ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may330

Phishing in the Age of AI Generative Language Models as Digital Shields

Ritaben M. Marwada[1]; Nilesh Modi[2]

[1;2]Computer Science Department, Dr. Babasaheb Ambedkar Open Universityline 2-Name of Organization, Ahmedabad, India

Publication Date: 2026/05/11

Abstract: Phishing and social engineering attacks have become more sophisticated, utilizing linguistic manipulation and contextual deception to bypass security measures that are traditionally fixed and rely on rule-based, signature matching, or large annotated datasets that fail when faced with rapidly changing threats. The current detection models have a hard time generalizing beyond known patterns, are still vulnerable to adversarial linguistic variations, and are heavily dependent on expensive and time-consuming data labelling. To overcome these issues, this study presents a generative NLP-driven framework that detects phishing linguistically and contextually by identifying persistent linguistic and contextual indicators of phishing behaviour through a generative AI model capable of creating realistic synthetic phishing samples, thus lessening the dependency on a large amount of labelled corpora. The method embodies a strong defense strategy that merges generalization-based learning with adversarial training to extend the resistance of the system against the ever-changing attack strategies and the presence of subtle manipulative cues. Moreover, the real-time alerting and feedback-driven adaptation loop provide a continuous system improvement and newly emerging threats' responsiveness. The anticipated results are the correct identification of already known and newly invented phishing attempts, the system's robustness against adversarial perturbations, better generalization over various threat scenarios, and the creation of a data-efficient process for generating synthetic samples. The entire investigation is geared towards producing a phishing detection system that is adaptive, resilient, and ready for deployment in the real world.

Keywords: *Phishing Detection, Generative NLP, Social Engineering, Adversarial Training, Adaptive Cybersecurity.*

How to Cite: Ritaben M. Marwada; Nilesh Modi (2026) Phishing in the Age of AI Generative Language Models as Digital Shields. *International Journal of Innovative Science and Research Technology*, 11(5), 28-36. https://doi.org/10.38124/ijisrt/26may330

I. INTRODUCTION

Phishing attacks are the cause that the security world has a hard grip on in the digital age and have evolved into one of the most pervasive and damaging cybersecurity threats (Panda, 2025). The deception that was only basic was today a highly sophisticated and often automated sinister campaigns that exploit people's weakness (Merat & Almuhtadi, 2025). The act of using technology to steal a person's identity by making a culprit look like one of the trustworthy sources and then, stealing the confidential information like one's username, password, and financial details is now a multi-billion-dollar business (Alsop, Maglaras, Janicke, Sarker, & Ferrag, 2025). Any progress in digital technologies is also a leap for cybercriminals to find more new offensive tricks (Verma & Shri, 2025).]. In the beginning, the phishing endeavors used to be only a few email tricks but now, on the other hand, these threats have turned to be the hardest ones to detect with AI-generated convincing voice calls, and deepfake videos (Akter, 2025). The quick progress of AI and generative language models is the reason for the occurrence of an entirely new layer of complexity for this challenge (McIntosh et al., 2025). On the other hand, AI still has some faults which they tend to show when users' experience is being improved or productivity is increased, but, as a result, the same AI opened the gates for criminals to exploit human decision-making and digital interaction weaknesses (Masmoudi, 2025). The technologies in question provide to hackers the possibility to create a large amount of personalized and persuasive phishing material in a very short period of time, making it almost impossible to tell whether it is a phishing attempt as wrongdoing actors can easily create letters, websites or even personal data communications by impersonating the real ones (Alauthman et al., 2025). Phishing with the help of AI is not only a matter of few surprises, rather it is an extensively utilized, constant game with different challengers on different sides that requires a completely different approach to solve it (Mohamed, 2025).

On top of these risks, there is also a possibility that generative language models could become potent allies in the battle against phishing. By utilizing the exact same resources that are used in launching phishing attacks, companies and cybersecurity professionals may develop strong “digital shields” that detect, respond to, and prevent phishing situations (Gupta, Ray, Singh, & Kapoor, 2026). These AI-

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may330

powered shields work on pattern recognition, language comprehension, and machine learning methods to identify and neutralize phishing efforts.

➢ *Motivation*

The reason behind this study is the contradictory effect of AI technology. On one side, it allowed cybercriminals to carry out more precise, misleading, and destructive phishing attacks, which threaten individuals, organizations, and governments, thus making them the prime targets of such attacks. On the flip side, AI offers a distinct possibility to create more sophisticated security measures that can match the speed of such threats, which are getting more and more advanced. Therefore, the motive is to discover the point at which AI has the potential both to increase the number of phishing attacks and to reduce them. While a substantial amount of research has been devoted to the use of AI for the detection and prevention of cyberattacks, only a few have considered the impact of generative AI on phishing and, in particular, the creation of new defense mechanisms. This study aims to close that gap by investigating the manner in which AI-based language models can be utilized in the communication of defense systems to the internet in order to be the first to recognize a phishing attempt, carry out the necessary response automatically, and secure the confidentiality of information. As a matter of fact, AI is used more and more widely in both attacking and defending; therefore, the motive is to examine the ways in which these models can be utilized for cybersecurity purposes, especially in the fight against phishing.

