Array 30 (2026) 100775

Contents lists available at ScienceDirect

Array

journal homepage: www.elsevier.com/locate/array

A systematic literature review of large language models in phishing attack generation and detection ☆

Dinushan Sivaneswaran [a,*], Chaminda T.E.R. Hewage [b,*], H.M.K.K.M.B. Herath [c], Rajkumar Singh Rathore [b], Vishal Krishna Singh [d,*], Weiwei Jiang [e]

[a] School of Computing and Mathematical Sciences, University of Greenwich, London, UK
[b] Cybersecurity and Information Networks Centre (CINC), Cardiff School of Technologies, Cardiff Metropolitan University, Cardiff, UK
[c] Industry 4.0 Convergence Bionics Engineering, Pukyong National University, Busan, Republic of Korea
[d] School of Computer Science and Electronics Engineering, University of Essex, Colchester, UK
[e] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, China

ARTICLE INFO

Keywords:
Cybersecurity
Generative AI
Large language models
Phishing attacks
Phishing detection
Social engineering
Threat intelligence

ABSTRACT

Phishing attacks continue to grow in scale and sophistication, causing substantial financial losses and privacy breaches worldwide. Recent advances in large language models (LLMs) have brought significant changes to the generation and detection of phishing content. This study systematically investigates the dual role of LLMs in facilitating phishing attacks and strengthening countermeasures. Using the PRISMA methodology, authors screened 142 records published between January 2023 and April 2025 and identified 36 eligible studies from major academic databases, including IEEE Xplore, ScienceDirect, ACM Digital Library, Web of Science, and Scopus. A comprehensive and rigorous analysis was conducted of research trends/themes over time, dataset characteristics, and the LLM architectures/models employed. The findings reveal that most studies relied on manually generated datasets rather than publicly available benchmark datasets, and that GPT-based models received considerably more attention than other LLM architectures. The review demonstrates that LLMs substantially enhance the generation of phishing content by producing coherent, contextually relevant, and persuasive email and website content. This capability lowers the technical barrier for attackers and potentially increases attack effectiveness. Conversely, LLMs also strengthen defensive strategies by enabling more effective analysis of textual and visual content for phishing detection. In many cases, LLM-based approaches outperform traditional machine learning and deep learning methods and, in certain contexts, approach or match human-level performance. Overall, the findings suggest that LLMs have accelerated and automated phishing-related processes, simultaneously intensifying the threat landscape and advancing defensive capabilities.

1. Introduction

Cyberattacks have emerged as a significant threat in the digital age, affecting individuals, corporations, and governments worldwide. These attacks encompass a wide range of malicious activities, including malware infections, ransomware campaigns, denial-of-service (DoS) incidents, and various forms of social engineering [1,2]. Among these, phishing stands out as one of the most prevalent and dangerous techniques used by cybercriminals to deceive individuals into disclosing sensitive information. These attacks are typically carried out through emails, fake websites, or messaging platforms, where the attacker poses as a trusted entity to deceive the victim [3–5]. As phishing methods continue to evolve, they have become more targeted and convincing, making detection increasingly difficult. Due to its widespread use, low cost, and high success rate, phishing remains a primary method for initiating larger cyberattacks such as credential theft, financial fraud, and data breaches [6,7].

In recent years, large language models (LLMs) have undergone a revolution and are now widely used by people to simplify and accelerate various tasks. LLMs have enabled individuals to complete tasks such as drafting poems, writing code, and detecting errors, even without deep technical knowledge. Commonly used LLM tools include ChatGPT, Gemini, and Claude. Today, LLM is not only capable of processing text, but also images, videos, and audio [8,9]. Several studies have shown that LLMs can match or even outperform humans on certain

☆ This article is part of a Special issue entitled: 'ARRAY_InSeTMA' published in Array.

* Corresponding authors.
E-mail addresses: ds5533s@greenwich.ac.uk (D. Sivaneswaran), chewage@cardiffmet.ac.uk (C.T.E.R. Hewage), kasunkh@pukyong.ac.kr (H.M.K.K.M.B. Herath), rsrathore@cardiffmet.ac.uk (R.S. Rathore), v.k.singh@essex.ac.uk (V.K. Singh), jww@bupt.edu.cn (W. Jiang).

https://doi.org/10.1016/j.array.2026.100775
Received 19 January 2026; Received in revised form 26 February 2026; Accepted 20 March 2026
Available online 21 March 2026

tasks [10–12]. These tools help to speed up workflows, improve outcomes, and make processes more efficient and accessible [12,13]. It is worth noting that LLM tools can be used for both beneficial and harmful purposes. Fortunately, most modern LLM systems are equipped with security measures that help prevent misuse. However, malicious users sometimes find ways to bypass these protections by manipulating prompts or using other techniques [13].

LLMs are widely used by researchers, AI engineers, and data scientists across diverse applications, such as building intelligent chatbots, performing sentiment analysis, and generating or debugging code [1]. Prominent models such as GPT-3, GPT-4, LLaMA-3, and Claude-2 offer diverse parameter sizes and multiple versions, allowing for flexibility across different applications [1]. To tailor these models for specific tasks, practitioners employ techniques like prompt engineering, fine-tuning, and even developing new models based on existing architectures to achieve optimized performance [6,7,14]. Additionally, LLM models are leveraged to fully automate complex workflows, such as writing and deploying code, retrieving information from the internet, and generating structured documents or summaries [15,16].

Phishing attacks in the pre-LLM era were primarily carried out manually, often via email or web, and relied on traditional social engineering and technical methods to extract sensitive information from users with limited cybersecurity awareness [17,18]. Common techniques included tab-napping, email spoofing, Trojan horses, and hacking [18]. Detection during this period focused on statistical and lexical analysis, user education, and, in some studies, traditional machine learning models and NLP-based feature extraction [17,19].

With the emergence of LLMs, both phishing attacks and their detection mechanisms have evolved significantly. LLMs enable the generation of more sophisticated, context-aware, and human-like phishing content, while defensive approaches increasingly leverage AI capabilities to enhance detection accuracy and adaptability. National and sector-level reports further underscore the persistence of phishing as a dominant cyber threat. The Federal Bureau of Investigation's Internet Crime Complaint Center (IC3) ranked phishing/spoofing as the top complaint category in 2024, with 193,407 reported cases. Similarly, the UK Government's Cyber Security Breaches Survey 2025 estimates that millions of phishing-related cybercrimes continue to affect businesses and charities. In this context, LLMs are best understood as amplifying an already entrenched problem, lowering the barriers to producing more convincing and scalable attacks while simultaneously enabling new tools for phishing detection and analysis.

The COVID-19 pandemic marked a significant turning point, resulting in a substantial increase in both the volume and impact of phishing attacks. Reports from the Anti-Phishing Working Group (APWG) indicate sustained high levels of activity prior to the pandemic, with 158,574 attacks recorded in Q4 2015 and 162,155 phishing sites detected in Q4 2019. Moreover, phishing incidents increased sharply in 2020, with activity approximately doubling during the early phase of the pandemic. More recently, record-breaking figures have been observed, including 1,003,924 attacks in Q1 2025, reflecting a continued upward trend throughout 2024. The global lockdowns and rapid transition to remote work and online services created new opportunities for attackers, leading to a surge in cyber incidents during the pandemic [20,21]. Several studies report that phishing campaigns intensified during this period, often exploiting COVID-19-related themes and heightened levels of fear, anxiety, and stress among users [22]. In addition, the rapid adoption of remote work without fully strengthened security infrastructures further exposed organizations and individuals to phishing and related attacks [23]. These findings indicate that the pandemic marked a significant shift in the scale and behavioral dynamics of phishing attacks.

LLM has been misused by some malicious actors to carry out phishing attacks. Several studies have reported that attackers use prompt engineering and evasion techniques to bypass security filters [13,24]. Research indicates that LLM tools can accelerate the phishing content generation process, assist inexperienced users in crafting convincing phishing messages, support multilingual phishing campaigns, and even automate the entire phishing workflow [5,25,26]. These capabilities highlight the potential threat posed by LLM in the creation and execution of phishing attacks.

Conversely, LLMs have demonstrated significant potential for detecting phishing attacks. Studies have demonstrated that LLMs can match, and in some cases outperform, human performance in identifying phishing content. Some research further highlights their capability for multi-modal analysis, such as interpreting screenshots alongside URLs to improve detection accuracy [7,27]. In the context of phishing email detection, LLM models can analyze metadata, grammatical and spelling errors, tone and writing style, and indicators of AI-generated content [3,13]. These comprehensive analytical capabilities make LLMs a robust and reliable tool for identifying phishing threats. Moreover, several studies have reported that LLM-based models outperform traditional machine learning and deep learning approaches in phishing detection tasks [28,29]. Beyond detection, LLM tools are also used to analyze cyber threat intelligence (CTI) reports, assist in developing training materials, and enhance security awareness programs [30,31]. These applications underscore the valuable role that LLMs can play in enhancing cybersecurity defenses, particularly in phishing prevention and user education.

The structure of this article consists of a detailed explanation of the study conducted. In Section 2, a comparison between existing related survey papers and this work is presented. Section 3 outlines all the methods followed to carry out the survey. Section 4 presents a detailed statistical analysis of the research articles gathered for the study. The following Section 5 clearly explains how LLM is used in the generation of phishing attacks, while the subsequent Section 6 discusses the use of LLM in phishing detection and other cybersecurity applications. Section 7 presents the general analysis of the study and, finally, the Conclusion section summarizes the key findings and contributions.

The main contributions of this review are as follows:

- A systematic synthesis of literature retrieved from major academic databases, ensuring comprehensive coverage of peer-reviewed research.
- Quantitative analysis of publication trends, datasets, and LLM architectures to illustrate the evolution of research in this domain.
- A critical evaluation of LLM-enabled phishing attacks, including examining realism, text-based and multimodal threats, scalability, automation, hybrid approach, and human-AI collaboration.
- A structured assessment of LLM-based phishing detection approaches, including comparisons with human experts and traditional models, alongside analysis of prompting, fine-tuning, agentic, and human-in-the-loop frameworks, as well as applications in cyber threat intelligence and security awareness.

2. Related work

This section positions our systematic literature review (SLR) relative to recent studies on phishing and LLMs. Prior work addresses parts of the problem, specific phishing modalities, selected detection settings, or broader cybersecurity applications, but few studies jointly and systematically examine both phishing attack generation and phishing detection under modern generative AI capabilities across multiple attack types.

Blancaflor et al. [32] investigated AI voice-cloning threats in phone-based social engineering, with a specific emphasis on potential impact on the Filipino population. While valuable for understanding vishing risk, the work is largely limited to voice-based deception and does not extend to other phishing modalities or to LLM-enabled attack generation and detection more broadly.

Table 1
Comparison of prior survey studies and the present systematic literature review (SLR), showing each study's scope and differences in phishing modality coverage and consideration of LLM-based attack generation and detection.

Work	Scope and Key Difference vs. Our SLR
Blancaflor et al. [32]	Focuses on AI voice cloning for voice phishing (vishing) and population impact; our SLR covers multiple phishing types and analyzes both LLM-enabled generation and detection.
Chen et al. [33]	Surveys LLMs for cyber threat detection/CTI with limited phishing-specific depth; our SLR provides a phishing-centered synthesis across modalities and across generation and detection.
Ding et al. [34]	Reviews LLMs for cyber resilience and discusses spam/threat detection tasks, with brief coverage of phishing; our SLR provides a focused analysis of phishing generation and detection.
Veit et al. [35]	Systematizes email deception techniques and client susceptibility without examining LLM roles; our SLR analyzes how LLMs enable and mitigate email-based phishing.
Hasanov et al. [36]	Includes phishing email/message generation and detection within a broad review, but lacks depth in web-based phishing; our SLR is phishing-focused and includes web-based phishing.
Li et al. [37]	Reviews phishing website detection (including LLM influence in detection) but not LLM-driven website generation; our SLR covers both detection and generation across phishing types.

Several reviews discuss LLMs in cybersecurity, but treat phishing only partially. Chen et al. [33] surveyed LLM use for cyber threat intelligence (CTI) and threat detection in textual artifacts (e.g., emails and messages), but provided limited phishing-specific analysis. Ding et al. [34] focused on cyber resilience and evaluated transformer-based models on tasks such as spam classification and threat detection; however, phishing is discussed only briefly, primarily through adjacent datasets and settings, rather than as a dedicated end-to-end area covering both generation and detection.

Other work examines phishing techniques without explicitly addressing LLMs. Veit et al. [35] analyzed email deception strategies (e.g., sender identity spoofing, links, and attachments) and the susceptibility of email clients, but did not consider the role of LLMs in enabling or countering these techniques.

Finally, two studies are closely aligned with our topic but remain narrower in scope. Hasanov et al. [36] reviewed LLM applications in cybersecurity and included phishing email/message generation and detection; however, phishing was presented as one component within a broader survey and did not include a focused treatment of web-based phishing. Li et al. [37] provided a detailed review of phishing website detection techniques and discussed LLM-related influences in detection, but did not address the use of LLMs for phishing website generation.

