IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

Generative AI–Enabled Phishing: Emerging Attack Capabilities, Detection Limitations, and Research Directions for Adaptive Cybersecurity

Pallvi Mahajan
University of Jammu
Pallvimahjanlaw1@gmail.com

Abstract: *Generative artificial intelligence (AI) is changing how phishing messages are written and adapted. Large language models can produce clear, convincing, and varied messages, making some traditional warning signs less reliable. These tools may also help security teams examine suspicious emails and support decisions. This paper reviews research on AI-enabled phishing, focusing on phishing emails, message generation, detection methods, and limits in current evaluation practices. It compares approaches that use message content, context, machine learning, and large language models to identify phishing attempts. The review finds that strong results on one dataset do not prove that a method will work against new campaigns or in different settings. Data quality, false alarms, missed attacks, privacy, cost, and clear explanations also matter. The paper identifies research gaps and suggests future priorities, including testing across time and sources, combining useful signals, and updating detection systems carefully. The proposed framework needs further testing and validation.*

Keywords: Generative artificial intelligence; phishing; large language models; cybersecurity; adversarial robustness; phishing detection; adaptive defence.

I. INTRODUCTION

Phishing remains a serious cybersecurity problem because it exploits decisions made during routine communication. A message may be trusted because it resembles a familiar organisation, refers to an expected transaction, or appears at a moment when a recipient is likely to respond quickly. The attack therefore depends on more than the technical design of a link or website; it also relies on context, trust, timing, and the recipient's interpretation of the request. Generative artificial intelligence (GenAI) adds a further dimension to this problem. Large language models can produce fluent messages, vary wording at scale, and help tailor communications to a particular role or situation. These capabilities do not make every phishing attempt sophisticated, nor do they remove the need for attacker knowledge and infrastructure. They do, however, reduce some of the effort traditionally associated with producing plausible deceptive content and may make campaigns less uniform.

The implications for defence are not one-sided. Language models and transformer-based classifiers can help analyse semantic intent, identify contextual inconsistencies, and support security analysts in prioritising alerts. Yet their performance is shaped by the data on which they are trained, the signals available at decision time, and the way evaluation is conducted. A detector that performs well on a familiar benchmark may be less reliable when confronted with a new campaign, a different language, a compromised legitimate account, or a message whose wording has been deliberately varied. The practical question is consequently broader than model accuracy: it concerns the reliability of an entire detection process under changing conditions. This distinction matters for organisations deciding whether, where, and how to deploy AI-assisted security controls.

This article addresses four questions: how generative AI changes the production and presentation of phishing; what limitations affect current phishing-email detection approaches; whether existing evaluation practices establish

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-40759

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 2, November 2026 Impact Factor: 8.2

generalisation to changing campaigns; and which research priorities could improve the reliability of adaptive defence. The article's contribution is a critical synthesis organised around the relationship between attack evolution, available detection evidence, evaluation validity, and operational consequences. It develops a conceptual research agenda for testing phishing defences across time, sources, and modalities. It does not introduce a new classifier, report original experiments, or claim empirical validation of the proposed framework.

II. RESEARCH METHODOLOGY

2.1 Research design

This study adopts a literature-based comparative analytical design to examine generative-AI-enabled phishing and the strengths and limitations of contemporary detection approaches. The analysis draws on published review articles, empirical studies, and methodological work addressing phishing generation, phishing-email detection, contextual language-model systems, explainability, dataset quality, and mitigation. The purpose is not to estimate a pooled effect or to rank every available detector, but to compare the types of claims made in the literature and the conditions under which those claims are supported.

2.2 Analytical framework

The literature is examined through four linked dimensions. The first concerns attack capability: fluency, variation, personalisation, impersonation, and movement across communication channels. The second concerns detection evidence: message semantics, sender and domain reputation, URL and webpage characteristics, contextual information, and multimodal signals. The third concerns evaluation validity: dataset provenance, class balance, deduplication, baselines, train–test separation, temporal testing, and performance on external sources. The fourth concerns operational suitability, including false-positive and false-negative costs, calibration, explainability, privacy, latency, computational expense, and human oversight. These dimensions enable comparison without assuming that studies using different tasks or datasets are directly equivalent.

A reported accuracy score cannot, by itself, establish operational reliability. Studies may differ in the definition of phishing, data collection period, class distribution, feature availability, labelling procedures, and evaluation design. A random split may also place related messages or campaign artefacts in both training and test sets, creating an optimistic estimate of generalisation. Accordingly, numerical results are interpreted within the conditions reported by each study; they are not treated as universal rankings. The analysis distinguishes experimental findings from conceptual proposals and avoids inferring real-world victimisation from a model's ability to generate or classify text.