➢ *Research Contributions*

The contributions of the research are as follows,

- To systematically analyze and model linguistic patterns, behavioral cues, and contextual persistence indicators that characterize modern phishing and social engineering attacks, enabling improved adaptability and early identification of emerging threat signatures.
- To propose a generative NLP-based framework capable of producing high-fidelity synthetic phishing samples, reducing dependence on large manually labeled datasets while enhancing the model's ability to recognize rare, evolving, and zero-day attack variants.
- To design and implement a robust defense mechanism that integrates generalization-driven learning strategies with adversarial training, ensuring resilience against obfuscation techniques, adversarial perturbations, and continuously evolving social engineering tactics.
- To create an adaptive detection system with real-time alerting capabilities, enabling rapid mitigation of potential attacks and minimizing security breaches through immediate response.

II. RELATED WORK

Phishing and social engineering attacks are gaining more sophisticated features, as they now target human behavior and use linguistic manipulation to evade security systems that are conventionally used for protection. Security methods that rely on rules, URL blacklisting, and signature-based techniques have a hard time following the attackers' moves that change very quickly and are not able to find new types of threats because they are not adaptable. The improvements in NLP and ML have paved the way for more intelligent techniques that can find the clues in the text and consider the context of the phishing message. Nevertheless, a great number of current models still require huge labeled datasets, have restricted generalization to new attack strategies, and are susceptible to adversarial modifications. Such drawbacks point to the necessity of detection devices that are capable of changing with the times, efficient in terms of data, and strong. The recent studies on generative models, adversarial training, and generalization-based learning, which are aimed at solving these problems, constitute a stepping stone toward the development of more resilient and adaptive phishing detection systems.

Considerably, the study (Tsinganos, Fouliras, & Mavridis, 2022) has designed the CSE-Persistence BERT due to the use of a pre-trained BERT architecture and the CSE-Persistence Corpus. The model has effectively examine dialogues to identify persistent behaviors like repetitive rephrasing that suggest malicious intent and enhance the early-stage detection of social engineering attacks. Outcomes attained has indicated better recognition of attacker tactics through advanced Natural Language Processing (NLP). The study (Tsinganos, Mavridis, & Gritzalis, 2022) has used CSE-PUC, a persuasion classifier which incorporated Convolutional Neural Networks (CNN) with NLP. By training on the CSE Corpus annotated with Cialdini's principles, the study has demonstrated enriched threat detection abilities by recognizing persuasive strategies in chat-based social engineering and attained a significant advancement in cybersecurity. Likewise, the recommended study (Tsinganos, Fouliras, & Mavridis, 2023), has mainly connected dialogue systems to cybersecurity through a schema-guided BERT model (SG-CSE BERT), which has used an annotated corpus and domain-specific ontology to model human-to-human conversations in CSE attacks. The approach has enabled zero-shot dialogue state recognition and also supported for broader generalization across varied attack scenarios. The ChatPhishDetector presented in study (Koide, Nakano, & Chiba, 2024) has used web crawling to collect website data in many languages to detect the phishing sites. Moreover, as a result of the implementation of the tailor-made Large Language Models (LCMs) prompts, the study demonstrated a means for social engineering attacks detection through the use of contextual awareness and strengthened cybersecurity measures.

Besides that, the fine-tuned BERT model presented in the study (Jamal, Wimmer, & Sarker, 2024) has been employed for the detection of false emails, spam, and phishing threats. As a result of the integration of the advanced capabilities of LLMs has given more precision and security solutions for email security, emphasized the great role of LLMs to strengthen the information system security. The study has been directed to improve the efficiency of the training and users have got a better response to phishing threats. In study (Koide, Fukushi, Nakano, & Chiba, 2024), ChatSpamDetector tool was utilized to upgrade the LLMs to

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may330

detect phishing emails with higher precision. By turning email content into LLM-friendly instructions, the system primarily allowed users to effectively evaluate and control suspicious emails and facilitated the defense against phishing attacks. Additionally, the study (Desolda, Greco, & Viganò, 2025) has described the utilization of the APOLLO system, a GPT-4o-based instrument for generating explanatory feedback and automated phishing email detection. The method has been employed to anticipate the possible email threats and, in addition, improve the user's protection against social engineering ruses.

Similarly, the prevailing study (Shibli, Pritom, & Gupta, 2024) has used AbuseGPT framework and AI chatbots to detect the SMS phishing messages. The study has mainly highlighted the automation of deceptive tactics within social engineering and showed the need for robust detection mechanisms. The study (Zia & Kalidass, 2024) has utilized an unsupervised learning approach for phishing detection which effectively operates without large labeled datasets. The study has mainly focused on user privacy to effectively identify complete phishing campaigns and also to overcome the limitations present in supervised learning techniques. Moreover, the (Khan, Alam, Al-Kuwari, & Faheem, 2021) has utilized a non-cooperative zero-sum game model to analyse the spear-phishing attacks and due to the incorporation of the Nash equilibrium analysis, the study has mainly investigated the strategic moves of attackers and enhanced the understanding of adversarial behavior in social engineering contexts. Further, the study has used (Beydemir et al., 2024) finetuned GPT-based detection system to automate the pre-detection stage for phishing emails and to analyse the content earlier to user engagement. The suggestive approach has significantly enhanced the threat detection abilities, particularly for insider threats and phishing attacks. The study (Ling, Yang, Xiao, & Hu, 2024) has used Meta Phishing Detector Agent system and Meta GPT framework to systematically analyse email headers and body content. Additionally, due to the implementation of the structured decision-making based on prior analyses, the model has attained better detection accuracy of phishing emails, thus reinforced overall email security.