In contrast to these works, our SLR provides a dedicated, integrated analysis of LLM-driven phishing attack generation and detection across multiple phishing forms (e.g., email, SMS, voice, and web-based). By consolidating findings across modalities and across both offensive and defensive uses of LLMs, our review aims to offer a comprehensive reference for researchers and practitioners studying how generative AI reshapes the phishing threat landscape. Table 1 summarizes the scope and key differences of the related work with respect to our SLR.

3. Source selection and inclusion

This study followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) protocol to guide the selection of articles, ensuring transparency, thoroughness, and reproducibility throughout the review.

Table 2
Search queries applied in academic databases.

Database	Search Query
IEEE Xplore	((“Document Title”:phishing) AND (“Document Title”:“Generative AI” OR “Generative Artificial Intelligence” OR “Large Language” OR GenAI OR “Gen AI” OR LLM OR LLMs OR ChatGPT OR GPT)) OR ((“Abstract”:phishing) AND (“Abstract”:“Generative AI” OR “Generative Artificial Intelligence” OR “Large Language” OR GenAI OR “Gen AI” OR LLM OR LLMs OR ChatGPT OR GPT)) OR ((“Author Keywords”:phishing) AND (“Author Keywords”:“Generative AI” OR “Generative Artificial Intelligence” OR “Large Language” OR GenAI OR “Gen AI” OR LLM OR LLMs OR ChatGPT OR GPT))
ScienceDirect	Title, abstract, keywords: phishing AND (“Generative AI” OR “Generative Artificial Intelligence” OR “Large Language” OR GenAI OR “Gen AI” OR LLM OR LLMs OR ChatGPT OR GPT)
ACM	((Title: phishing) AND ((Title: “generative ai”) OR (Title: “generative artificial intelligence”) OR (Title: “large language”) OR (Title: genai) OR (Title: “gen ai”) OR (Title: llm) OR (Title: llms) OR (Title: chatgpt) OR (Title: gpt))) OR ((Keyword: phishing) AND ((Keyword: “generative ai”) OR (Keyword: “generative artificial intelligence”) OR (Keyword: “large language”) OR (Keyword: genai) OR (Keyword: “gen ai”) OR (Keyword: llm) OR (Keyword: llms) OR (Keyword: chatgpt) OR (Keyword: gpt))) OR ((Abstract: phishing) AND ((Abstract: “generative ai”) OR (Abstract: “generative artificial intelligence”) OR (Abstract: “large language”) OR (Abstract: genai) OR (Abstract: “gen ai”) OR (Abstract: llm) OR (Abstract: llms) OR (Abstract: chatgpt) OR (Abstract: gpt)))
WoS	TS=(phishing AND (“Generative AI” OR “Generative Artificial Intelligence” OR “Large Language” OR GenAI OR “Gen AI” OR LLM OR LLMs OR ChatGPT OR GPT))
Scopus	TITLE-ABS-KEY(phishing AND (“Generative AI” OR “Generative Artificial Intelligence” OR “Large Language” OR GenAI OR “Gen AI” OR LLM OR LLMs OR ChatGPT OR GPT))

Table 3
Inclusion criteria applied during the PRISMA-based screening process.

Criterion	Description
Years	January 2023 – April 2025
Language	English only
Type	Research articles
Title	Relevant to LLM and phishing
Full Text	Full text available

The initial step involved creating a search query to effectively gather relevant articles from various academic databases, including IEEE Xplore, ScienceDirect, ACM Digital Library, Web of Science (WoS), and Scopus. The search query was applied across all five databases. The query used was: phishing AND ("Generative AI" OR "Generative Artificial Intelligence" OR "Large Language" OR GenAI OR “Gen AI” OR LLM OR LLMs OR ChatGPT OR GPT). This query was applied to the title, keywords, and abstract fields. This query was applied to the databases as summarized in Table 2.

Filters were applied to get relevant research articles. The articles were gathered between January 2023 and April 2025. Also selected English-only articles. Moreover, excluded review papers from the analysis. Furthermore, after careful reading, relevant articles were selected for the study. the criteria is presented in Table 3.

The PRISMA diagram of this study is shown in Fig. 1 [38]. Initially, 142 articles were collected from the selected databases. After applying the selection criteria, excluding articles not directly related to the topic and removing duplicates, 66 records remained. Following a review of the abstracts, 16 articles were excluded for irrelevance to the study's

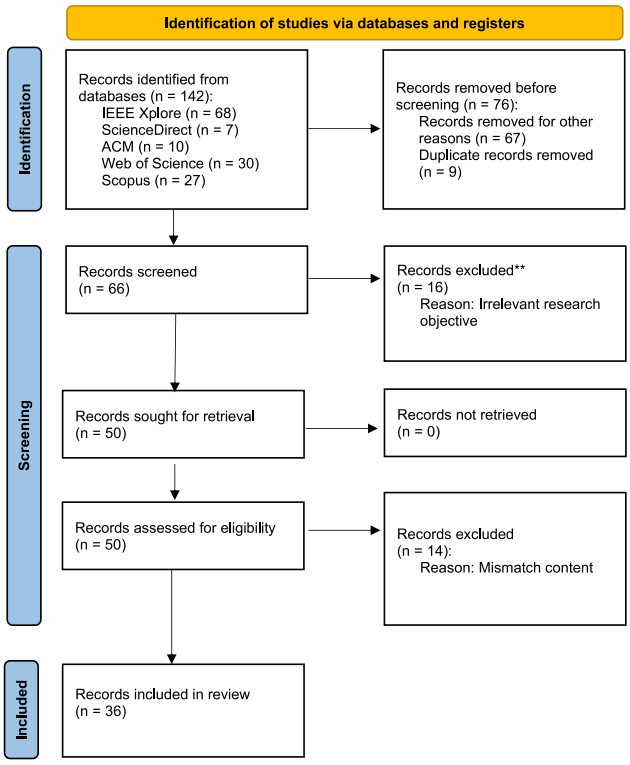

Fig. 1. PRISMA 2020 flow diagram of the study selection process. A total of 142 records were identified from IEEE Xplore, ScienceDirect, ACM, Web of Science, and Scopus, illustrating the systematic screening and review procedure.

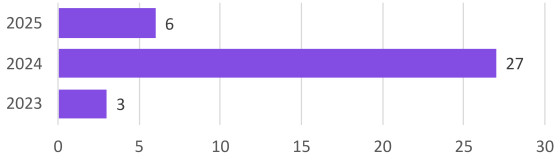

Fig. 2. Distribution of the 36 included studies by publication year (2023–2025), showing a marked increase in 2024 and reflecting growing academic interest in LLM-based phishing attack generation and detection.

focus, leaving 50. All remaining articles were retrieved in full. After a detailed review of the full text, 14 articles were excluded due to content mismatch. A total of 36 articles were ultimately selected for inclusion in the study.

4. Statistical analysis of the literature

4.1. Time frame of included studies

This study encompasses research articles on phishing attack generation and detection published between 2023 and 2025, as this period represents recent research following the widespread adoption of generative AI and LLMs in phishing attacks and detection. The number of articles analyzed for each year is shown in Fig. 2.

According to the statistics, 3 articles published in 2023, 27 in 2024, and 6 in 2025 were included in this study. According to the statistics, the majority of the analyzed articles were published in 2024. Since this study includes only articles up to the end of March 2025, only a limited number of 2025 publications were considered. Additionally, only a small number of relevant articles from 2023 were included, as per the selection and exclusion criteria.

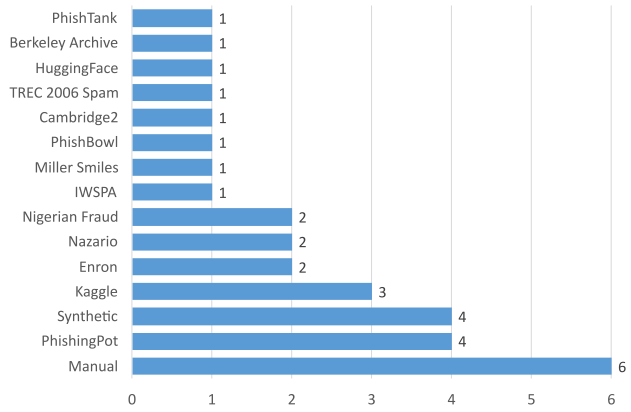

Fig. 3. Frequency of dataset sources used in phishing email detection studies across the reviewed articles, highlighting a predominant reliance on manually curated or synthetic datasets rather than standardized benchmarks.

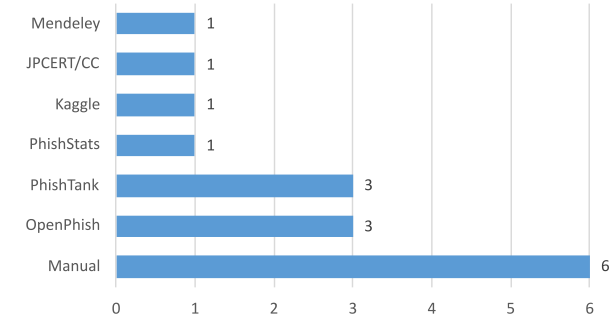

Fig. 4. Dataset sources in phishing website detection studies, with manual collection most common (6 studies), OpenPhish and PhishTank next (3 each), and PhishStats, Kaggle, JPCERT/CC, and Mendeley used once each.

4.2. Dataset used in research publications

4.2.1. Dataset used in phishing email detection

Various datasets from different sources were used across the analyzed articles related to phishing email detection. Table 4 provides details on the sources of the datasets, the number of articles that used each dataset, the web links to the datasets, and the corresponding references. Moreover, a bar chart (Fig. 3) is included to illustrate the number of articles that used each dataset.

According to the analysis, most of the articles used manually collected datasets for their experiments. Notably, the second-most-common type of data was synthetic. Among publicly available datasets, PhishingPot was the most frequently used, with a usage count equal to that of synthetic datasets. Kaggle was the second-most-used source. Additionally, datasets such as Enron, Nazario, and Nigerian Fraud were also commonly referenced. Other datasets, including IWSPA, Miller Smiles, PhishBowl, Cambridge2, TREC 2006 Spam, HuggingFace, Berkeley Archive, and PhishTank, were utilized in limited studies.

4.2.2. Dataset used in phishing web detection

Several articles have utilized various datasets for phishing website detection. The Table 5 summarizes these datasets, and Fig. 4 provides a usage count analysis.

According to the analysis, most of the articles utilized manually collected datasets. Following that, the OpenPhish and PhishTank datasets were the most commonly used. In addition, a limited studies employed datasets from PhishStats, Kaggle, JPCERT/CC, and Mendeley.

Table 4
Summary of datasets used in phishing email detection studies, including dataset source, usage frequency, links (where available), and corresponding references. The analysis shows that manual and synthetic datasets were most frequently used, indicating limited standardization in the adoption of benchmark datasets.

Dataset/Source	Count	Links	Research Publications
Manual	6	*N/A*	[6,14,39–42]
PhishingPot	4	https://github.com/rf-peixoto/phishing_pot	[6,9,41,43]
Synthetic	4	*N/A*	[14,28,44,45]
Kaggle	3	https://www.kaggle.com/datasets	[6,9,46]
Enron	2	https://www.cs.cmu.edu/~enron/	[3,28]
Nazario	2	https://zenodo.org/records/8339691	[29,44]
Nigerian Fraud	2	https://figshare.com/articles/dataset/Phishing_Email_11_Curated_Datasets/24952503	[29,44]
IWSPA	1	https://github.com/ReDASers/IWSPA-2023-Adversarial-Synthetic-Dataset	[3]
Miller Smiles	1	https://www.zbh.uni-hamburg.de/forschung/amd/datasets/smarts-dataset.html	[3]
PhishBowl	1	https://informationsecurity.princeton.edu/phish-bowl	[3]
Cambridge2	1	https://www.cambridgecybercrime.uk/data.html	[3]
TREC 2006 Spam	1	https://plg.uwaterloo.ca/~gvcormac/treccorpus06/	[43]
HuggingFace	1	https://huggingface.co/datasets	[44]
Berkeley Archive	1	https://www.lib.berkeley.edu/visit/bancroft/university-archives	[45]
PhishTank	1	https://phishtank.org/	[42]

Table 5
Datasets used in phishing website detection research, with frequency counts and references. Manual datasets dominate, followed by OpenPhish and PhishTank, highlighting variability in dataset selection across studies.

Dataset/Source	Count	Links	Research publications
Manual	6	N/A	[4,7,27,47–49]
OpenPhish	3	https://www.openphish.com/phishing_database.html	[7,27,47]
PhishTank	3	https://phishtank.org/	[4,7,47]
PhishStats	1	https://phishstats.info/	[47]
Kaggle	1	https://www.kaggle.com/datasets	[4]
JPCERT/CC	1	https://blogs.jpcert.or.jp	[4]
Mendeley	1	https://data.mendeley.com/	[50]

Table 6
Datasets used in Cyber Threat Intelligence (CTI) analysis studies, including manual and public intelligence sources, highlighting the limited diversity in CTI-focused LLM research.