Sources were selected for their direct relevance to the article's research questions and for their contribution to understanding phishing capability, detection, generalisation, or mitigation. The approach is a selective narrative review, not a systematic review: it does not claim exhaustive database coverage, a fully reproducible screening protocol, or statistical meta-analysis. The discussion therefore presents a critical synthesis of representative work rather than a complete census of the field. This scope is appropriate for developing an analytical research agenda, but it also means that relevant studies may have been omitted and that the conclusions should be read as interpretive rather than exhaustive.

2.3 Phishing-Focused Synthesis of the Literature

The literature addresses three connected problems: how phishing messages are constructed to appear credible, how recipients interpret those messages, and how technical systems distinguish malicious communication from legitimate activity. Schmitt and Flechais discuss the capacity of generative AI to amplify social engineering by making plausible content easier to produce and adapt. Jabir, Le and Nguyen emphasise the human factors that shape susceptibility to phishing, including trust, urgency, and the interpretation of context. Together, these perspectives show why phishing should not be reduced to poor grammar or suspicious keywords. Risk emerges from the relationship between message content, apparent sender identity, requested action, delivery channel, and the recipient's expectations.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

Research on large language models also complicates the distinction between attack generation and defence. LLMs can generate fluent and varied messages, potentially reducing the usefulness of older awareness cues such as obvious spelling errors or repeated templates. The same class of systems may assist with semantic interpretation, contextual triage, and explanation of suspicious features. Neither capability establishes that every AI-generated message is more persuasive or that LLM-assisted detection is necessarily superior. Such claims depend on the threat model, the data used, the availability of contextual signals, and the way success is measured. A defensible analysis must therefore distinguish the capability to generate deceptive content from evidence that the content increases attack success, and distinguish a promising detector from one shown to work reliably in deployment.

Research on phishing-email detection examines message content, sender information, contextual signals, and patterns associated with malicious requests. Traditional filters and reputation systems can identify known threats, while machine-learning and transformer-based models may detect less familiar patterns in message text. Context-augmented language models can also assess whether a request fits an expected relationship or organisational process. These approaches rely on different evidence and should be compared only when their tasks, inputs, and evaluation conditions are clear. This paper therefore concentrates on phishing emails and the detection of deceptive communication, rather than treating website classification as a separate literature stream.

A further concern is the extent to which benchmark results transfer to unseen campaigns. Performance can be affected by collection practices, duplicated examples, source-specific patterns, class imbalance, and the separation of training and test data. Research on domain adaptation and dataset quality indicates why temporal and cross-source evaluation is important: attackers can vary wording, infrastructure, sender identities, and delivery channels, while legitimate communication also changes over time. A detector that recognises familiar artefacts may therefore perform less reliably when deployed against a new campaign or a different organisation. High performance on one dataset is useful evidence about that dataset, but it is not sufficient evidence of broad operational robustness.

The practical consequences of detection errors should also shape evaluation. False negatives may permit credential theft, fraudulent payments, or unauthorised access; false positives can interrupt legitimate communication and encourage users to disregard warnings. Accuracy alone does not capture this trade-off. Where data permit, studies should report precision, recall, F1 score, calibration, false-positive burden, and performance across time periods, sources, languages, and phishing categories. Explanations should be assessed for fidelity and usefulness to analysts rather than accepted merely because they sound plausible. Privacy, latency, cost, and escalation to human review are likewise part of the security value of a detection system.

This synthesis identifies the central issue addressed by the paper: whether phishing defences remain dependable as generative tools make deceptive communication more fluent and variable and as attack patterns evolve. Existing studies contribute evidence about individual detection methods, LLM applications, and mitigation controls, but methodological differences limit direct comparisons. The analysis that follows therefore connects attack capability with detection evidence and evaluation quality. On this basis, the paper develops a research agenda centred on temporal and cross-source testing, complementary signals, calibrated uncertainty, explainable escalation, and governed adaptation. These priorities are proposed for future validation, not presented as the results of an experiment conducted for this article.