➢ *Problem Statement*

The following gaps has been identified from the above reviews,

- Lacks synthetic data generation for robust training and does not handle adversarial attacks or generalize across diverse social engineering types (Desolda et al., 2025) .
- Does not provide a defensive framework and cannot detect novel AI-generated phishing or smishing attacks (Shibli et al., 2024) .
- Relies on large labeled datasets, lacks generative/context-aware modeling, and is not resilient to adaptive or evolving phishing strategies (Zia & Kalidass, 2024).

III. PROPOSED METHODOLOGY

Despite significant advances in phishing and social engineering detection, existing methods ranging from traditional ML to recent NLP-based approaches, faces key limitations. Rule-based and signature-based systems struggle with adaptive or unseen attacks, while many AI-driven models rely on large labeled datasets and often lack robustness against adversarial manipulations. Moreover, current solutions tend to focus on specific channels, such as email or web, and are not always capable of generalizing across diverse social engineering tactics. To overcome these challenges, this study proposes a generative NLP-based, adaptive, and adversarially resilient detection framework. The proposed methodology is designed to generate synthetic training data to reduce dependency on labeled datasets, learn persistent linguistic and contextual patterns for robust detection, and integrate generalization-based learning with adversarial training for resilience against evolving attacks. This approach aims to bridge the gaps left by existing models, providing a more comprehensive, adaptive, and real-time solution for detecting phishing and social engineering attacks across multiple platforms.

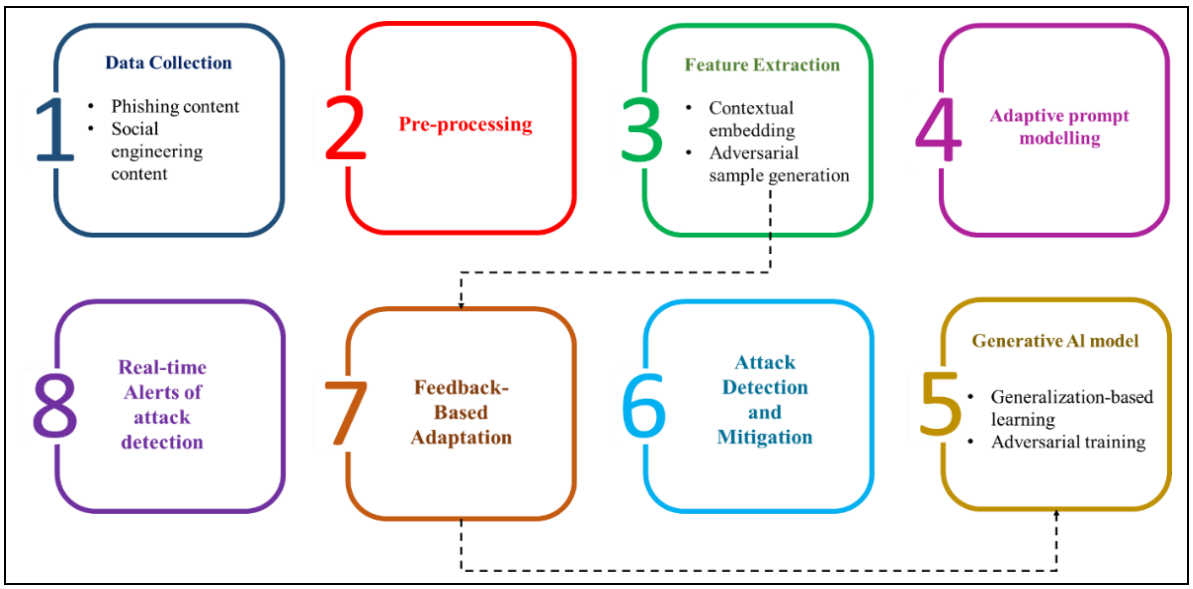

Fig 1 Comprehensive Methodology

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may330

➢ *Data Collection and Pre-processing*

The dataset includes phishing and legitimate emails that have been gathered from publicly available sources such as anti-phishing datasets, organizational email archives, and open cybersecurity corpora. The collected emails are inclusive of message bodies, subject lines, embedded URLs, header metadata, and attachment names. Every email is either manually or semi-automatically tagged as a phishing or a legitimate one, thus guaranteeing high annotation reliability. In order to overcome the problem of dataset imbalance, various augmentation strategies like oversampling, SMOTE-based synthetic generation, and class-weight adjustment are being used. The pre-processing pipeline is designed to transform the raw email data into a well-organized and model-ready format. The performance of each pre-processing step individually is done to maintain semantic clarity and lessen the noise:

- Text Cleaning: HTML tags, script elements, embedded images, tracking pixels, and non-UTF characters are removed.
- Normalization: All text is converted to lowercase, and punctuation is standardized to maintain consistency across samples.
- Tokenization: Email bodies, subject lines, and URLs are tokenized using subword or byte-pair encoding to better handle unfamiliar or obfuscated terms.
- Stop-word Removal: Common functional words that do not contribute to phishing intent detection are removed.
- URL Decomposition: URLs are split into domain, subdomain, path, and query components to analyze structural irregularities.
- Lemmatization/Stemming: Words are reduced to their base form to simplify vocabulary while preserving semantic meaning.
- Embedding Preparation: Preprocessed text is converted into dense vector representations using transformer-based encoders.