Dataset/Source	Count	Links	Research publications
Manual	1	N/A	[30]
Cisco Talos Intelligence Group	1	https://www.talosintelligence.com	[16]
Microsoft Security Intelligence Center	1	https://www.microsoft.com/en-us/security/blog/	[16]
MITRE ATT&CK	1	16.https://attack.mitre.org/	[16]

4.2.3. Dataset used in cyber threat intelligence (CTI) report analysis

Some articles have experimented with using LLMs for CTI report analysis. For these experiments, the studies used the MITRE ATT&CK CTI reports [51]. Table 6 presents a summary of the datasets used in these studies.

According to the statistics, a research article has utilized manually created datasets. A study has utilized datasets from sources such as the MITRE ATT&CK, Cisco Talos Intelligence Group, and the Microsoft Security Intelligence Center.

4.2.4. Dataset used for user training and security awareness programs

This study reviews several research articles that examine the application of LLMs in user training and security awareness programs. Out of these, Yamin et al. [52] used organizational documents to perform a better training program. while Hafzullah [31] used only LLM to perform the training program.

4.3. LLMs used in research publications

4.3.1. LLM used in phishing attack generation

Many studies have conducted experiments on phishing attack generation, employing various LLMs from different model families. Table 7 summarizes the number of times each LLM has been used across these research articles, while Fig. 5 provides a visual representation of their usage frequencies.

According to the analysis, many articles have used GPT models in their experiments. The second most commonly used LLM is Gemini.

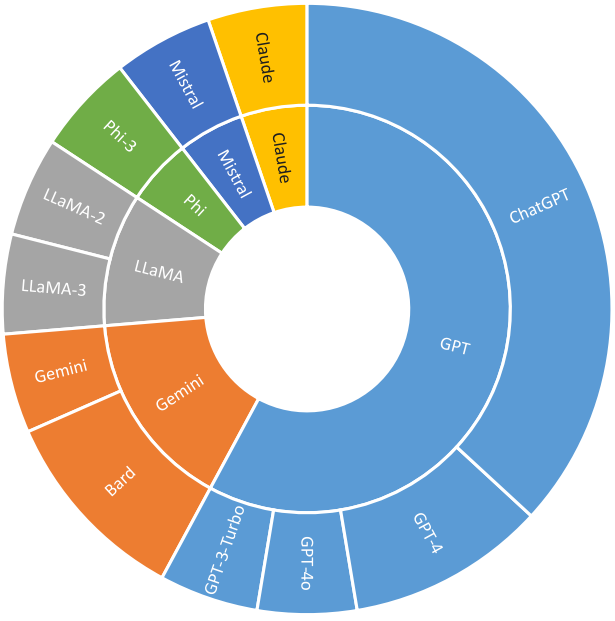

Fig. 5. Usage frequency of LLM families in phishing attack generation studies, showing GPT-family models as most common, followed by Gemini and LLaMA, with limited use of Claude, Mistral, and Phi.

Table 7
LLM models used in phishing attack generation studies, grouped by model family and usage frequency, showing GPT-family models as dominant and highlighting the concentration on specific proprietary systems.

Family	LLM Model	Count	Research publications
GPT	ChatGPT	7	[5,10,11,13,15,25,26]
	GPT-4	2	[11,45]
	GPT-3-Turbo	1	[53]
	GPT-4o	1	[29]
Gemini	Bard	2	[11,13]
	Gemini	1	[25]
LLaMA	LLaMA-2	1	[53]
	LLaMA-3	1	[53]
Claude	Claude	1	[11]
Mistral	Mistral	1	[53]
Phi	Phi-3	1	[53]

Table 8
LLM models used in phishing email detection studies, with frequency counts and references, showing a wider variety of models compared to generation studies and reflecting broader experimentation in defensive applications.

Family	LLM Model	Count	Research Publications
GPT	GPT-3.5	4	[9,29,43,46]
	GPT-4	4	[29,40,45,46]
	GPT-4o	3	[6,9]
	GPT-3.5-Turbo	3	[14,28,40]
	GPT-4-Turbo	2	[14,40]
	GPT-4V	1	[40]
Gemini	Gemini	2	[6,29]
	Gemini-1.5	2	[44]
	Gemini-1.5-Pro	1	[9]
	Bard	1	[45]
LLaMA	LLaMA-3	3	[6,9,29,41]
	LLaMA-2	2	[3,45]
Claude	Claude-3	2	[6,29]
	Claude-1	1	[45]
	Claude-3.5	1	[9]
Qwen	Qwen-2	2	[6,9]
	Qwen-Max	1	[43]
Mistral	Mistral	2	[3,6]
CyberGPT	CyberGPT	2	[39,46]
T5	T5-Large	1	[3]
	Flan-T5-Large	1	[3]
Gemma	Gemma-2	1	[9]
	Gemma	1	[6]
Phi	Phi-3-Medium	1	[9]
	Phi-3-Mini	1	[6]
GLM	GLM-4	1	[43]
Falcon	Falcon	1	[3]
Nous-Hermes	Nous-Hermes-2	1	[9]
Open-Hermes	Open-Hermes-2.5	1	[9]
Yi	Yi-1.5	1	[9]
Aya	Aya-101	1	[6]

Additionally, some articles have utilized LLaMA. A limited number of articles have used Claude, Mistral, and Phi models.

4.3.2. LLMs used in phishing email detection

Numerous studies have conducted experiments on phishing email detection, employing various LLMs from different model families. Table 8 summarizes the frequency of each LLM used across these research articles, while Fig. 6 provides a doughnut chart for a visual representation of the counts.

According to statistics, many articles have used GPT-based LLMs in their experiments, particularly GPT-3.5, GPT-4, GPT-4o, GPT-3.5-Turbo, GPT-4-Turbo, and GPT-4V. The second-most-common LLM family was Gemini. Some articles have also used LLaMA, Claude, Qwen, Mistral, CyberGPT, T5, Gemma, and Phi models. A limited number of studies have used GLM, Falcon, Nous-Hermes, OpenHermes, Yi, and Aya LLMs in their experiments.

Table 9
LLM models used in phishing website detection research, categorized by model family and usage frequency, with GPT and LLaMA models most prevalent, reflecting dominance patterns similar to those in email detection studies.

Family	LLM Model	Count	Research Publications
GPT	GPT-4	3	[7,27,47]
	GPT-4o	2	[7,48]
	ChatGPT	1	[54]
	GPT-3.5-Turbo	2	[7,50]
	GPT-4V	1	[7]
	GPT-3.5	1	[48]
	GPT-4o-Mini	1	[49]
	GPT	1	[50]
	GPT-2	1	[50]
	GPT-2-Medium	1	[50]
	DistilGPT-2	1	[50]
LLaMA	LLaMA-3	3	[7,47,48]
	LLaMA-2	2	[7,47]
	Baby-LLaMA	1	[50]
Gemini	Gemini-1.0-Pro	2	[7,27]
	Gemini	1	[47]
Claude	Claude-3	2	[27,47]
	Claude-2	1	[50]
Gemma	Gemma-2	2	[7,49]
Command-R+	Command-R+	1	[7]
Bloom	Bloom-560 m	1	[50]
Mistral	Mistral-NeMo	1	[49]
Phi	Phi-3-Mini	1	[49]

4.3.3. LLMs used in phishing website detection

Several studies have conducted experiments on phishing website detection, utilizing various LLMs from different model families. Table 9 summarizes the number of times each LLM has been used across these research papers, and Fig. 7 provides a visual representation of the usage counts.

According to the analysis, similar to phishing detection-related articles, the most commonly used LLM in web phishing detection articles is GPT. The second-most-used LLM is LLaMA, which was used in 6 articles. Gemini and Claude were also among the frequently used models. Some articles used Gemma, while a limited number of articles used Command-R+, Bloom, Mistral, and Phi.

4.3.4. Open and closed LLMs

LLMs differ in openness, which influences their role in phishing research. Open models, such as LLaMA (Meta), Mistral (Mistral AI), and T5 (Google), provide access to weights and architectures, allowing researchers to create phishing content for study and develop detection methods through fine-tuning and experimentation. Closed models, including GPT (OpenAI), Gemini (Google), Claude (Anthropic), Qwen (Alibaba Cloud), and Phi (Microsoft), are primarily accessible via APIs, which limits direct experimentation but enables controlled, safer deployment. Together, this mix of open and closed models shapes a research landscape where phishing generation and detection can advance while balancing ethical, security, and practical deployment considerations.

5. LLM-driven phishing attack generation and associated risks

This section examines the role of LLMs in generating phishing attacks. It covers how LLMs can produce convincing phishing emails, websites, and visual and voice-based attack content. The discussion also examines how LLMs impact the scale and quality of phishing campaigns. Additionally, it considers the risks posed by LLM vulnerabilities that attackers might exploit for phishing.

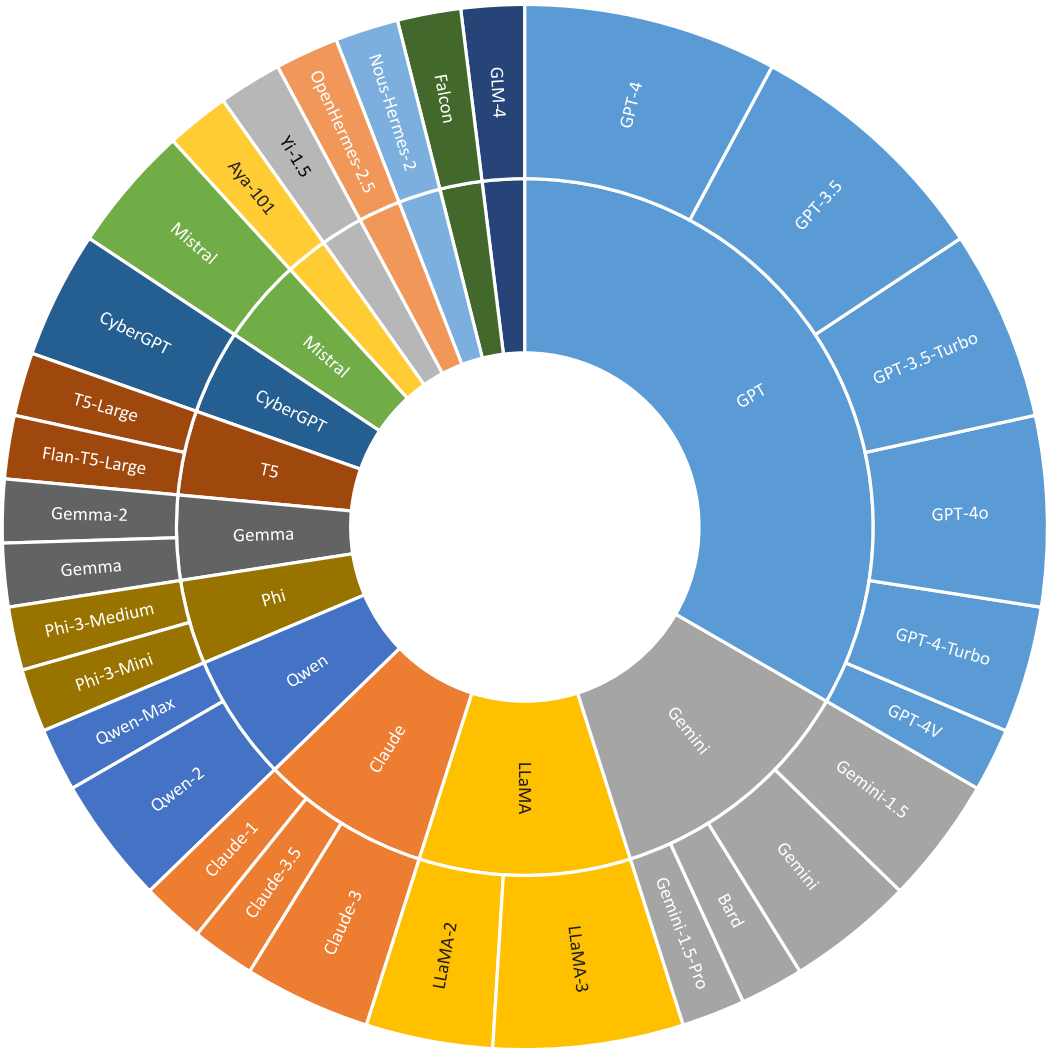

Fig. 6. Distribution of LLM models in phishing email detection studies, with GPT-family models (GPT-3.5, GPT-4, GPT-4o, and Turbo variants) most commonly evaluated, followed by Gemini, LLaMA, Claude, Qwen, and Mistral.

5.1. Generating realistic phishing emails using LLMs

The emergence of Large Language Models (LLMs) has significantly enhanced the realism of phishing emails, enabling attackers to craft messages that closely resemble legitimate communications, with high levels of personalization and professionalism. This subsection examines the performance of LLMs in generating phishing emails, comparing their outputs with human-created content, exploring hybrid approaches, methods to enhance phishing emails, and their capability to generate multilingual phishing content.

5.1.1. LLM performance in phishing email generation

Nagarajan et al. [25] noted that generative AI can produce highly realistic phishing emails that effectively mimic reputable sources in a more professional and convincing manner. It can also incorporate personalized, distinctive details, making the content appear significantly more authentic. Moreover, the authors have highlighted that it can draft high-quality phishing content with minimal or no grammar or spelling errors.

