \#(3)$$

$$Precision = \frac{TP}{TN + FP} \quad \#(4)$$

$$F1 - score = \frac{2 \times \text{Precision} \times \text{Sensitivity}}{Precision + Sensitivity} \quad \#(5)$$

The terms TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

4. Results and Discussion

The PhiUSIIL or the primary dataset consists of 235,795 URLs, distributed into 134,850 legitimate URLs and 100,945 phishing URLs. The dataset was strategically partitioned into three subsets: a 70% training set comprising 165,056 URLs and a 15% for both validation and testing set containing 35,369

URLs and 35,370 URLs respectively. The parameters employed are showcased in Table 4.

Table 4
Parameters settings for the LLMs

Sr. No.	Parameter	Value	Sr. No.	Parameter	Value
1	Optimizer	AdamW	**4**	Learning rate	2e-5
2	Training Batch size	32, 16 (DeBERTa)	**5**	Weight decay	5e-4
3	Testing Batch size	32, 16 (DeBERTa)	**6**	Number of epochs	2

On the showcased parameters settings, all the models were fine-tuned on the primary dataset, the comprehensive results of the models on the primary dataset's testing subset are given in Table 5.

Table 5
LLMs results on the primary dataset's testing subset

Model Name	AUC	F1-score (%)	Accuracy (%)	Precision (%)	Sensitivity (%)
BERT	0.99	99.990	99.988	99.980	100
DistilBERT	0.99	99.992	99.991	99.985	100
RoBERTa	0.99	99.987	99.985	99.975	100
DeBERTa	0.99	99.990	99.988	99.980	100

On analyzing the performance of four prominent language models—BERT, DistilBERT, RoBERTa, and DeBERTa—in the classification of phishing versus legitimate URLs. The AUC scores for all models are consistently high, at 0.99, indicating that each model possesses an almost perfect ability to discriminate between phishing and legitimate URLs. This high AUC value underscores the effectiveness of these models in ranking URLs by their likelihood of being phishing, demonstrating robust overall performance. The F1-scores of the models are exceptionally high, with DistilBERT achieving the highest F1-score of 99.992%, closely followed by BERT and DeBERTa at 99.990%, and RoBERTa at 99.987%. In terms of accuracy, all models perform exceedingly well, with DistilBERT recording the highest accuracy of 99.991%, while BERT and DeBERTa both achieve 99.988%, and RoBERTa slightly lower at 99.985%. Precision metrics are very similar across the models, with DistilBERT again leading with a precision of 99.985%, followed by BERT and DeBERTa at 99.980%, and RoBERTa at 99.975%. Sensitivity, or recall, is perfect across all models, each achieving a score of 100%. This indicates that the models are adept at identifying all phishing URLs, leaving no phishing instances undetected, which is crucial for effective phishing detection. Overall, while all models demonstrate exceptional performance, DistilBERT stands out

due to its highest F1-score, accuracy, and precision, suggesting it may be the most effective choice in practical applications. DistilBERT's efficiency in terms of size and computational resources, combined with its high performance, makes it particularly suitable for deployment in real-world scenarios. However, BERT and DeBERTa, with their similar high-performance metrics, remain strong contenders, especially in contexts where computational resources are less constrained. RoBERTa, despite slightly lower precision, still offers high performance and could be considered depending on specific application needs. The external dataset serves as a rigorous benchmark for evaluating the robustness and generalization capabilities of the models, with over 450,176 URLs. Unlike the primary dataset, which includes a larger set of features, the external dataset contains only URLs and their respective labels, along with two additional features extracted directly from the URLs: the length of the URL and HTTPS protection status. This minimal feature set poses a significant challenge, requiring the models to rely heavily on these limited attributes, but helping in rigorously testing the models over such less information on a large URL dataset, providing more nuanced insights. As shown in Table 6, the performance metrics across various models reflect their ability to navigate these constraints.