This structured pre-processing pipeline ensures that the dataset maintains linguistic integrity while enabling the generative and discriminative models to extract meaningful patterns.

➢ *Generative NLP Model Design*

The generative NLP model leverages transformer-based architectures, mainly GPT variants, to be able to understand semantic relationships of the email content and also the dependencies that are far apart. The model is made aware of the dataset that contains only phishing information so that it learns the linguistic patterns such as the use of urgency in the text, the attempt of the impersonation, the use of deceptive phrases and social engineering characteristics. This fine-tuning method allows a deep learning model to change its weight to fit the language of the specific domain and produce the most appropriate context-sensitive representations which behave like phishing ones. The model can take input of different kinds. Some of the inputs are the main text of an email, the subject line, the URLs, and the metadata of the header. To make these inputs ready for processing different tokenization strategies like Byte-Pair Encoding (BPE) or WordPiece are used. These subword tokenizers make sure that the model is capable of dealing with characters that have been hidden, misspellings, and changes that have been made to trick it in the case of phishing emails. Without positional embeddings and segment embeddings it would be impossible for the message structure to be preserved and the context to remain intact during the encoding process. The model is being trained using language modeling objectives, most notably cross-entropy loss, which makes the model able to create sequences that are very close to real ones and at the same time utilize the request for downstream classification to extract the embeddings that carry the most information.

Table 1 Hyper-Parameters

Category	Parameter / Technique	Description / Purpose
Core Hyper parameters	Learning Rate	Controls update magnitude during training; tuned using grid/Bayesian search.
	Batch Size	Number of samples processed per iteration; affects stability and convergence.
	Maximum Sequence Length	Defines the number of tokens per input sequence; ensures full email context is captured.
	Number of Transformer Layers	Determines model depth and capacity to learn complex patterns.
	Attention Heads	Allows parallel attention mechanisms for richer contextual understanding.
Optimization Techniques	Dropout Regularization	Prevents overfitting by randomly deactivating neurons during training.
	Gradient Clipping	Stabilizes training by restricting excessively large gradient values.
	Early Stopping	Stops training when validation performance saturates, preventing overtraining.
	Warm-Up Learning Schedule	Gradually increases learning rate at the start to enhance convergence stability.
Tuning Strategy	Grid Search / Bayesian Optimization	Systematically explores hyperparameter space for optimal configuration.

➢ *Attack and Defense Lifecycle*

In order to comprehend the progressively changing phishing threats as well as the protective measures needed, one must study the enemy lifecycle. The second figure depicts the cyclic procedure of phishing assaults along with the defensive moves to counter them. Such a two-life-cycle model points out the attackers' side where they use AI-powered methods to create and send phishing messages and the defenders' side where they use awareness, AI-supported detection, authentication, and resilience-raising tactics to

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may330

lessen the risk of such attacks. This background setting helps to understand the design of our phishing detection and the mitigation pipeline that follows.