III. LITERATURE REVIEW

Schmitt and Flechais (2024) examine the implications of generative AI for social engineering and phishing. Their analysis presents generative systems as tools that can make deceptive communication more persuasive and easier to adapt, while recognising that the same technologies may support defensive analysis. The important contribution is the emphasis on amplification: GenAI can reduce the effort needed to produce plausible content, but the resulting threat still depends on the credibility of the scenario, the information available to the attacker, and the surrounding delivery infrastructure. This distinction cautions against treating language generation as a complete explanation for phishing success.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

Jabir, Le and Nguyen (2025) examine phishing in the GenAI era through the lens of human factors. Their review draws attention to attention, trust, urgency, and contextual interpretation in recipients' responses. This perspective is important because technical sophistication does not determine whether a message succeeds on its own. A message can be linguistically polished yet fail if the request is implausible, while a relatively simple message may succeed when it arrives during a credible workflow. The implication for detection is that systems should consider the relationship between the message, sender, request, and communication context instead of treating language quality as a sufficient proxy for risk.

3.1. Generative AI, LLMs, and Human Factors

Sivaneswaran and colleagues (2026) review the use of large language models in phishing generation and detection. Their synthesis notes that the evidence base often relies on manually assembled datasets and that evaluation practices vary across studies. These limitations make it difficult to infer how well a detector will perform beyond the data used to develop it. The review is particularly relevant because it considers both offensive and defensive applications, highlighting a continuing tension: techniques that improve content generation may also alter the characteristics on which detection systems have been trained. Progress in this area therefore depends not only on new models, but also on stronger and more transparent evaluation practices.

Popescul and Radu (2025) use bibliometric analysis to examine research on AI-based phishing detection. This approach helps identify influential topics, publication patterns, and areas of growing interest. Its value, however, lies in mapping the research landscape rather than testing the real-world effectiveness of individual methods. A topic's visibility or citation frequency should not be confused with operational readiness. For the present analysis, the study is useful for locating the field's main research strands, while performance claims must be assessed against empirical evaluation and deployment conditions.

3.2. Phishing-Email Detection Approaches

Thakur, Ali, Obaidat and Kamruzzaman (2023) survey deep-learning approaches to phishing email detection. Neural models can capture contextual and sequential features that are difficult to represent through manually specified rules. Their performance, however, depends on factors such as label quality, text preprocessing, the balance of classes, and the relationship between training and test data. These considerations are especially important when detection is extended to GenAI-generated content, because changes in writing style may not correspond to changes in malicious intent. The review supports the use of learned representations, but also reinforces the need for careful testing beyond a single benchmark.

Naqvi and colleagues (2023) review mitigation strategies against phishing, considering organisational and technical measures rather than detection algorithms alone. This broader scope is useful because a classifier cannot prevent every successful attack. Sender authentication, phishing-resistant multifactor authentication, staff reporting, transaction verification, access controls, and incident response can reduce harm even when a suspicious message reaches the inbox. The review supports a defence-in-depth approach. It does not, however, settle how these controls should be integrated with GenAI-specific detection or how their combined effectiveness should be measured under changing attack conditions.

3.3. Generalisation, Context, and Transformer-Based Detection

Rashid and colleagues (2024) investigate generalisation in phishing URL detection using unsupervised domain adaptation. Their work challenges the assumption that a random train–test split provides sufficient evidence of generalisation. When training and test examples share a similar distribution, performance may appear strong even though a change in source or domain weakens the model. This concern is directly relevant to adaptive phishing, where URLs, hosting patterns, and campaign conventions can change. Domain adaptation may help, but its practical value

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

depends on the availability and quality of target-domain signals and on whether the evaluation reflects the conditions in which the system will be used.

Nicklas, Ventulett and Conrad (2024) evaluate context-augmented open LLMs for phishing-email detection using few-shot learning and retrieval-augmented generation. Their findings suggest that relevant contextual information can improve recognition, including when smaller models are used. The result is promising because the meaning of a message often depends on whether the request fits an established relationship or business process. Context is not automatically trustworthy, however: retrieved material may be outdated, incomplete, or compromised, and a model may still produce an incorrect decision with a persuasive explanation. Further evaluation should therefore examine context provenance, retrieval quality, error rates, and the effect of missing or conflicting evidence.

Jamal, Wimmer and Sarker (2024) propose a transformer-based model for distinguishing phishing, spam, and legitimate emails, reporting improvements over baseline methods across the datasets they examine. Their work illustrates how contextual representations can support email classification. At the same time, results depend on the dataset, preprocessing decisions, class distribution, and the choice of baseline. The next step is not simply to reproduce a high score on a similar corpus, but to test whether the model remains useful on later campaigns, adversarially varied messages, and legitimate communications that resemble common phishing patterns. False-positive rates are particularly important where automated blocking could interrupt ordinary business.