Table 6
LLMs results on the external test dataset

Model Name	AUC	F1-score (%)	Precision (%)	Sensitivity (%)
BERT	0.66	1.14	91.68	0.57
DistilBERT	0.65	42.44	99.51	26.97
RoBERTa	0.60	68.03	92.81	53.70
DeBERTa	0.66	2.78	84.75	1.41

BERT and DeBERTa both achieve an AUC of 0.66 but show poor recall, with BERT having a very low F1-score of 1.14% and DeBERTa at 2.78%. Despite high precision (91.68% for BERT and 84.75% for DeBERTa), both models miss most true positives. DistilBERT performs slightly better with an F1-score of 42.44%, combining high precision (99.51%) with a modest sensitivity (26.97%). RoBERTa, with an AUC of 0.60, stands out for its higher F1-score of 68.03%, demonstrating a more balanced performance between precision (92.81%) and sensitivity (53.70%). RoBERTa is notably better at identifying positive instances compared to the others, albeit with a slight trade-off in precision. Despite a lower AUC, RoBERTa's relatively higher recall suggests a less conservative approach, making it the most robust model in this limited feature scenario. Figure 4 presents a bar plot comparing the F1-score metrics across the models over the external test dataset. Overall, the models demonstrate a significant trade-off between precision and recall, with each model's performance shaped by the stringent conditions of the external dataset. The external dataset's role as a critical stress test is evident in the stark differences in performance metrics compared to results over the primary dataset, though the major reason for the contrast in metrics is due to external dataset having very less features. Overall, LLMs demonstrated strong performance on a

dataset containing only URLs, despite being initially trained on a feature-rich primary dataset. This performance was notably better compared to traditional machine learning algorithms under similar conditions [22]. The findings highlight the importance of developing models that can maintain robust performance even when operating under severe feature constraints.

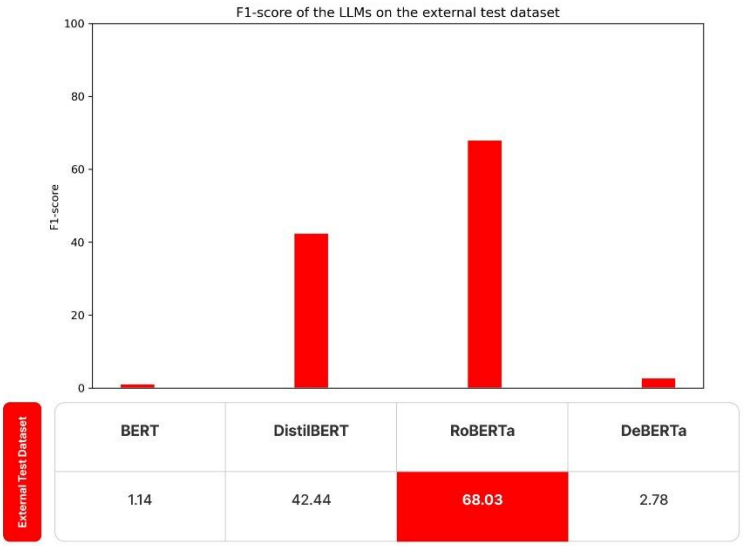

Figure 4
F1-score metrics across the models over the external test dataset

Conclusion

This research demonstrates that Large Language Models (LLMs)—BERT, DistilBERT, RoBERTa, and DeBERTa—are highly effective for classifying URLs as either legitimate or phishing, particularly when provided with a feature-rich dataset. The consistent high performance across all models on the primary dataset, with AUC scores of 0.99 and exceptional F1-scores, precision, and accuracy metrics, underscores the models' capability to accurately discern between phishing and legitimate URLs. DistilBERT, in particular, stands out for its remarkable efficiency and effectiveness, achieving the highest F1-score, accuracy, and precision, making it an ideal choice for real-world applications where both performance and computational efficiency are critical. BERT and DeBERTa also show strong performance, though with slightly less efficiency compared to DistilBERT. RoBERTa, while slightly lower in precision, demonstrates competitive performance and could be a viable option depending on specific application requirements. However, the performance metrics shift notably when evaluated using the external dataset, which is constrained by a minimal set of features. The stark decline in performance, especially in recall and F1-scores for BERT and DeBERTa, highlights the challenge of maintaining

classification efficacy under severe feature limitations. DistilBERT's moderate performance and RoBERTa's superior recall under these conditions point to the importance of balancing precision and recall in feature-scarce environments. The findings emphasize the need for continued development of models that not only excel with rich datasets but also exhibit robustness and adaptability when faced with limited features. This research contributes valuable insights into the trade-offs between precision and recall and highlights the necessity of optimizing models to perform reliably across a spectrum of real-world scenarios. Future work should focus on enhancing model resilience to feature constraints and exploring methods to further improve performance under such conditions.