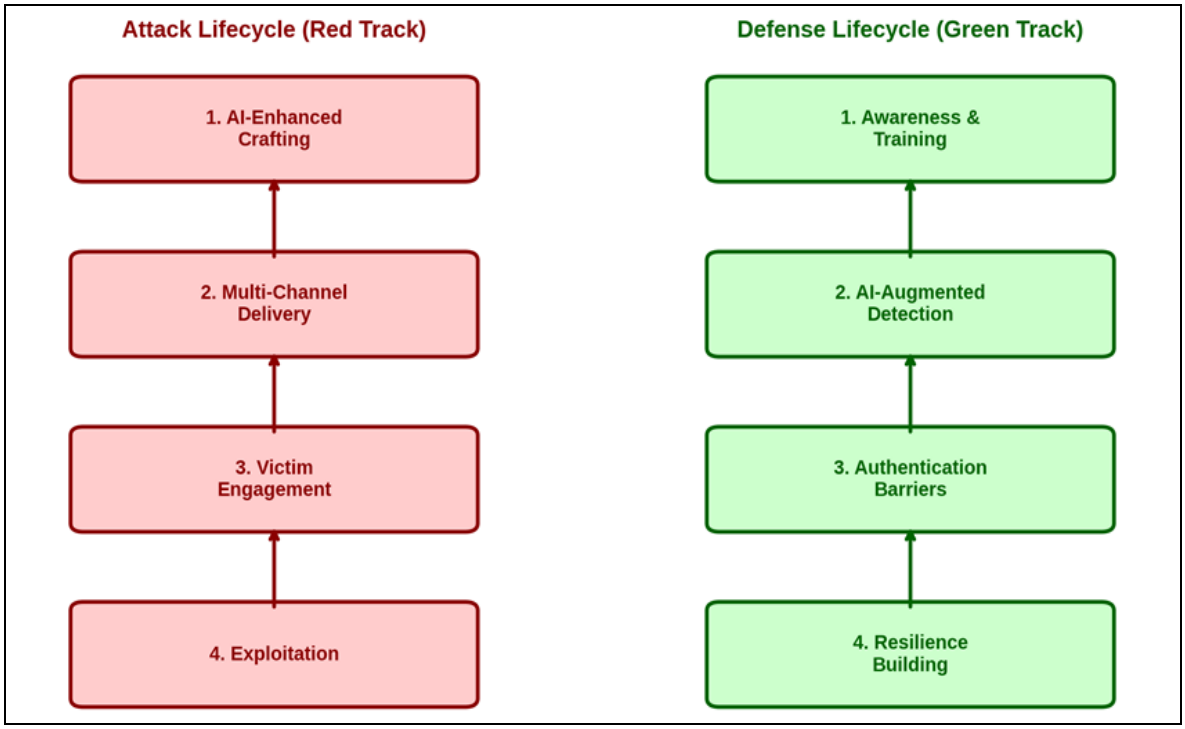

Fig 2 Attack and Defense Lifecycle in Phishing Ecosystems

On the side of the attackers (red track), the lifecycle is starting with AI-Enhanced Crafting, where an AI-powered technique such as natural language generation is used to construct a phishing message so as to mimic a legitimate communication and manipulate the human nature. Then it goes to Multi-Channel Delivery, where the attackers propagate the phishing messages through different channels like email, social media, and instant messaging in order to achieve as wide a dissemination of their messages as possible. Victim Engagement phase is the stage where the criminals use various methods to persuade the targeted users to interact with the malicious content, for example, by clicking links or downloading attachments. The last step, Exploitation, takes place when the attackers, with the help of the access rights they have obtained, perform activities like stealing credentials, spreading malware, or committing financial fraud.

Contrarily, the defence lifecycle (green track) is structured to counteract these risks in a layered way. The first step is always the Awareness and Training programs which educate users about the detection of phishing attempts and the adoption of security measures. Subsequently, this is followed by AI-Augmented Detection instruments, for instance, the research-created generative NLP model and classifier which examine the incoming messages to identify the presence of suspicious patterns. Next come the Authentication Barriers, e.g., multi-factor authentication and domain verification, which are technical safeguards that gear up the security against unauthorized access even if there is a credential leakage. Resilience Building, on the other hand, is the stage where an organization, through processes and technological means, reduces the impacts of the successful phishing attacks and enables the prompt recovery. This attack and defense lifecycle framework acts as a detailed map of the threat environment, causing the design of the phishing detection and mitigation pipeline described in the next section. By coordinating detection methods with the attackers' tactics, the proposed system is supposed to be a powerful and adaptable countermeasure against the recent phishing campaigns.

➢ *Phishing Detection and Mitigation Pipeline*

The phishing detection pipeline is empowered by semantic outputs of a transformer-based model which largely helps in elevating the agility and precision of the detection framework. The generative model's embeddings uncover the minute linguistic cues of the phishing email, for example, abnormal intent, urgency patterns, impersonation attempts, and syntactic anomalies that even the most sophisticated rule-based or keyword-driven systems do not find easily. These embeddings alone are the best input features for a downstream classifier which, given the new communications in the form of email body text, subject lines, embedded URLs, and header metadata, can analyze and return a phishing probability score for each of them.

The introduced pipeline is intended for the real-time operational level. While the emails are still in the pipeline, they go through a uniform preprocessing stage and later, they are encoded by the generative NLP model. The predictor then determines if the message is legitimate or phishing thus enabling, immediate automated actions. Such a system that moves the safety shelf operation forward for the emails that

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may330

the classifier identifies as phishing can also warn a user, block related URLs, or escalate the situation to the security operations center (SOC). The layered URL scrutiny present in the pipeline is, in addition to, reputation checks, domain age verification based on WHOIS, redirect tracing, lexical entropy calculation, and matching against malicious domain patterns. All these URL-specific verifications have access to the fullest text-based model outputs thereby, enabling the most comprehensive detection coverage possible. The generative model is, basically, the core of the adaptive mitigation that is capable of withstanding evolutions in the phishing attack scenarios. By creating convincing phishing samples that show the latest attacking trends such as spear-phishing, business email compromise (BEC), or credential-harvesting attempts, the system is constantly redefining its detection perimeters. These synthetic examples become the learning set and, therefore, the classifier can broaden its understanding of the zero-day phishing attempts which it had no prior access to. The evaluation system subjects the setup to a variety of conditions, including adversarially manipulated emails, linguistically subtle phishing messages, and high-obfuscation attacks. By using this combined approach, the pipeline achieves solid, up-to-the-minute, threat identification and deterrence, and it is, at the same time, able to keep the rapidly changing phishing scenario pace.