Patel, Rehman and Iqbal (2024) compare multiple large language models on a collection of scam emails and report meaningful variation in detection performance. The study is valuable because it evaluates models on a defined detection task rather than assuming that general language ability translates into security competence. Its conclusions nevertheless remain bounded by the composition of the corpus and the tested prompts and models. Scam emails are not interchangeable with all forms of phishing, and performance on one collection cannot establish reliability across languages, channels, or later campaigns. Independent evaluation, transparent prompts, and explicit reporting of false positives and false negatives are needed before such results can inform deployment decisions.

Opara, Modesti and Golightly (2025) assess spam filters and stylometric methods against AI-generated phishing emails. They report that an XGBoost model using stylometric features achieved high accuracy in their experimental setting and make a dataset available for further research. This provides evidence that writing patterns may contribute useful signals, but it does not establish stylometry as a durable standalone solution. Such features can change as generators evolve or when generated text is edited by a person. Accuracy also gives an incomplete account of deployment value unless it is considered alongside class prevalence, false-positive burden, temporal testing, and performance on hybrid messages.

3.4. Explainability, Dataset Quality, and Research Gaps

Al-Subaiey and colleagues (2024) present an interpretable web-based approach to phishing email detection and report a high F1 score. Combining classification with an explanation addresses a practical need: security analysts must understand why a communication has been flagged, especially when decisions affect legitimate users. Yet a high score is meaningful only in relation to the dataset and evaluation design, and an explanation is not necessarily a faithful account of the model's reasoning. Further assessment should examine explanation fidelity, calibration, stability across model versions, and whether analysts can use the explanation to make better decisions.[16]

Wilk-Jakubowski and colleagues (2025) review machine-learning and neural-network approaches to phishing detection and identify recurring concerns about dataset quality and reporting. These include incomplete descriptions of data acquisition, cleaning, deduplication, and train–test separation. Such omissions make results difficult to reproduce and can inflate apparent performance when near-duplicate or related samples occur across training and test sets. Their analysis reinforces a basic methodological point: benchmark quality is part of the security evidence, not a secondary technical detail. Models cannot be compared reliably unless researchers describe how their data were obtained and how leakage was prevented.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

IV. EMERGING CAPABILITIES OF GENERATIVE AI–ENABLED PHISHING

The reviewed literature points to several connected changes in phishing capability. First, GenAI can reduce the linguistic effort required to produce coherent messages in different tones and languages. This weakens awareness advice that treats spelling mistakes, awkward grammar, or generic phrasing as reliable warning signs. Such cues may still be useful, but their absence does not make a message safe, and polished writing does not establish that AI was used. Language quality should therefore be considered alongside sender identity, the requested action, links, and the context in which the message appears.[3,4]

Second, generative systems can support rapid variation and personalisation. A message can be adapted to a recipient's role, a plausible business process, or an earlier interaction, making a campaign less dependent on a single repeated template. The principal change is not that every attack becomes highly sophisticated; it is that attackers may be able to test more variations at lower cost. Detection systems that depend on fixed signatures or narrow textual patterns may consequently face a moving-target problem. Third, synthetic audio and video can extend impersonation beyond written messages. The degree of risk depends on the quality of the media, access to contextual information, and the opportunity for independent verification. These capabilities strengthen the case for identity and transaction controls that do not rely solely on judging how convincing a communication sounds.

Image 1 presents a conceptual lifecycle for a GenAI-enabled phishing attempt; it should not be read as a universal or inevitable sequence. An attack may begin with publicly available context, move through message or media generation, and reach a recipient through one or more communication channels before a credential, payment, or other security objective is pursued. Controls can interrupt this process at several points: limiting access to sensitive information, authenticating senders, checking URLs and contextual signals, independently verifying identity, and confirming high-risk transactions through a trusted channel. A defence based only on textual classification may miss attacks that rely on phone calls, collaboration platforms, QR codes, or compromised accounts.