Roy et al. [11] tested the effectiveness of ChatGPT-3.5, GPT-4, Claude, and Bard in generating phishing attacks, including phishing emails. The study employed several NLP evaluation metrics, including BLEU, ROUGE-1, Perplexity, and Topic Coherence. According to the results, GPT-4 outperformed the other models, achieving the highest scores in BLEU (0.54), ROUGE-1 (0.68), and Topic Coherence (0.72), while also having the lowest Perplexity (15), indicating the production of high-quality and coherent phishing content. Claude closely matched performance, while GPT-3.5 achieved moderate results. Bard, on the other hand, performed slightly lower than the other models, with scores of BLEU (0.46), ROUGE-1 (0.58), Topic Coherence (0.62), and Perplexity (20).

Moreover, Fairbanks et al. [53] have compared the performance of LLM models in generating phishing emails. According to the results, the LLM models such as GPT-3-Turbo, LLaMA-2-7B, LLaMA-2-70b, and Mistral-7B performed well compared to LLaMA-2-13B, LLaMA-3-7B, LLaMA-3-70B, and Phi-3.

5.1.2. LLM-generated emails vs human-crafted emails

To better understand the impact of LLM-generated phishing emails, it is essential to compare them with those created by cybersecurity experts.

Bethany et al. [10] discussed the threat of phishing emails in the context of the emergence of LLMs. This research was conducted at a large public university that employed approximately 9000 staff members across various fields, including faculty, staff, administrators, and student workers. The study compared three types of phishing scenarios: emails from a supervisor to a direct report, organizational-level emails on a timely event, and common phishing emails from an

Fig. 7. Frequency of LLM models in phishing website detection research, showing GPT-family models as most common, followed by LLaMA, Gemini, and Claude, with occasional use of Gemma, Bloom, Mistral, and Phi.

unfamiliar internal employee account. The study found that phishing emails generated by LLMs were as effective as, and in some cases even more effective than, those created by cybersecurity professionals. In terms of overall performance, LLM-generated emails had an open success rate of 66.08%, a link click success rate of 12.72%, and a data entry success rate of 5.12%. In comparison, human-written emails achieved an open success rate of 65.10%, a link click success rate of 10.11%, and a data entry success rate of 3.56%. In the "timely phishing" scenario, LLM-generated emails significantly outperformed human-crafted emails, particularly in open and link click rates. The LLM-generated emails achieved an open success rate of 66.08% and a link click success rate of 17.62%, whereas human-written emails had an open success rate of 53.33% and a link click success rate of 9.78%. This represents nearly double the success rate in link clicks for LLM-generated emails. These all demonstrate how effective LLMs can be in creating phishing emails.

5.1.3. The effectiveness of hybrid approaches

Interestingly, combining LLMs with traditional methods has also shown promise in creating highly effective phishing emails. This approach would help avoid detection if the LLM is also used in the detection process.

Heiding et al. [45] compared the effectiveness of phishing emails created automatically by the GPT-4, manually using the V-Triad framework, a combination of both, and generic phishing emails. This experiment involved simulating attackers and emailing 112 participants. The results showed that V-Triad-generated emails achieved the highest click-through rates, followed by hybrid-generated emails and GPT-4-generated emails. Specifically, the study found that emails created using the V-Triad achieved click-through rates between 69% and 79%, the hybrid approach ranged from 43% to 81%, GPT-4-generated emails from 30% to 44%, and the control group from 19% to 28%. The study highlighted that combining GPT-4 with the V-Triad framework provides an effective method for creating phishing emails. This approach achieved results comparable to those of LLMs alone for generating emails and outperformed the existing phishing emails in the control group. These findings underscore the growing effectiveness of LLM in generating highly convincing phishing emails.

5.1.4. Rewriting phishing emails using LLMs

In the study [29], the results indicate that LLMs can even be used to rephrase phishing emails, making them more difficult to detect using traditional systems such as Gmail Spam Filter, SpamAssassin, and Proofpoint. The study shows that, using the original Nigerian Fraud Dataset, Gmail Spam Filter achieved an accuracy of 97.88%, SpamAssassin 96.75%, and Proofpoint 95.88%. However, after rephrasing the emails using few-shot prompting with GPT-4o, the detection accuracy dropped to 92.00% for Gmail Spam Filter, 89.75% for SpamAssassin, and 93.00% for Proofpoint. This demonstrates that LLMs can also be exploited to enhance the effectiveness of phishing emails.

Fairbanks et al. [53] investigated how LLMs perform in rewriting phishing emails. The study selected detectable malicious phishing emails and applied a novel LLM-based automatic output optimization technique to rewrite them. The approach focused on preserving the original Indicators of Compromise (IOCs) and their semantic meaning while embedding prompt injections in email headers to render them invisible to the user. The results demonstrate that rewritten phishing emails were significantly more successful at evading detection compared to the original malicious emails. Specifically, the models GPT-3.5-Turbo, LLaMA-2 (7B, 13B, 70B), LLaMA-3 (7B, 70B), Mistral-7B, and Phi-3 achieved success rates of 0.91, 0.86, 0.58, 0.98, 0.50, 0.56, 0.85, and 0.19, respectively. These findings highlight that LLMs can substantially increase the effectiveness of phishing attacks when leveraged for adversarial rewriting.

5.1.5. Multilingual spear-phishing email generation using LLMs

Recent studies show that LLMs can craft convincing spear-phishing emails in multiple languages. Gradon and Kacper [5] tested ChatGPT's capabilities for creating spear-phishing emails in both English and Polish and found that it can generate sophisticated, convincing, and reasonable spear-phishing emails without any linguistic, stylistic, grammatical, or spelling errors. Further mentioned that this makes it difficult to differentiate phishing emails from genuine ones. Furthermore, the study highlighted that the accessibility and ease of use of LLMs have lowered the barrier for new offenders to conduct phishing attacks. The findings of this study highlight that LLMs have significantly contributed to the rise of more widespread and personalized phishing attacks.

5.2. Creating phishing websites using LLMs

LLMs are increasingly used not only to generate phishing emails but also to create entire phishing websites, enabling attackers to convincingly mimic legitimate pages and deceive users into revealing sensitive information. This subsection discusses how LLMs are leveraged to design phishing websites and automate their deployment.

5.2.1. HTML generation for phishing sites

One of the key contributions of LLMs in creating phishing websites is their ability to generate HTML code that is both accurate and professional. Nagarajan et al. [25] have noted that generative AI can generate HTML code to create highly realistic phishing websites without spelling or grammatical errors, making it harder to visually distinguish a genuine website from a phishing website.

Roy et al. [11] evaluated the effectiveness of ChatGPT-3.5, GPT-4, Claude, and Bard in generating phishing websites. For the assessment, three independent coders were asked to rate the appearance of each site on a scale from 1 (worst) to 5 (best). A total of 80 samples were reviewed per LLM, covering ten types of phishing attacks: Regular Phishing, ReCAPTCHA Attacks, QR Code Attacks, Exploiting DOM Classifiers, iFrame Injection/Clickjacking, Browser-in-the-Browser Attacks, Polymorphic URLs, and Text Encoding Exploits. According to the results, 80% of GPT-4's samples scored nearly 5, indicating near-perfect quality. In comparison, 80% of GPT-3.5 and Claude's samples scored around 4, reflecting good quality, while 80% of Bard's samples scored only around 2.8 or lower, indicating poor performance. Notably, GPT-4 outperformed all other models across all ten attack types, solidifying its position as the top-performing LLM. In the Exploiting DOM Classifiers category, GPT-4 successfully generated all 10 phishing samples, while GPT-3.5 achieved 7, Claude 8, and Bard only 4. The study also concluded that phishing websites generated by LLMs, except Bard, were comparable in quality to those created by humans.

5.2.2. End-to-end phishing attack automation

In addition to HTML generation, LLMs have been used to automate entire phishing attack workflows. In [15], the study investigated the use of ChatGPT to design and automate the deployment of advanced phishing attacks. The researchers stated that they successfully generated various components of a phishing site using ChatGPT, including cloning a website, integrating credential-stealing code, obfuscating code, automating website deployment, registering a phishing domain, and integrating a reverse proxy. While working, the researchers faced limitations in the token size of the GPT-3.5-Turbo-16K model, which can hinder code generation for very large websites. Moreover, the study highlighted that by simplifying tasks such as code generation, site cloning, LLMs have opened the door for even those with little technical background to create and launch phishing sites, making these kinds of attacks more widespread and successful.

5.3. Developing multimodal phishing using LLMs

In today's threat landscape, phishing attacks have evolved beyond text-based deception. This subsection discusses the capability of LLMs or GenAI in general to generate highly realistic fake images and to facilitate fully automated scam communications.

5.3.1. Image and voice-based attacks enabled by LLMs

According to articles [13,49], LLMs can be used to generate deceptive images to build false trust. Additionally, studies such as [49,55] have highlighted that LLMs can be employed to conduct scam calls, where TTS and STT modules enable complete automation of verbal interactions with victims. These articles underscore the critical need to raise public awareness about the growing threat posed by convincingly AI-generated images and voice-based phishing attacks.

5.4. The impact of LLMs on phishing attack scalability and quality

The automation capabilities of LLMs have greatly increased the scalability of phishing attacks, enabling even individuals with limited technical skills to create and launch sophisticated, convincing campaigns with minimal effort. This subsection discusses the impact of LLMs on the scalability and quality of phishing attacks.

5.4.1. Automation and ease of use in phishing attacks

Iturbe et al. [26] experimented with the LLM-based mechanism for automatically generating code that maps attack techniques and tactics defined by the MITRE ATT&CK framework. The study was conducted in three steps: AI-based code generation using ChatGPT-3.5, execution of the generated code in a sandbox environment across both Windows and Linux, and evaluation of the results with cybersecurity experts. The study's results indicate that 48.32% of the generated phishing attack code successfully breached operating systems. Furthermore, the study highlighted that the code generated for Linux had a higher success rate than that for Windows. Moreover, the article noted that LLMs make it much easier and more effective to perform phishing attacks.

Moreover, in [15], the study states that website deployment was automated using ChatGPT. Moreover, the researchers noted a limitation in token size, which can hinder the process when working on larger tasks.

Authors in [5] highlighted that events like the COVID-19 pandemic and the war have already shown how harmful the spread of false information and hybrid threats can be. It also pointed out that LLM tools like ChatGPT can worsen the situation by making it easier to produce convincing, targeted, and misleading content, such as phishing emails. The study further noted that LLMs enable rapid creation of high-quality, convincing, logical, and error-free phishing content, making it difficult to distinguish from genuine communications. Moreover, it was noted that these LLMs make it easier for attackers to create and execute phishing attacks more quickly, without requiring extensive expertise or domain knowledge. Overall, the reviewed studies highlighted that LLMs significantly enhance the scalability of phishing attacks by lowering the technical barrier to execution and accelerating content generation, thereby positively contributing to the wider spread of such threats.

5.4.2. Automation benefits vs. Practical limitations

Divakaran et al. [12] discussed how large language models influence cybersecurity, and the authors also addressed phishing attacks. This article mentioned that LLMs can be a powerful tool for hackers due to their ability to produce human-like natural language. Furthermore, the study noted that scammers can utilize LLMs to enhance the grammar, prose, and translation of phishing emails, as well as to randomize their content. However, author mentioned that LLMs are unlikely to drastically change the landscape of standard phishing attacks, for a few key reasons, which were, since there are already many tools to help write, spell-check, and translate phishing emails automatically,

LLMs do not offer a major advantage in this area., another was since modern anti-phishing systems do not just look at the message content, they also analyze metadata like the sender's identity, embedded URLs, and file attachments, making it harder for LLM-generated messages alone to succeed, another was even though making the emails too polished, it will not work with people who are unlikely to be fooled, which may increase unwanted operational time. Overall, the study concluded that LLMs have the potential to help hackers reduce cost and automate attacks, but likely increase the quantity rather than the quality. Moreover, the study highlighted that detection and signature generation will be challenging, as there will be a limited number of high-quality phishing emails generated through LLMs.

5.5. Exploiting LLM vulnerabilities

Although LLMs incorporate safety mechanisms, studies have shown that attackers can exploit model vulnerabilities and employ various techniques to bypass these safeguards, facilitating phishing and other malicious activities. This subsection provides a detailed discussion of these issues.

5.5.1. Bypassing ethical safeguards in chatgpt and bard

Gupta et al. [13] have described that LLMs have vulnerabilities that can be exploited by malicious users. The article mentioned several techniques that attackers widely use to bypass ethical restrictions and engage in harmful activities, including jailbreaks (DAN method, SWITCH method, CHARACTER Play), reverse psychology, ChatGPT-4 model evasion, and prompt injection attacks. The article compared the cybersecurity capabilities of ChatGPT and Google's Bard and presented some findings. While ChatGPT often declined to generate attack code due to its ethical guidelines, Bard proved more unpredictable and sometimes provided useful code snippets without jailbreaking, posing a real threat to humanity. These vulnerabilities underscore the urgent need to enhance both ethical and technical defenses in LLMs to prevent their misuse in phishing and other cyber threats.

6. LLM-based approaches for phishing detection and cybersecurity applications

This section analyzes how LLMs can be utilized to detect phishing attacks. It reviews existing research across several key areas, organized into subsections covering LLM performance in phishing email detection, phishing website detection, the analysis of Cyber Threat Intelligence (CTI) reports, and the use of LLMs in training and awareness programs.