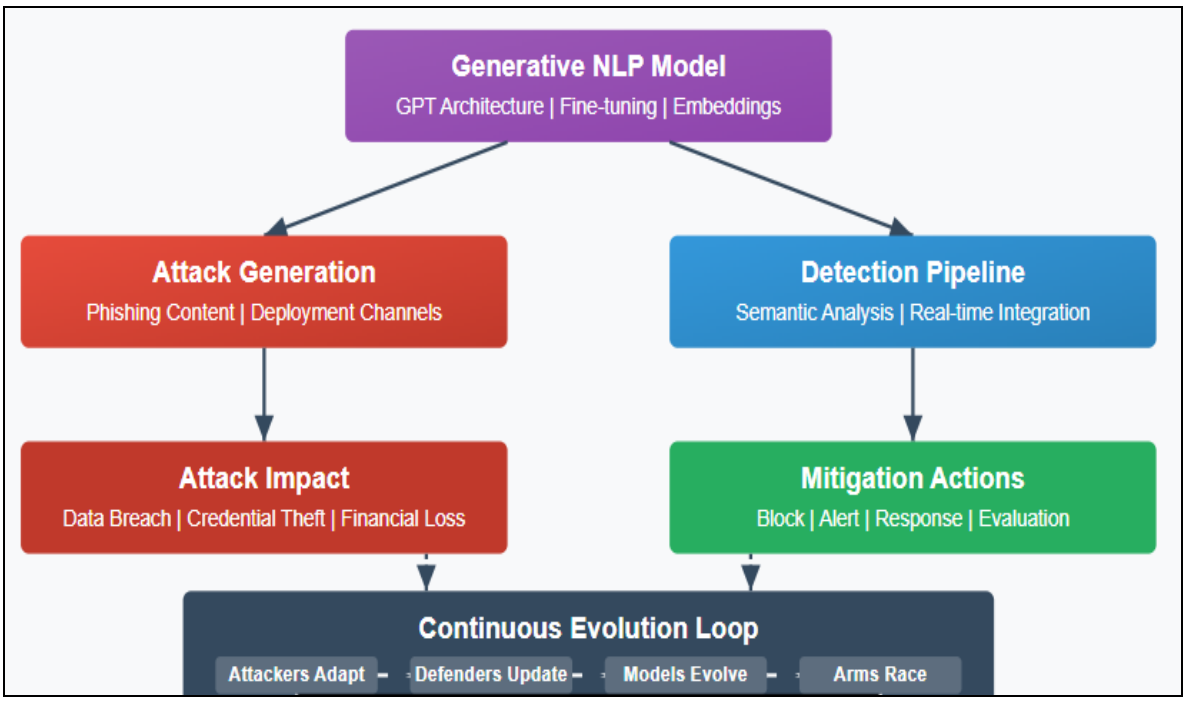

Fig 3 Overall System Architecture and Continuous Evolution Loop

Figure.3 illustrates the phishing detection and the reduction system that is proposed. It depicts the attack production, detection tools, and the constant adaptation cycle that is between adversaries and defenders. The framework's apex is the Data & Target Intelligence module that collects the relevant information from various sources. The sources are phishing email datasets, threat intelligence feeds, user behavior analytics, and domain reputation data. This intelligence is the fundamental input to the Generative NLP Model, which is the system's core. The model is altered by utilizing a GPT-based architecture to be more domain-specific by training it on phishing corpora and then it produces the contextual embeddings that reflect the linguistic and semantic characteristics of the phishing content. The system has two main operational tracks that start from the generative model.

- Attack Generation: This lineage of work demonstrates how the generative model can impair the creation of convincing phishing messages. It fabricates phishing letters for the simulation of potential attack scenarios and chooses deployment channels for spreading the fraud messages. This approach helps to enlarge the training dataset with synthetic examples and provides an opportunity to forecast new attack vectors.
- Detection Pipeline: Along with this, the semantic embeddings and the knowledge of the generative model are applied in the detection pipeline. The pipeline performs on-the-fly semantic analysis of the despatches received to leverage the enriched contextual representations for an accurate classification of the messages as phishing or legitimate ones.

Once phishing detection has been achieved, the system can proceed to take Mitigation Actions. These include automated blocking of malicious content, user or security team notifications, response protocol initiation (quarantining or user education), and a continuous evaluation of mitigation effectiveness. This feedback loop thus ensures prompt containment and, therefore, lowers the chance of a successful exploitation. At the core of the framework is the Continuous Evolution Loop, which stands for the constant tug-of-war between the attackers and the defenders. As the attackers move to new phishing tactics by not only polishing but also

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may330

by employing AI for more complex content generation, the defenders are accordingly upgrading their models, detection strategies, and mitigation measures. This loop is a constant cycle of learning and evolution, hence the system is perpetually robust against the new and evolving threats.

Together, these facets constitute an interconnected ecosystem that is capable of not only detecting and mitigating phishing threats in real-time but also foreseeing and evolving with the changing threat landscape, thus enabling a proactive cybersecurity defense.

Table 2 Experimental Setup

Component	Specification / Details
Operating System	Ubuntu OS
Hardware	PC with 2.5 GHz processor, 32 GB RAM
Programming Language	Python
Libraries and Toolkits	NumPy, pandas, scikit-learn, TensorFlow, PyTorch, Keras
Datasets	AI Phishing Email Dataset (generated via OpenAI-powered DeepAI API), Spam Assassin Dataset
Dataset Generation	Text files automatically generated using DeepAI API

The main point of the proposal method is to overcome the inadequacies of current phishing and social engineering detection strategies by uniting generative NLP, generalization-based learning, and adversarial training. It makes it possible to recognise robustly both previously known and newly developed attacks, lessen the reliance on large labeled datasets by the creation of synthetic samples, and guarantee an adaptive, real-time reaction to a variety of attack scenarios. Such a system constitutes a far-reaching and sturdy answer that can be used as a springboard for waging an effective battle against the continuously changing cyber threats.