References

[1] P. Yang, G. Zhao, and P. Zeng, "Phishing website detection based on multidimensional features driven by deep learning," *IEEE Access*, vol. 7, pp. 15196–15209 (2019).

[2] P. Yi, Y. Guan, F. Zou, Y. Yao, W. Wang, and T. Zhu, "Web phishing detection using a deep learning framework," *Wireless Communications and Mobile Computing*, vol. 2018, Article ID 4678746 (2018).

[3] M. A. A. Siddiq, M. Arifuzzaman, and M. S. Islam, "Phishing website detection using deep learning," in Proceedings of the 2nd International Conference on Computing Advancements (ICCA), pp. 83–88 (2022).

[4] F. Greco, G. Desolda, A. Esposito, and A. Carelli, "David versus Goliath: Can machine learning detect LLM-generated text? A case study in the detection of phishing emails," in Proceedings of The Italian Conference on CyberSecurity (ITASEC) (2024).

[5] F. Trad and A. Chehab, "Prompt engineering or fine-tuning? A case study on phishing detection with large language models," *Machine Learning and Knowledge Extraction*, vol. 6, no. 1, pp. 367–384 (2024).

[6] M. Bethany, A. Galiopoulos, E. Bethany, M. B. Karkevandi, N. Vishwamitra, and P. Najafirad, "Large language model lateral spear phishing: A comparative study in large-scale organizational settings," arXiv preprint arXiv:2401.09727 (2024).

[7] A. Prasad and S. Chandra, "PhiUSIIL Phishing URL (Website)," *UCI Machine Learning Repository* (2024). [Online]. Available: https://doi.org/10.1016/j.cose.2023.103545

[8] M. M. Uddin, N. Jahan, F. M. Shahin, M. R. Hashem, and M. S. Kaiser, "A comparative analysis of machine learning-based website phishing

detection using URL information," in Proceedings of the 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), pp. 69–74 (2022).

[9] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805 (2018).

[10] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108 (2019).

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692 (2019).

[12] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," arXiv preprint arXiv:2006.03654 (2020).

[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and A. M. Rush, "HuggingFace's transformers: State-of-the-art natural language processing," arXiv preprint arXiv:1910.03771 (2019).

[14] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in Adam," arXiv preprint arXiv:1711.05101 (2017).

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980 (2014).

[16] D. Goyal, F. Sheth, P. Mathur, and A. K. Gupta, "Discrete mathematical models for enhancing cybersecurity: A mathematical and statistical analysis of machine learning approaches in phishing attack detection," *Journal of Discrete Mathematical Sciences and Cryptography,* vol. 27, no. 2-B, pp. 569–599 (2024).

[17] P. Mathur, F. Sheth, D. Goyal, and A. K. Gupta, "Deep insight: Mathematical modeling and statistical analysis for mango leaf disease classification using advanced deep learning models," *Journal of Interdisciplinary Mathematics,* vol. 27, no. 2, pp. 317–342 (2024).

[18] A. Kumar, A. K. Gupta, D. Panwar, S. Chaurasia, and D. Goyal, "Operating system security with discrete mathematical structure for secure round robin scheduling method with intelligent time quantum," (2023)

[19] J. K. S. Kaitholikkal and B. Arthi, "Phishing URL dataset," Mendeley Data, V1 (2024). [Online]. Available: https://doi.org/10.17632/vfsz-bj9b36.1

[20] T. Teubner, C. M. Flath, C. Weinhardt, W. van der Aalst, and O. Hinz, "Welcome to the era of ChatGPT et al.: The prospects of large language models," *Business and Information Systems Engineering*, vol. 65, no. 2, pp. 95–101 (2023).

[21] P. Wang, L. Li, L. Chen, F. Song, B. Lin, Y. Cao, J. Li, Q. He, J. Wang, and Z. Sui, "Making large language models better reasoners with alignment," arXiv preprint arXiv:2309.02144 (2023).

[22] C. Song, T. Ristenpart, and V. Shmatikov, "Machine learning models that remember too much," in Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 587–601 (2017).

Received November, 2024