6.1. Email phishing detection using LLMs

This subsection reviews several research articles that evaluate the performance of LLMs in phishing email detection. It includes comparisons between LLMs and human detection capabilities, traditional machine learning models, and transformer-based models. Additionally, it examines performance differences among various LLMs. Moreover, the analysis also considers the impact of fine-tuning, the integration of complementary NLP techniques, and the use of persuasion principles. Furthermore, it explores agentic AI approaches and the effectiveness of human-in-the-loop systems.

To support the discussion, several tables are included. Table 10 presents a summary of studies comparing the performance of LLMs with traditional machine learning models. In this table, scores separated by commas represent results obtained using different methods or datasets within the same study. Table 11 summarizes studies that benchmark LLMs against Transformer-based models. Finally, Table 12 presents an overview of studies that compare the performance of various LLMs in phishing email detection.

6.1.1. LLMs vs. Human detection

Studies reported in [45,49] evaluate the effectiveness of LLMs in comparison to humans in detecting phishing emails, highlighting how these LLMs perform against humans in detection. Chang et al. [49] experimented with the usage of LLMs in detecting phishing emails, and the results showed that LLaMA-3-8B demonstrated the highest overall performance among the evaluated lightweight LLMs: GPT-4o-Mini, Mistral-NeMo, Gemma-2-9B, Phi-3-Mini. Furthermore, it has been stated that LLMs sometimes outperform humans in identifying phishing emails.

Moreover, Heiding et al. [45] have evaluated the ability of four LLMs (GPT, Claude, PaLM, and LLaMA) to detect phishing emails. It has been stated that LLMs demonstrated a strong ability to detect malicious emails, sometimes even surpassing human detection in accuracy, recall, and F1-score. Furthermore, it was noted that LLMs outperformed humans at identifying various types of phishing emails, especially real-world and V-Triad + GPT-generated phishing emails. Notably, the results indicate that the Claude LLM model achieved the best detection performance.

6.1.2. LLMs vs. Traditional machine learning models

Beyond human comparisons, a critical line of research focuses on how LLMs outperform compared to traditional machine learning and deep learning algorithms in phishing email detection.

Beydemir et al. [28] analyzed the performance of traditional machine learning models, such as Support Vector Machines (SVMs) and Random Forests, compared with a fine-tuned large language model, GPT-3.5-Turbo. This study focuses on early phishing email detection, analyzing email content rather than relying on hyperlinks or attachments. In this study, the proposed detection system categorizes phishing attempts into six threat types: information-gathering queries, verification requests, impersonation messages, deceptive proposals, emotional manipulation, and urgent requests. The overall results of this study demonstrated that the LLM significantly outperformed the traditional ML models in terms of accuracy.

Chataut et al. [39] also investigated the detection of phishing email using LLMs and traditional machine learning algorithms, namely CyberGPT, Logistic Regression Model, Random Forest Model, and Gradient Boosting Model. In this study, the author has considered several key features for detection, including sender and recipient information, timestamps, email body content, and the number of URLs. The study concluded that LLM (customized CyberGPT) outperformed traditional machine learning algorithms in phishing email detection based on key features.

Furthermore, in article [29], the results demonstrate that LLMs, specifically GPT-4, GPT-3.5, Claude 3 Sonnet, and LLaMA-3, outperformed traditional machine learning algorithms such as Support Vector Machines (SVM), Logistic Regression, and Naive Bayes in phishing email detection. The study also found that LLMs surpassed conventional phishing detection systems, including Google's Gmail Spam Filter, Apache SpamAssassin, and Proofpoint. Using the Nigerian Fraud Dataset, Naive Bayes achieved 95.38% accuracy, SVM 96.12%, and Logistic Regression 96.75%. Among conventional systems, Gmail's Spam Filter reached 97.88% accuracy, SpamAssassin 96.75%, and Proofpoint 95.88%. In contrast, the LLMs demonstrated superior performance, with GPT-4 achieving 99.12%, Claude 3 Sonnet 98.00%, LLaMA-3 98.50%, GPT-3.5 95.38%, and Gemini 96.00%. These findings underscore the effectiveness of LLMs in phishing detection tasks.

In contrast, Dumitras et al. [41] examined the effectiveness of combining traditional ML models with LLMs for phishing email detection, but their results showed that this integration did not improve detection performance.

6.1.3. LLM vs. Transformers models

In addition to traditional machine learning models, several studies have compared the performance of LLMs with that of transformer-based models. This section examines how transformer models compare to LLMs in phishing email detection tasks.

Beydemir et al. [28] not only compared the performance of a large language model (LLM) with traditional machine learning models, but also evaluated it against a Transformer-based model. The study found that both BERT and GPT-3.5-Turbo outperformed traditional ML models, with GPT-3.5-Turbo achieving the best performance, surpassing BERT during testing. Similarly, Authors in [4] compared a custom LLM with both traditional ML models and Transformer-based models, including BERT and all-MiniLM, and concluded that the custom LLM significantly outperformed them. Collectively, these articles highlight the superior capability of fine-tuned LLMs in phishing email detection when compared to both traditional machine learning and earlier Transformer models.

In the article [44], the study compares the performance of the transformer-based DeBERTa model and large language models for phishing detection. In the study, authors have used HuggingFace, the Nazario, and Nigerian Fraud phishing datasets, along with synthetic data. The findings show that LLMs and DeBERTa V3 performed similarly on public datasets, and LLMs notably outperformed DeBERTa V3 on the synthetic dataset. However, the author notes that the DeBERTa model trained on a subset of data showed slightly greater improvement in detecting synthetic data. The author concluded that DeBERTa performed better in generalization detection with properly aligned data, while LLMs performed well in recognizing new phishing patterns and emerging tactics.

6.1.4. Zero-shot and prompt-based LLM evaluation

This subsection evaluates the performance of various LLMs in detecting phishing emails in different prompt-based.

Siemerink et al. [14] stated that, in prompting strategies, zero-shot prompting consistently outperformed both multi-shot and Chain-of-Thought (CoT) prompting. According to the accuracy results, using GPT-3.5 with a generalized dataset, zero-shot prompting achieved 88.77%, compared to 87.43% for multi-shot prompting and 84.22% for CoT prompting. Similarly, zero-shot prompting also outperformed the other methods when evaluated on synthetic datasets using both GPT-3.5 and GPT-4. However, when GPT-4 was tested on a generalized dataset, multi-shot and CoT prompting slightly outperformed zero-shot prompting by approximately 1%. Overall, both LLMs demonstrated strong performance, with GPT-4 outperforming GPT-3.5 in phishing email detection.

Zhang et al. [9] conducted research on building a phishing email detecting system, particularly for Small and Medium-sized Enterprises (SMEs). The article proposes a prompt template that enables LLMs to detect phishing emails and provides explanations and suggestions to aid users in making informed decisions. The research results indicate that the LLaMA-3-8b-instruct model outperformed other LLM models. The article concludes that LLMs perform well in email phishing detection, making them a more valuable tool for Small and Midsize Enterprises.

Moreover, in study [3], the authors compared the detection accuracy of several LLM models, namely Mistral-7B, T5-Large, LLaMA-2-7B, and Falcon-7B. The study results indicate that among these, Mistral-7B outperformed the other models, achieving an accuracy of 80.8%.

Furthermore, in article [6], several LLMs were compared for accuracy across multilingual datasets in a zero-shot setting. The models included GPT-4o, Gemini-1.5, Claude-3, Gemma-7B, Mistral-2-7B, Aya-101, LLaMA-3-7B, Phi-3-Mini, and Qwen-2-7B. According to the results, GPT-4o, Gemini-1.5, and Claude-3 performed remarkably well on both datasets, achieving over 90% accuracy.

6.1.5. Fine-tuning strategies for improving LLM detection accuracy

Several studies have shown that fine-tuning strategies can significantly improve LLMs' detection accuracy in phishing email detection.

In the study [46], the study compared the performance of GPT-3.5, GPT-4, and a fine-tuned version of ChatGPT called CyberGPT in detecting phishing emails. According to the results, GPT-3.5 achieved 80.68% accuracy, GPT-4 achieved 97.22%, and CyberGPT achieved 97.46%. These findings indicate that GPT-4 and CyberGPT significantly outperformed GPT-3.5. Moreover, CyberGPT slightly outperformed GPT-4.

Similarly, Zhang et al. [40] have evaluated two versions of ChatGPT (GPT-3.5 and GPT-4) for detecting phishing content through systematic tests of text and image recognition. Further, the study experimented by fine-tuning the GPT-3.5-Turbo-0125 model. Overall, the study found that the fine-tuned model performed remarkably well at detecting phishing emails that contain URLs.

Moreover, study [6] examined the impact of fine-tuning on the performance of several small language models, including Mistral-2-7B, Aya-101, LLaMA-3-7B, Phi-3-Mini, and Qwen-2-7B. The experiments were conducted on two multilingual datasets: Phishing_Pot (comprising phishing emails in 19 languages) and a social media dataset from the Arabic platform Baaz. The results indicate that, in the zero-shot setting, models achieved moderate performance—approximately 70% accuracy on Phishing_Pot and 55% on the Baaz dataset. However, after fine-tuning, the accuracies significantly boosted to around 95% and 90% on the respective datasets. Notably, for the Arabic dataset, larger LLMs such as GPT-4o, Gemini-1.5, and Claude-3 performed strongly in zero-shot evaluations, achieving accuracy scores of 91.21%, 91.78%, and 90.11%, respectively. In contrast, smaller models such as Gemma-7B, Mistral-7B, Aya-101, LLaMA-3-7B, Phi-3-Mini, and Qwen-2-7B achieved lower zero-shot accuracy, typically ranging from 50% to 60%. After fine-tuning, these smaller models reached accuracy levels of around 90%, demonstrating the significant performance gains achievable through task-specific adaptation. Overall, these findings highlight the crucial role of fine-tuning in enhancing LLM effectiveness, especially for multilingual phishing detection tasks.

These studies highlight that fine-tuning significantly increases the accuracy of LLM.

6.1.6. Enhancing LLM-based phishing detection with additional NLP techniques

Sayyafzadeh et al. [42] mentioned that integrating Natural Language Processing (NLP) techniques with Large Language Models (LLMs), significantly improves phishing email detection. The results showed that GPT-4 combined with VADER achieved 95.8% accuracy, GPT-4 with RoBERTa achieved 92%, and GPT-4 with ALBERT achieved 90%, compared to GPT-4 alone, which achieved 88%. These findings indicate that incorporating NLP techniques with LLM can significantly enhance the performance in phishing detection.

6.1.7. Use of persuasion principles

Several studies have analyzed the key approach of enhancing phishing email detection in LLMs by leveraging principles of persuasion.

The study presented in [3], suggests that a fine-tuned LLM model that also considers Cialdini's Principles of Persuasion (Scarcity, Authority, Consistency) and Leakage Cues (Enticement, Urgency Tactics, Personalized Greeting) significantly improves the detection of anomalies in phishing messages.

6.1.8. Agentic architectures

One emerging advancement in LLM evolution is the adoption of agentic approaches. In the era of LLMs, this approach has significantly enhanced AI's ability to autonomously analyze and respond to complex tasks such as phishing detection.

In the paper [43], the study focused on developing an email phishing-detection system using agentic AI. Three agents were used for detection: one to analyze header information to detect spoofing.

Table 10
Overview of the comparison between LLMs and traditional machine learning models in email phishing detection, highlighting the overall dominance of GPT models.

Article	LLM		Traditional ML		Best Model(s)
	Model	Performance	Model	Performance	
[28]	GPT-3.5-Turbo	95 %	SVM	85 %	GPT-3.5-Turbo
			Random Forest	88 %	
[39]	CyberGPT	98.47%	Logistic Reg.	98.4%	CyberGPT
			Random Forest	97.5%	
			Gradient Boosting	97.2%	
[41]	CB + LLM	0.9058	CB	0.9058	CB
	EBM + LLM	0.8834	EBM	0.8879	CB+LLaMA-3
	KNN + LLM	0.8744	KNN	0.9058	
	LR + LLM	0.8520	LR	0.8655	
	RF + LLM	0.9103	RF	0.9148	
	SVM + LLM	0.8700	SVM	0.8879	
	XGB + LLM	0.8924	XGB	0.8924	
[29][a]	GPT-4	99.12%, 95.75%, 94.88%	Naive Bayes	95.38%, 90.50%, 88.88%	GPT-4
	GPT-3.5	95.38%, 91.62%, 89.50%	SVM	96.12%, 90.25%, 87.12%	
	Claude-3	98.00%, 95.00%, 94.62%	Logistic Reg.	96.75%, 89.50%, 88.50%	
	LLaMA-3	98.50%, 96.62%, 94.88%			
	Gemini	96.00%, 88.25%, 86.62%			

[a] Three values are reported due to evaluation on three different datasets.

Table 11
Overview of the comparison between LLMs and transformer models in email phishing detection, highlighting LLM dominance.