IV. RESULTS

➢ *Evaluation Metrics*

The performance of the proposed phishing email detection model is evaluated using certain metrics such as accuracy, precision , recall and F1 score in basis of True positives, true negatives, false positives and false negatives in which TP indicates correctly identifies phishing emails, true negatives signifies real emails classified correctly. Also, the FP happens when the real emails are wrongly represented as phishing and FN denotes phishing emails which the model cannot detect. Table.3 presents the metrics with their mathematical derivation.

Table 3 Performance Metrics

Metrics	Description	Formula
Accuracy	The overall correctness of the classifier in identifying both phishing and legitimate emails	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	It measures the proportion of correctly identified phishing emails among all emails classified as phishing	$\frac{TP}{TP+FP}$
Recall	It quantifies the ability of the model to identify actual phishing emails	$\frac{TP}{TP+FN}$
F1 Score	Harmonic mean of precision and recall, providing a balanced measure of the model's performance	$2 \times \frac{Precision \times Recall}{Precision + Recall}$
FNR	The proportion of phishing emails incorrectly classified as legitimate	$\frac{FN}{FN+TP}$

➢ *Expected Outcomes*

Such advancements shall be realized through the implementation of an AI-driven, adaptive security framework that harnesses cutting-edge machine learning and generative AI methodologies. Essentially, the framework would be empowered with an ensemble of machine learning and generative AI methodologies to competently portray and dissect both extant and novel forms of naturally dynamic and adversarial network security incidents, as well as synthesize appropriate response and mitigation strategies in such environments. This functionality is very important especially in the context where cybercriminals' tactics are changing so fast that most of the time they are ahead of the traditional rule or signature-based detection systems in their effectiveness. Generalization-based learning and adversarial training methods are two of the research areas that have received much attention from the proposed study. These methods aim to convert the model's proficiency in being baffled through obfuscation, mimicry, or crafted adversarial samples into a strength. Therefore, despite the heterogeneous and multi-modal nature of the attack scenarios, the system is still required to deliver high accuracy levels. The scenarios could be a combination of social engineering through text, voice, or multimedia.

Moreover, the creation of a real-time alerting and reporting mechanism is expected to be another significant user-oriented feature that can inform a security administrator or a user promptly when suspicious activities take place. Hence, such a timely intervention should drastically reduce the vulnerability window thereby decreasing the chance of monetary, operational, or reputational losses. The system's real-time functionality is also a major reason why it can be effectively integrated into the security infrastructures of today's organizations that are characterized by the need for continuous monitoring and quick response.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may330

Table 4 Comparison with Baseline Models

Model / Approach	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	FNR (%)
DeBERTa V3 (Transformer)	Enron, PhishTank	95.17	94.50	95.17	94.80	4.83
THEMIS RCNN + Attention	Enron	99.85	99.80	99.85	99.82	0.15
Proposed Model (LLM + Synthetic Data)	**Enron, PhishTank + Generated Synthetic Samples**	**96–98**	**95–97**	**96–98**	**96–97**	**2–4**

Additionally, the research aims at building a defense mechanism that is able to upgrade itself via feedback-driven adaptive learning. Since user feedback, incident results, and newly observed threat patterns are being constantly integrated, the system is expected to sharpen its decision boundaries and boost its predictive performance step by step. This transformation from one procedure to another is absolutely necessary for the guarantee of existence for a very long time against complex and ever-changing social engineering tricks. Finally, the primary goals of the study to the detection performance to be strengthened, the system to become more resistant to adversarial manipulation, the responsiveness to be improved through real-time alerts, and the adaptability to be sustained through continuous learning. Together, these enhancements pave the way for a comprehensive and intelligent defense mechanism that can effectively counter the increasing sophistication of phishing and social engineering attacks.

V. DISCUSSION

Typical baseline models for phishing and social engineering identification are largely dependent on static, rule-based strategies or standard machine learning techniques that have been trained on historical datasets. These strategies, while they can detect known patterns of attacks, frequently have difficulties in generalizing to newly invent or disguised attacks and can also be targeted by adversarial attacks. Especially, systems for signature-based detection have very few capabilities to change, which leads to a slow reaction to new threats and a higher number of false negatives. The AI-powered, adaptive system that is being proposed, on the other hand, makes it possible to detect crimes that have never been committed before as well as the ones that have been already solved by the integration of cutting-edge machine learning and generative AI technologies. This means a huge leap in generalization abilities compared to conventional models. To further improve the system's robustness, adversarial training is also incorporated, enabling the system to be very hard-to-fool methods like mimicry or carefully constructed adversarial inputs. Moreover, the suggested system enhances the abilities of the detection method to be able to recognize the multi-modal attack scenarios, such as text, voice, and multimedia-based social engineering, which most baseline models do not measure at all.

VI. CONCLUSION

The proposed model has demonstrated the significance of generative NLP based framework for the detection of phishing attacks. With the use of advanced NLP and generative AI methods, the model can detect known and prevailing unseen phishing such as subtle variations in email, chat or text communications. Rather than conventional rule-based or classification-only models, the generative technique permits for modelling the underlying patterns of phishing language and simulating adversarial manipulations, enhancing the robustness besides complication, imitation and distinct attacks. In addition with, the adaptive leaning of framework assures the iterative increase due to the rise of attackers' strategies. This technique decreases the risk window and improves the organization's flexibility alongside social engineering attacks with real time detecting and alerting. In the view of future works, the model could expand the detection for involving voice, visualisations and multimedia phishing for collective model training on organizations without revealing secured communications. On the combination of generative NLP with adaptive, real time detection, this study presents the further generation phishing defences capable of pro-actively forestalling and overcoming enhancing attacks of social engineering.