Article	LLM		Transformer		Best Model(s)
	Model	Performance	Model	Performance	
[28]	GPT-3.5-Turbo	95 %	BERT	92 %	GPT-3.5-Turbo
[4]	Custom LLM	0.8390	BERT	0.7151	Custom LLM
			all-MiniLM	0.5678	
[44]	Gemini-1.5	0.77311	DeBERTa V3	0.28570	Gemini-1.5

Another tool for analyzing body content to check whether it contains any social engineering tactics or suspicious elements. For the prompt, the author has used a chain-of-thought approach. The final agent makes a decision based on the results of the two agents above and outputs the result in JSON format. The experimental results demonstrate that the agentic method performed remarkably well in detecting phishing emails.

6.1.9. Human-in-the-loop systems

In recent years, the Human-in-the-Loop approach has emerged as a significant step in enhancing LLM performance by integrating human interaction into the process.

Nguyen et al. [3] discussed utilizing LLMs and a human-in-the-loop approach to generate warnings for phishing emails. The article highlighted that involving a human in the loop to provide information that cannot be gleaned from the email header or body, such as whether the user recognizes the sender, significantly increases detection accuracy. According to the results, without human-in-the-loop, the LLM models, namely Mistral-7B, T5-Large, Flan-T5-Large, LLaMA-2-7B, and Falcon-7B, scored 80.8%, 77.9%, 77.9%, 76.7%, and 75.6%, respectively. After integrating the human-in-the-loop, Mistral-7B, T5-Large, Flan-T5-Large, LLaMA-2-7B, and Falcon-7B scored 82.1%, 81.1%, 79.1%, 77.9%, and 76.4%, respectively, highlighting that incorporating human feedback can enhance model detection accuracy.

Overall, these studies demonstrate that LLMs provide a strong baseline for phishing email detection, particularly when traditional models struggle with newer attack formats.

6.2. Malicious URLs and websites using LLMs

Recent research has expanded the application of LLMs beyond traditional text classification tasks to encompass the detection of phishing URLs and malicious websites. This subsection reviews studies that evaluate the effectiveness of LLMs in identifying web-based phishing threats. Specifically, it covers comparisons between LLMs and traditional machine learning models, evaluations in zero-shot and prompt-based settings, and multimodal approaches. Additionally, the subsection examines techniques such as obfuscation-resistant URL detection and domain squatting detection. To support these discussions, Table 13 presents accuracy scores of LLMs across various web phishing detection scenarios.

6.2.1. LLMs vs. Traditional machine learning models

Fu et al. [4] evaluated a custom LLM against various machine learning and deep learning models, including Deep Neural Networks (DNN), Long Short-Term Memory networks (LSTM), a combined DNN+LSTM model, K-Nearest Neighbors (KNN), and Feedforward Neural Networks (FNN). The study reported that the custom LLM achieved the highest accuracy of 84%, while most other models ranged from 40% to 65%, with KNN performing best at 78%. These results suggest that LLMs outperform traditional ML models in phishing web detection.

Moreover, Fu et al. [4] suggested that LLMs aid in building phishing detection systems, especially for obfuscated URLs, which are difficult for traditional machine learning and deep learning algorithms to effectively handle. In this study, the authors have focused on two prevalent techniques used in real-world obfuscation scenarios: Domain Obfuscation and Redirects. The study stated that the model achieved over 92% accuracy on the Redirects obfuscation test and over 83% accuracy on the Domain Obfuscation tests, indicating the effectiveness of LLM in detecting phishing URLs.

6.2.2. LLM models evaluation

Several studies have examined the performance of LLMs in detecting phishing websites using various prompt-based approaches, including zero-shot settings. This subsection discusses the LLMs that have demonstrated effective performance in detecting phishing websites.

Table 12
Overview of LLM model performance across different settings for email phishing detection, showcasing LLM effectiveness.

Article	LLM		Best Model(s)
	Model	Performance	
[3]	Mistral-7B	82.1%	Mistral-7B
	T5-Large	81.1%	
	Flan-T5-Large	79.1%	
	LLaMA-2-7B	77.9%	
	Falcon-7B	76.4%	
[43]	GLM-4	99.30%	GLM-4
	GPT-3.5	98.55%	
	Qwen-Max	98.00%	
[6][a]	GPT-4o	98.82%, 91.21%	Ensemble
	Gemini	99.01%, 91.78%	
	Claude-3	99.21%, 90.11%	
	Gemma-7B	73.25%, 52.10%	
	Mistral-2-7B	71.24%, 53.03%	
	Aya-101	70.04%, 59.83%	
	LLaMA-3-7B	69.15%, 48.30%	
	Phi-3-Mini	70.39%, 55.20%	
	Qwen-2-7B	72.45%, 52.10%	
	Mistral-2-7B (FT)	95.11%, 88.21%	
	Aya-101 (FT)	95.20%, 93.95%	
	Llama-3-7B (FT)	96.72%, 89.26%	
	Phi-3-Mini (FT)	95.40%, 88.15%	
	Qwen-2-7B (FT)	92.12%, 89.24%	
	Ensemble (SLMs)	98.50%, 95.25%	
[40]	GPT-4 (0125)	High	GPT-3.5 (0125)
	GPT-4-Turbo	Moderate	GPT-4 (1106)
	GPT-4 (1106)	Very High	
	GPT-4V	High	
	GPT-3.5 (0125)	Very High	
	GPT-3.5	High	
[9]	LLaMA-3-8b	0.975	LLaMA-3-8b
	Gemini-1.5-Pro	0.925	
	Phi-3-Medium-4k	0.925	
	LLaMA-3-70b	0.900	
	GPT-3.5	0.900	
	Qwen-2-7B	0.850	
	Gemma-2-9b	0.800	
	Nous-Hermes-2	0.800	
	Yi-1.5-9B-Chat	0.825	
	Claude-3.5	0.700	
	GPT-4o	0.650	
	OpenHermes 2.5	0.625	
[45]	GPT-4	25%	Claude-1
	Claude-1	75%	
	Bard	25%	
	LLaMA-2	0%	
[14]	GPT-4-Turbo	94.93%	GPT-4-Turbo
	GPT-3.5-Turbo	89.18%	
[46]	GPT-3.5	80.68%	CyberGPT
	GPT-4	97.22%	
	CyberGPT (FT)	97.46%	
[42][b]	GPT-4 (Vader, Roberta, Albert)	95.80%, 92.00%, 90.00%	GPT-4 & Vader
	GPT-4 (1-shot, 5-shot, 10-shot)	88.00%, 93.00%, 94.00%	

[a] Two values are reported due to evaluation on two different datasets.
[b] GPT-4 is evaluated using different NLP-based configurations and varying numbers of shots.

Rashid et al. [47] experimented with various prompting methods (zero-shot, one-shot, and few-shot) and various LLMs (GPT-4-Turbo, Claude-3 Opus, Gemini, LLaMA-3, and LLaMA-2) in detecting phishing URLs. The results showed that the GPT-4 Turbo model performed remarkably well across all prompting methods. Additionally, researchers have noted that the accuracy of all LLMs was lower in the zero- and five-shot settings than in the one-shot setting. Overall, the study concluded that GPT-4-Turbo with one-shot performed remarkably well at detecting phishing URLs, achieving an accuracy of 0.87 and an average F1 score of 0.92 on a random sample of 1000 URLs from the ISCX-2016, EBBU-2017, and HISPAR-Phishstats (HP) Datasets.

Trad et al. [50] analyzed the accuracy of LLMs in detecting phishing websites. Several LLMs were evaluated through fine-tuning, including Bloom-560M, DistilGPT-2, GPT, Baby-LLaMA, GPT-2, and GPT-2-Medium. The respective accuracy scores for these models were 92.40%, 95.90%, 96.10%, 96.60%, 96.60%, and 97.30%. According to the results, the fine-tuned GPT-2-Medium outperformed the other models in phishing URL detection. Moreover, Trad et al. [50] explored

prompt-based approaches by experimenting with zero-shot prompting, role-playing prompting, and Chain-of-Thought (CoT) prompting using the GPT-3.5-Turbo and Claude-2 models. The GPT-3.5 achieved accuracy scores of 81.68% in zero-shot, 85.19% in role-playing, and 89% with CoT prompting. Meanwhile, Claude-2 achieved higher accuracy scores of 91.40% in zero-shot, 92.70% in role-playing, and 92.90% with CoT prompting. These results demonstrate that CoT prompting significantly improves the accuracy of phishing URL detection. Moreover, in this case, Claude-2 outperformed GPT-3.5. Overall, the study concluded that there were significant performance differences between prompt-engineered and fine-tuned LLMs, with fine-tuning generally leading to better results.

6.2.3. Multimodal webpage analysis with LLMs

Another major approach to phishing website detection involves using a multimodal model that accepts both text and images as inputs, taking detection capabilities to the next level.

Lee et al. [27] experimented with the effectiveness of Multimodal Large Language Models in detecting phishing websites by considering webpage contents, such as logos, themes, favicons, and website URLs. This was conducted in two phases: the first phase involved performing brand identification by analyzing webpage screenshots and HTML content, and the second phase involved domain verification by comparing the identified brand with the domain name in the URL. The study concluded that this approach performed significantly better than the real-world brand-based phishing detection system, "VisualPhishNet". This indicates that multimedia features (screenshot, HTM) of a webpage utilizing a multi-model LLM significantly improve the performance of the phishing detection system.

Koide et al. [7] developed a system called ChatPhishDetector to detect phishing websites by leveraging an LLM and web-crawling techniques. Here, the authors have employed a web crawler to gather various information about the website (e.g., URL, HTML, screenshot). Here, the system operated in two modes: Normal mode, which accepts only text input, and Vision mode, which accepts both text and image inputs. The experimental results showed that GPT-4V achieved exceptional performance in vision mode, and GPT-4 demonstrated strong performance in normal mode. Furthermore, the study highlighted that in open-source models, LLaMA-3-70B performed so close to GPT-4, making it a cost-effective model.

Moreover, Study [54] proposed a LLM-based phishing detection approach that builds upon the work of Koide et al. [7] by extending prompts with URLs, adapting the Chain-of-Thought (CoT) technique, and incorporating additional parameters such as HTML content, OCR data, and embedded links. The study concluded that these enhancements led to improved detection performance, enabling the system to identify previously unrecognized phishing websites [7].

6.2.4. Domain squatting detection using LLMs

Chiba et al. [48] experimented with combining an LLM with domain-specific knowledge to build a robust domain squatting detection system called DomainLynx. This model consisted of four main components: Input Data Processing, Domain Name eXpansion (DNX), Threat Recognition Validation (TRV), and Output Generation. The DNX component, Retrieval Augmented Generation, identifies potential squatting domains, while the TRV component evaluates their squatting risk using techniques like must-pass domain injection to test the system's ability to detect malicious domains. The results indicate that DomainLynx achieved 94.7% accuracy in detecting squatting domains, making it a valuable tool for phishing detection.

These studies highlight how LLMs can complement traditional URL-based features by incorporating context from domain names, webpage text, or surrounding messages.

6.3. LLMs for cyber threat intelligence (CTI) analysis and generation

Beyond phishing detection, LLMs have also been used to analyze CTI reports. This subsection reviews studies that leverage LLMs to extract, structure, and summarize information from CTI reports and, in some cases, to generate CTI reports by extracting relevant threat details.

6.3.1. Cyber threat intelligence (CTI) report analysis using LLM-based multi-module pipelines

Zhang et al. [16] experimented with developing a model based on cyber threat intelligence (CTI) reports. The proposed model, AttacKG+, consists of four consecutive modules: rewriter, parser, identifier, and summarizer. All were implemented using instruction prompting and in-context learning empowered by LLMs. The rewrite module acts as a filter and organizer, which reads the CTI report and removes unnecessary parts and reorganizes the remaining information into sections based on the different stages of a cyberattack, following the MITRE ATT&CK framework's tactics. The parser module extracts key details from the rewrite module's output in a structured way. The identifier module links the attack actions from the parser module's output to known attack techniques in the MITRE ATT&CK framework for better understanding. The summarizer module provides a brief overview of what happened at each stage of the attack. This module gives a higher-level understanding of the attacker's progress and the impact at each stage. In conclusion, the study highlighted that the AttacKG+ model significantly outperformed existing CTI parsing solutions.

6.3.2. Comparative evaluation of LLMs and NER models for generating cyber threat intelligence (CTI) reports

Zacharis et al. [30] experimented on automating and streamlining the process of CTI generation by extracting key intelligence with limited human intervention and utilizing LLM. The study compared the performance of several LLM models, including Gemini 1.5 Pro, GPT-4o, and BERT-based models like TRAM and TTPHunter, along with custom Named Entity Recognition (NER) models called AiCEF NER, in extracting key attributes for CTI reports such as Attack Type, Motivation, Threat Name, Sector Affected, and Tactics, Techniques, and Procedures (TTPs) from cybersecurity texts. The study noted that these tools could help automate certain aspects of threat intelligence workflows, thereby reducing manual effort while maintaining a reasonable level of accuracy.

6.4. User training and security awareness programs using LLMs

LLMs are also being explored for more user-interacting applications like generating phishing simulations or training content. This subsection analyzes how LLMs can be utilized to conduct user training and awareness programs.

6.4.1. LLM-based security awareness training evaluation

Hafzullah [31] has examined the effectiveness of LLMs in security awareness training. Here, the author has calculated the Phish Prone Percentage (PPP) before and after training to quantitatively assess the efficacy of the LLM-driven SAT. In the study, platforms such as Caniphish and KnowBe4 were used for phishing simulations. The results demonstrate a remarkable decrease in the PPP from 18.3% to 6.3%, highlighting the potential use of LLM. This demonstrates a significant improvement in participants' ability to distinguish subtle phishing indicators through the use of real-world scenarios in training.

Table 13
Overview of LLM model performance across various settings for web phishing detection, highlighting GPT dominance.

Article	LLM		Best Model(s)
	Model	Performance	
[47]	GPT-4	0.87	GPT-4
	Claude-3	0.81	
	Gemini	0.72	
	LLaMA-3	0.71	
	LLaMA-2	0.50	
[4][a]	Custom LLM	0.92, 0.83	Custom LLM
[27]	GPT-4	0.92	GPT-4
	Gemini-1.0-Pro	0.90	
	Claude-3	0.81	
[7]	GPT-4V	99.20%	GPT-4V
	GPT-4o	98.60%	
	GPT-4	98.40%	
	LLaMA-3	98.30%	
	Command-R+	93.80%	
	Gemini-1.0-Pro	89.10%	
	GPT-3.5-Turbo	86.70%	
	LLaMA-2	74.10%	
	Gemma-2	64.00%	
[48]	GPT-3.5	89.90%	GPT-4o
	GPT-4o	96.20%	
	LLaMA-3	94.70%	
[54]	ChatGPT	94.10%	ChatGPT
[50]	GPT-3.5-Turbo	89.00%	GPT-2-Medium
	Claude-2	92.90%	
	Bloom-560 m	92.40%	
	DistilGPT-2	95.90%	
	GPT	96.10%	
	Baby-LLaMA	96.60%	
	GPT-2	96.60%	
	GPT-2-Medium	97.30%	

[a] Two values are reported due to evaluation on two different datasets.

6.4.2. LLM-based scenario generation using retrieval-augmented generation (RAG)

Yamin et al. [52] experimented with LLMs to generate dynamic, complex, and adaptable cybersecurity exercise scenarios. The study employs the Retrieval-Augmented Generation (RAG) approach to construct the model. Here, the author intended to build phishing attack scenarios using LLM for training purposes. This study involved two LLMs interacting in parallel: LLM1 (CISO) generates the organizational context, which LLM2 (Cybersecurity Expert) then refines by applying cybersecurity frameworks. The study concluded that using LLMs to create cybersecurity exercise scenarios was an innovative and effective approach for training security professionals.

7. Discussion

Several articles have analyzed the influence of LLMs on both the generation and detection of phishing attacks, as detailed in Sections 5 and 6. This section identifies and discusses the main themes and findings gathered from this SLR.

7.1. Influence of LLMs on the evolution of phishing attacks

Many research articles have highlighted that LLMs significantly influence and advance phishing attacks. For example, articles [11,25,53] emphasized that LLMs can generate highly realistic emails. The authors in [25] specifically noted that LLMs can even produce personalized emails. articles [11,53] analyzed multiple LLMs and found that GPT-3.5 and Mistral-7B performed the best, respectively. Moreover, authors in [10] compared the performance of LLM-generated emails with human-crafted ones and concluded that LLM-generated emails were as effective and in some cases even more effective than those written by humans. Additionally, authors in [45] highlighted that integrating LLMs with traditional methods, such as the V-Triad framework, leads to significantly better performance than using LLMs alone. Several studies have explored how LLMs can enhance phishing emails. articles [29,53] showed that refined phishing emails generated by LLMs achieved higher success rates compared to the original versions, demonstrating the advanced capabilities of LLMs. Furthermore, authors in [5] revealed that LLMs can generate phishing emails not only in English but also in other languages. These findings collectively indicate the powerful role of LLMs in generating effective phishing emails.

Furthermore, several articles have analyzed LLMs' capabilities for creating phishing websites. Roy et al. [11] specifically examined the performance of LLMs in generating various types of phishing websites. The study concluded that GPT-3.5 and Claude performed particularly well in this task, demonstrating the effectiveness of LLMs in phishing website creation. Moreover, Begou et al. [15] highlighted that LLMs are not only capable of generating sophisticated phishing emails but also of integrating credential-stealing code, obfuscating malicious scripts, automating website deployment, registering phishing domains, and incorporating reverse proxies. These findings demonstrate the extensive capabilities of LLMs in facilitating comprehensive phishing attacks.

Recent studies [23,41,43,53] indicate that LLMs not only improve the quality of phishing messages but also automate multiple stages of the phishing lifecycle, reducing the expertise and time needed to execute campaigns. In the reviewed literature, LLMs are used to (i) generate persuasive social-engineering content (including multilingual variants), (ii) produce and modify phishing code/HTML, and (iii) support end-to-end workflow automation. Importantly, some work reports that LLMs can assist with highly operational tasks such as cloning a target website, integrating credential-stealing logic, obfuscating code, automating deployment, registering phishing domains, and integrating

reverse proxies' steps that traditionally require specialist knowledge. Beyond site creation, LLMs can also automate the production of attack scripts aligned with known adversary tactics; for example, one study generated executable attack code mapped to the MITRE ATT&CK framework and evaluated it in sandboxed environments, showing that a substantial portion of generated code could successfully breach test systems. Collectively, these findings support the claim that LLMs facilitate phishing automation by turning complex, multi-step tasks (content, code, and infrastructure setup) into prompt-driven workflows, thereby increasing scalability and lowering the barrier to entry for less experienced attackers.

Moreover, articles [13,49,55] highlighted that LLMs can even generate visuals to build trust, making phishing attacks more advanced. These studies also highlighted that LLMs can simulate entire phone conversations, thereby elevating phishing attacks to a new level.

Many articles have discussed the scalability and quality of LLM outputs. articles [5,15,26] highlight that LLMs can automate phishing attacks, making them feasible even for beginners. Moreover, these studies reported promising results regarding the effectiveness of LLM-generated content. However, Divakaran et al. [12] noted that while LLMs help automate the process and generate a higher volume of phishing emails, the focus tends to be more on quantity than on producing high-quality emails.

Notably, Gupta et al. [13] highlighted that LLMs have certain vulnerabilities that attackers can exploit to create phishing emails. The study analyzed various methods attackers use to bypass security measures and found that some LLMs lack robust security protections. This lack of robustness poses significant risks and underscores the need to address these vulnerabilities.

While most current studies focus on single-model prompting, agentic AI LLM-based systems capable of planning, calling tools, and executing multi-step goals are increasingly relevant to phishing because they can operationalize entire workflows rather than producing isolated outputs. Survey research on LLM agents in cybersecurity contexts indicates that agentic systems can reduce attack costs and scale multi-stage operations by decomposing objectives into subtasks and leveraging external tools or APIs. In phishing attack generation, an agentic pipeline could autonomously perform OSINT-based reconnaissance, including gathering victim profiles and organizational context, draft highly tailored lures, generate multiple variants for A/B testing, select delivery channels (email, SMS, or web), and orchestrate infrastructure tasks such as domain registration, website cloning, deployment automation, and evasion-driven revisions. This approach extends the end-to-end automation already reported for LLM-assisted phishing website workflows, demonstrating how agentic AI can significantly increase the sophistication, scalability, and operational efficiency of phishing campaigns.

Altogether, studies show that LLMs have had a strong influence on how phishing attacks are created, whether through emails, fake webpages, visuals, or even voice-based methods. They have changed the way phishing works, often performing better than traditional machine learning models, transformer-based approaches, and even human-crafted content. These attacks are more realistic and have a higher success rate, making them harder for detection systems to catch. Some articles have also pointed out that LLMs have security gaps that attackers can exploit to circumvent safety filters. What is more, these tools make it easier for beginners with little knowledge to carry out sophisticated phishing attacks. Overall, it is clear that LLMs have taken phishing to a new level, introducing new cybersecurity challenges.

7.2. Influence of LLMs on the evolution of phishing detection

Several articles have discussed the use of LLMs in phishing email detection. Numerous studies have examined the effectiveness of LLMs in this task. articles [45,49] reported that LLMs are capable of detecting phishing emails even more accurately than humans. Additionally, articles [28,29,39] highlighted that LLMs significantly outperformed traditional machine learning algorithms in phishing detection. Similarly, articles [28,44] showed that LLMs also outperformed transformer-based models.

articles such as [3,6,9,14] compared the performance of various LLMs in phishing email detection, with GPT, Mistral, and Claude frequently emerging as the top-performing models. Furthermore, some studies explored how to further improve LLM performance. Studies [6, 40,46] emphasized that fine-tuning significantly enhances LLM performance, especially in detecting phishing emails written in languages other than English.

Other techniques have also been proposed to boost detection capabilities. For instance, authors in [42] suggested using NLP techniques, while authors in [3] highlighted that incorporating persuasion principles can enhance detection accuracy. Authors in [43] introduced the concept of "Agentic AI" as a potential next step in phishing detection, and authors in [3] also emphasized that incorporating human-in-the-loop mechanisms can elevate detection to a new level.

Furthermore, several articles discussed the performance of LLMs in detecting phishing websites. Study [4] noted that LLMs performed significantly better than traditional machine learning and deep learning models, especially when dealing with obfuscated URLs. Authors in [47] claimed that GPT-4, even with a one-shot prompt, achieved better results, while authors in [50] found that fine-tuning further improved LLM performance.

Studies such as [7,27,54] have highlighted the benefits of multimodal approaches, in which LLMs analyze both textual content and visual elements to improve detection accuracy. Additionally, Chiba et al. [48] mentioned the use of a custom LLM model specifically designed to detect squatting domains and phishing URLs.

Moreover, Zhang et al. [16] mentioned that LLMs perform well in analyzing and summarizing CTI reports. Similarly, authors in [30] found that LLMs are effective at extracting key details for generating CTI reports. Furthermore, authors in [31] suggested that using LLMs is a strong choice for enhancing cybersecurity training programs. In the same context, authors in [52] highlighted that the Retrieval-Augmented Generation (RAG) method can help create dynamic, complex cybersecurity exercise scenarios, thereby making training more adaptable and effective.

Recent advances in agentic AI have enabled phishing detection systems to adopt more proactive and automated capabilities. Investigation co-pilots can automatically enrich suspicious emails and URLs with WHOIS and brand-similarity checks, sandbox links, correlate signals with CTI, and generate analyst-ready explanations and response actions, such as block rules or takedown requests. This trend reflects broader industry observations: while attackers increasingly leverage AI to craft obfuscated phishing campaigns, defenders are similarly adopting AI to detect and mitigate these threats more effectively. However, agentic AI systems also introduce new risks, including misuse of tools, prompt injection, and unsafe autonomous behavior. Consequently, we recommend implementing safeguards when deploying agentic AI in both offensive and defensive phishing contexts. Such measures may include human-in-the-loop oversight, least-privileged access to tools, audit logging, and robust policy constraints to minimize potential harm.

Altogether, studies highlight that LLMs have a significant positive impact on detecting and preventing phishing attacks. Research has shown that LLMs perform well in identifying various types of phishing, including email-based and website-based attacks. Additionally, some studies have noted that LLMs are useful in analyzing CTI reports and extracting key information to generate new CTI reports. Moreover, LLMs have been shown to support user training and security awareness efforts, helping to educate users and reduce the success rate of phishing attacks. Overall, the evidence suggests that LLMs play an important role in mitigating phishing threats.

8. Conclusion and future work

This systematic literature review aimed to investigate how the emergence of large language models has impacted both phishing attack strategies and the corresponding defense mechanisms. This study reviewed articles from leading peer-reviewed journals, following the PRISMA protocol for selecting articles.

According to a statistical analysis of the literature, more relevant articles were published in 2024 than in 2023, with strong engagement continuing in 2025. The analysis of datasets used in the research indicates that many studies relied on manual collection methods rather than publicly available datasets. Notably, several researchers employed synthetic methods to generate new datasets, while others gathered data from platforms such as Kaggle, GitHub, and HuggingFace. The analysis of LLM usage in research articles shows that authors generally preferred GPT models, although there has also been growing interest in Gemini, LLaMA, and Claude models.

LLMs have significantly influenced the generation of phishing attacks. Numerous studies have demonstrated that LLMs, particularly GPT and LLaMA, can be highly effective in generating phishing emails. LLM-generated phishing emails were often found to be as effective, or even more effective, than those written by human experts. Their effectiveness increased further when combined with frameworks such as V-Traid. LLMs can also rewrite existing phishing emails to make them more persuasive and can generate content in multiple languages. Beyond emails, LLMs can create phishing websites by automating HTML generation and implementation, lowering the skill barrier for attackers. Moreover, LLMs can generate multimodal phishing content, including images, videos, and voice recordings, thereby enhancing the scalability and quality of attacks. Although LLMs have built-in safety mechanisms, attackers continue to find ways to bypass these restrictions.