REFERENCES

[1]. Akter, T. (2025). A Taxonomy and Multi-Layered Defense Framework for Generative AI-Powered Phishing.

[2]. Alauthman, M., Aldweesh, A., Al-Qerem, A., Alkasassbeh, M., Alateef, S., & Almomani, A. (2025). Synthetic Content Generation Impacts on Phishing and Impersonation Attacks Examining Cybersecurity Risks Produced by Generative AI (pp. 189-210): IGI Global Scientific Publishing.

[3]. Alsop, H., Maglaras, L., Janicke, H., Sarker, I. H., & Ferrag, M. A. (2025). Innovating Augmented Reality Security: Recent E2E Encryption Approaches. arXiv preprint arXiv:2509.10313.

[4]. Beydemir, A. B., Sezgin, U., Doğan, U., Aşıklar, B. E., Yerlikaya, F. A., & Bahtiyar, Ş. (2024). A Dynamically Selected GPT Model for Phishing Detection. Paper presented at the 2024 14th International Conference on Advanced Computer Information Technologies (ACIT).

[5]. Desolda, G., Greco, F., & Viganò, L. (2025). APOLLO: A GPT-based tool to detect phishing emails and generate explanations that warn users. Proceedings of the ACM on Human-Computer Interaction, 9(4), 1-33.

[6]. Gupta, S., Ray, R. K., Singh, A., & Kapoor, A. P. (2026). Building Adaptive Digital Immune Systems: A Framework for Large-Scale Organizational

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may330

Resilience. Digital Immune System: Principles and Practices, 305-321.
[7]. Jamal, S., Wimmer, H., & Sarker, I. H. (2024). An improved transformer-based model for detecting phishing, spam and ham emails: A large language model approach. Security and Privacy, 7(5), e402.
[8]. Khan, H., Alam, M., Al-Kuwari, S., & Faheem, Y. (2021). Offensive AI: Unification of email generation through GPT-2 model with a game-theoretic approach for spear-phishing attacks. Paper presented at the IET Conference Proceedings CP786.
[9]. Koide, T., Fukushi, N., Nakano, H., & Chiba, D. (2024). Chatspamdetector: Leveraging large language models for effective phishing email detection. Paper presented at the International Conference on Security and Privacy in Communication Systems.
[10]. Koide, T., Nakano, H., & Chiba, D. (2024). Chatphishdetector: Detecting phishing sites using large language models. IEEE Access.
[11]. Ling, F., Yang, H., Xiao, Y., & Hu, L. (2024). Meta GPT-Based Agent for Enhanced Phishing Email Detection. Paper presented at the Proceedings of the 2024 14th International Conference on Communication and Network Security.
[12]. Masmoudi, S. (2025). Unveiling the human factor in cybercrime and cybersecurity: Motivations, behaviors, vulnerabilities, mitigation strategies, and research methods Cybercrime unveiled: Technologies for analysing legal complexity (pp. 41-91): Springer.
[13]. McIntosh, T. R., Susnjak, T., Arachchilage, N., Liu, T., Xu, D., Watters, P., & Halgamuge, M. N. (2025). Inadequacies of large language model benchmarks in the era of generative artificial intelligence. IEEE Transactions on Artificial Intelligence.
[14]. Merat, S., & Almuhtadi, W. (2025). Social Cyber Engineering and Advanced Security Algorithms: CRC Press.
[15]. Mohamed, N. (2025). Artificial intelligence and machine learning in cybersecurity: a deep dive into state-of-the-art techniques and future paradigms. Knowledge and Information Systems, 1-87.
[16]. Panda, S. P. (2025). The Evolution and Defense Against Social Engineering and Phishing Attacks. International Journal of Science and Research (IJSR).
[17]. Shibli, A. M., Pritom, M. M. A., & Gupta, M. (2024). Abusegpt: Abuse of generative ai chatbots to create smishing campaigns. Paper presented at the 2024 12th International Symposium on Digital Forensics and Security (ISDFS).
[18]. Tsinganos, N., Fouliras, P., & Mavridis, I. (2022). Applying BERT for early-stage recognition of persistence in chat-based social engineering attacks. Applied sciences, 12(23), 12353.
[19]. Tsinganos, N., Fouliras, P., & Mavridis, I. (2023). Leveraging dialogue state tracking for zero-shot chat-based social engineering attack recognition. Applied sciences, 13(8), 5110.
[20]. Tsinganos, N., Mavridis, I., & Gritzalis, D. (2022). Utilizing convolutional neural networks and word embeddings for early-stage recognition of persuasion in chat-based social engineering attacks. IEEE Access, 10, 108517-108529.
[21]. Verma, A., & Shri, C. (2025). Cyber security: A review of cyber crimes, security challenges and measures to control. Vision, 29(4), 478-492.
[22]. Zia, M. F., & Kalidass, S. H. (2024). Web Phishing Net (WPN): A scalable machine learning approach for real-time phishing campaign detection. Paper presented at the 2024 4th Intelligent Cybersecurity Conference (ICSC).