LLMs are not only used to generate phishing content but also to detect and defend against it. Studies have shown that LLMs are highly effective in detecting phishing emails, often matching or outperforming human performance, particularly with GPT models. LLMs have significantly outperformed traditional machine learning algorithms and even transformer models, especially in zero-shot prompting. Fine-tuned LLMs performed even better in detection tasks. Research also found improved results when combining LLMs with NLP techniques and persuasion principles. Further, agentic AI approaches enhanced accuracy and effectiveness, while human-in-the-loop methods provided additional improvements. LLMs also showed strong performance in detecting malicious phishing URLs and websites, surpassing traditional machine learning approaches. Among the models, GPT, Claude, and LLaMA consistently performed well. Some studies have found that detection accuracy improves when sharing both the HTML code and a screenshot of a webpage, along with its URL. LLMs also proved effective in detecting domain squatting, analyzing CTI reports, automating CTI generation, and supporting user training and security awareness programs to mitigate phishing threats.

In conclusion, LLMs have profoundly transformed both phishing attack generation and detection. While they have enabled more sophisticated, harder-to-detect attacks, they have also powered advanced defensive techniques, making them a double-edged sword in cybersecurity. To ensure sustainable progress, the field now requires standardized evaluation protocols, consolidated benchmark datasets, and robustness testing against adversarial attacks. Without coordinated dataset development and reproducible evaluation frameworks, advances in LLM-based phishing research risk remaining fragmented and difficult to compare. Establishing unified benchmarks and defensible evaluation standards will be essential for guiding the next phase of AI-driven cybersecurity research while mitigating malicious use.

In the future, we plan to expand the scope of this review by covering more research articles from additional databases and a longer timeframe. Moreover, investigating how agentic AI approaches influence phishing attacks will be a key direction for future work. GenAI-based multimodal phishing attacks will play a significant role in the future. Therefore, a rigorous review of GenAI-based multimodal phishing attack generation and detection would add an important dimension to the existing literature.

CRediT authorship contribution statement

Dinushan Sivaneswaran: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Chaminda T.E.R. Hewage:** Validation, Supervision, Software, Resources, Formal analysis. **H.M.K.K.M.B. Herath:** Visualization, Validation, Methodology, Investigation, Data curation, Conceptualization. **Rajkumar Singh Rathore:** Visualization, Validation, Supervision, Project administration, Investigation. **Vishal Krishna Singh:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Weiwei Jiang:** Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Falowo OI, Ozer M, Li C, Abdo JB. Evolving malware and DDoS attacks: Decadal longitudinal study. IEEE Access 2024;12:39221–37.
[2] Yin S, Fu C, Zhao S, Li K, Sun X, Xu T, Chen E. A survey on multimodal large language models. Natl Sci Rev 2024;11(12):nwae403.
[3] Nguyen QH, Wu T, Nguyen V, Yuan X, Xue J, Rudolph C. Utilizing large language models with human feedback integration for generating dedicated warning for phishing emails. In: Proceedings of the 2nd ACM workshop on secure and trustworthy deep learning systems. 2024, p. 35–46.
[4] Fu Z, Acharya S, Ding SH, Zhu Y, Fu J, Xu C. Leveraging human knowledge in large language model for obfuscation-resisted phishing URL detection. In: 2024 ninth international conference on mobile and secure services. IEEE; 2024, p. 1–9.
[5] Gradon KT. Electric sheep on the pastures of disinformation and targeted phishing campaigns: The security implications of chatgpt. IEEE Secur Priv 2023;21(3):58–61.
[6] Al Daoud E, Al Daoud L, Asassfeh M, Al-Shaikh A, Al-Sherideh AS, Afaneh S. Enhancing cybersecurity with transformers: Preventing phishing emails and social media scams. In: 2024 IEEE conference on dependable and secure computing. IEEE; 2024, p. 31–6.
[7] Koide T, Nakano H, Chiba D. ChatPhishDetector: Detecting phishing sites using large language models. IEEE Access 2024.
[8] Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, Min Y, Zhang B, Zhang J, Dong Z, et al. A survey of large language models, 1(2). 2023, arXiv preprint arXiv:2303.18223.
[9] Zhang J, Wu P, London J, Tenney D. Benchmarking and evaluating large language models in phishing detection for small and midsize enterprises: A comprehensive analysis. IEEE Access 2025.
[10] Bethany M, Galiopoulos A, Bethany E, Karkevandi MB, Beebe N, Vishwamitra N, Najafirad P. Lateral phishing with large language models: A large organization comparative study. IEEE Access 2025.
[11] Roy SS, Thota P, Naragam KV, Nilizadeh S. From chatbots to phishbots?: Phishing scam generation in commercial large language models. In: 2024 IEEE symposium on security and privacy. IEEE; 2024, p. 36–54.
[12] Divakaran DM, Peddinti ST. Large language models for cybersecurity: New opportunities. IEEE Secur Priv 2024.
[13] Gupta M, Akiri C, Aryal K, Parker E, Praharaj L. From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. IEEE Access 2023;11:80218–45.
[14] Siemerink A, Jansen S, Labunets K. The dual-edged sword of large language models in phishing. In: Nordic conference on secure IT systems. Springer; 2024, p. 258–79.

[15] Begou N, Vinoy J, Duda A, Korczyński M. Exploring the dark side of ai: Advanced phishing attack design and deployment using chatgpt. In: 2023 IEEE conference on communications and network security. IEEE; 2023, p. 1–6.

[16] Zhang Y, Du T, Ma Y, Wang X, Xie Y, Yang G, Lu Y, Chang E-C. AttacKG+: Boosting attack graph construction with large language models. Comput Secur 2025;150:104220.

[17] Yadav S, Bohra B, et al. A review on recent phishing attacks in internet. In: 2015 international conference on green computing and internet of things. IEEE; 2015, p. 1312–5.

[18] Gupta S, Singhal A, Kapoor A. A literature survey on social engineering attacks: Phishing attack. In: 2016 international conference on computing, communication and automation. IEEE; 2016, p. 537–40.

[19] Buber E, Dırı B, Sahingoz OK. Detecting phishing attacks from URL by using NLP techniques. In: 2017 international conference on computer science and engineering. IEEE; 2017, p. 337–42.

[20] Ahmed J, Tushar Q. COVID-19 pandemic: A new era of cyber security threat and holistic approach to overcome. In: 2020 IEEE Asia-Pacific conference on computer science and data engineering. IEEE; 2020, p. 1–5.

[21] Alkhalil Z, Hewage C, Nawaf L, Khan I. Phishing attacks: A recent comprehensive study and a new anatomy. Front Comput Sci 2021;3:563060.

[22] Abroshan H, Devos J, Poels G, Laermans E. Covid-19 and phishing: Effects of human emotions, behavior, and demographics on the success of phishing attempts during the pandemic. IEEE Access 2021;9:121916–29.

[23] Khode P, Gahane S, Kapse A, Anawade P, Sharma D. Cybersecurity risks and challenges in transition to remote work during the COVID-19 pandemic: A focus on employee behavior and organizational vulnerabilities. In: International conference on smart trends for information technology and computer communications. Springer; 2025, p. 323–33.

[24] Das BC, Amini MH, Wu Y. Security and privacy challenges of large language models: A survey. ACM Comput Surv 2025;57(6):1–39.

[25] Nagarajan S, Kamalbabu K. Analysis of how ChatGPT and gemini help in the generation of cyber attacks. In: 2024 8th international conference on i-SMAC. IEEE; 2024, p. 663–7.

[26] Iturbe E, Llorente-Vazquez O, Rego A, Rios E, Toledo N. Unleashing offensive artificial intelligence: Automated attack technique code generation. Comput Secur 2024;147:104077.

[27] Lee J, Lim P, Hooi B, Divakaran DM. Multimodal large language models for phishing webpage detection and identification. 2024, arXiv preprint arXiv:2408.05941.

[28] Beydemir AB, Sezgin U, Doğan U, Aşıklar BE, Yerlikaya FA, Bahtiyar Ş. A dynamically selected GPT model for phishing detection. In: 2024 14th international conference on advanced computer information technologies. IEEE; 2024, p. 481–4.

[29] Afane K, Wei W, Mao Y, Farooq J, Chen J. Next-generation phishing: How LLM agents empower cyber attackers. In: 2024 IEEE international conference on big data. IEEE; 2024, p. 2558–67.

[30] Zacharis A, Gavrila R, Patsakis C, Douligeris C. Optimising AI models for intelligence extraction in the life cycle of cybersecurity threat landscape generation. J Inf Secur Appl 2025;90:104037.

[31] Hafzullah İ. LLM-driven SAT impact on phishing defense: A cross-sectional analysis. In: 2024 12th international symposium on digital forensics and security. IEEE; 2024, p. 1–5.

[32] Blancaflor EB, Abaleta RM, Achacoso LMD, Amper ACC, Ampiloquio PIR. Emerging threat: The use of AI voice cloning software and services to deceive victims through phone conversations and its potential effects on the filipino population. In: Proceeding of the 2024 5th Asia service sciences and software engineering conference. 2024, p. 137–46.

[33] Chen Y, Cui M, Wang D, Cao Y, Yang P, Jiang B, Lu Z, Liu B. A survey of large language models for cyber threat detection. Comput Secur 2024;145:104016.

[34] Ding W, Abdel-Basset M, Ali AM, Moustafa N. Large language models for cyber resilience: A comprehensive review, challenges, and future perspectives. Appl Soft Comput 2025;170:112663.

[35] Veit MF, Wiese O, Ballreich FL, Volkamer M, Engels D, Mayer P. SoK: The past decade of user deception in emails and today's email clients' susceptibility to phishing techniques. Comput Secur 2025;150:104197.

[36] Hasanov I, Virtanen S, Hakkala A, Isoaho J. Application of large language models in cybersecurity: A systematic literature review. IEEE Access 2024.

[37] Li W, Manickam S, Chong Y-W, Leng W, Nanda P. A state-of-the-art review on phishing website detection techniques. IEEE Access 2024.

[38] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Bmj 2021;372.

[39] Chataut R, Usman Y, Rahman CMA, Gyawali S, Gyawali PK. Enhancing phishing detection with AI: A novel dataset and comprehensive analysis using machine learning and large language models. In: 2024 IEEE 15th annual ubiquitous computing, electronics & mobile communication conference. IEEE; 2024, p. 0226–32.

[40] Zhang D, Jain K, Singh P. Guarding against ChatGPT threats: Identifying and addressing vulnerabilities. In: 2024 IEEE 7th international conference on multimedia information processing and retrieval. IEEE; 2024, p. 612–5.

[41] Dumitras A, Mocan CM, Oprisa C. A feature engineering approach for detecting phishing emails. In: 2024 IEEE 20th international conference on intelligent computer communication and processing. IEEE; 2024, p. 1–8.

[42] Sayyafzadeh S, Weatherspoon M, Yan J, Chi H. Securing against deception: exploring phishing emails through ChatGPT and sentiment analysis. In: 2024 IEEE/aCIS 22nd international conference on software engineering research, management and applications. IEEE; 2024, p. 159–65.

[43] Ling F, Yang H, Xiao Y, Hu L. Meta GPT-based agent for enhanced phishing email detection. In: Proceedings of the 2024 14th international conference on communication and network security. 2024, p. 78–84.

[44] Mahendru S, Pandit T. Securenet: A comparative study of deberta and large language models for phishing detection. In: 2024 IEEE 7th international conference on big data and artificial intelligence. IEEE; 2024, p. 160–9.

[45] Heiding F, Schneier B, Vishwanath A, Bernstein J, Park PS. Devising and detecting phishing emails using large language models. IEEE Access 2024.

[46] Chataut R, Gyawali PK, Usman Y. Can ai keep you safe? a study of large language models for phishing detection. In: 2024 IEEE 14th annual computing and communication workshop and conference. IEEE; 2024, p. 0548–54.

[47] Rashid F, Ranaweera N, Doyle B, Seneviratne S. LLMs are one-shot URL classifiers and explainers. Comput Netw 2025;258:111004.

[48] Chiba D, Nakano H, Koide T. DomainLynx: Advancing LLM techniques for robust domain squatting detection. IEEE Access 2025.

[49] Chang Y-C, A¨ımeur E. Chat or trap? Detecting scams in messaging applications with large language models. In: 2024 8th cyber security in networking conference. IEEE; 2024, p. 92–9.

[50] Trad F, Chehab A. Prompt engineering or fine-tuning? a case study on phishing detection with large language models. Mach Learn Knowl Extr 2024;6(1):367–84.

[51] Al-Sada B, Sadighian A, Oligeri G. MITRE ATT&CK: State of the art and way forward. ACM Comput Surv 2024;57(1):1–37.

[52] Yamin MM, Hashmi E, Ullah M, Katt B. Applications of llms for generating cyber security exercise scenarios. IEEE Access 2024.

[53] Fairbanks J, Serra E. Generating phishing attacks and novel detection algorithms in the era of large language models. In: 2024 IEEE international conference on big data. IEEE; 2024, p. 2314–9.

[54] Schesny M, Lutz N, J¨agle T, Gerschner F, Klaiber M, Theissler A. Enhancing website fraud detection: A ChatGPT-based approach to phishing detection. In: 2024 IEEE 48th annual computers, software, and applications conference. IEEE; 2024, p. 1494–5.

[55] Gressel G, Pankajakshan R, Mirsky Y. Discussion paper: Exploiting llms for scam automation: A looming threat. In: Proceedings of the 3rd ACM workshop on the security implications of deepfakes and cheapfakes. 2024, p. 20–4.