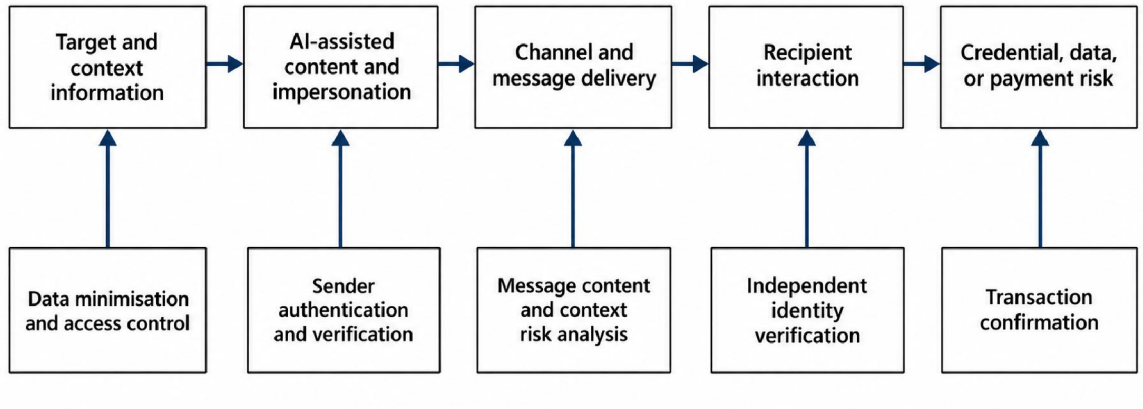


Image 1. Conceptual lifecycle of generative AI-enabled phishing.

V. DETECTION LIMITATIONS AND EVALUATION CHALLENGES

The most consequential limitation is often the validity of the evaluation rather than the absence of sophisticated algorithms. High accuracy on a randomly divided dataset can give a misleading impression when training and test samples share templates, domains, collection periods, or near-duplicates. Research on domain adaptation and dataset quality shows why performance should be tested under conditions that better approximate future use.[11,17] At a minimum, studies should document data provenance, remove duplicates, separate training from later time periods, and evaluate on an external source where possible. These practices cannot eliminate all bias, but they make claims about generalisation more credible.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

Metric choice also affects interpretation. Accuracy may conceal weak recall when phishing is relatively uncommon, or an excessive false-positive rate that makes a system difficult to use. Precision, recall, F1, receiver-operating-characteristic measures, calibration, and workload-sensitive outcomes should be reported alongside class prevalence and decision thresholds. False negatives can expose users to compromise; false positives can interrupt legitimate communication and erode trust in the security system. The acceptable trade-off will differ between consumer email, financial services, public institutions, and internal enterprise communication. Evaluation should therefore make the operational costs of errors explicit rather than presenting a single score as a complete measure of success.

LLM-based detection brings additional practical and methodological concerns. Outputs may change with prompt wording, model version, context-window limits, and the evidence supplied to the system. A model can produce a confident but incorrect explanation, while retrieval may introduce outdated or malicious material. Model access, inference cost, response latency, data retention, and confidentiality also affect whether an approach is suitable for deployment. Context augmentation and explainable outputs are promising, but they should be assessed as parts of a complete decision process, including what happens when the model is uncertain, unavailable, or contradicted by other signals.

Testing robustness requires more than asking a model to classify a few paraphrased messages. Evaluations should include campaigns not seen during training, multilingual content, compromised legitimate accounts, redirected URLs, messages containing both legitimate and suspicious context, and attacks that move between communication channels. Tests should be designed and reported in ways that support defensive research without disclosing unnecessary operational guidance for abuse. Performance should be broken down by attack type and time period, and uncertainty intervals should be reported where the data permit. The relevant question is whether a detector continues to support sound decisions when the environment changes, not whether it recognises examples similar to those on which it was trained.

Detection of malicious intent must also be distinguished from detection of AI authorship. The two tasks overlap only imperfectly. A human-written message may be a phishing attempt, while AI-generated text may be a legitimate customer-service response. A detector trained to identify stylistic features of generated text may fail when an attacker edits that text or combines human and machine-written material. A security system should therefore prioritise the risk presented by the communication, drawing on evidence such as sender identity, destination, requested action, message context, and known indicators of compromise. Authorship detection may provide supplementary information, but it should not substitute for risk assessment.

Evaluation should specify the point at which each decision is made and which signals are available at that moment. A detector that uses a rendered webpage, a domain-reputation service, or later account behaviour has access to evidence that may not be available when an email first arrives. Comparing such a system with a text-only classifier without describing signal availability can obscure the reasons for different results. Studies should report feature-acquisition time, dependencies on external services, latency, and performance when some signals are missing. These details are necessary to judge whether a proposed system can function within a real mail gateway or incident-response workflow.

VI. RESEARCH GAP AND PROPOSED ADAPTIVE DEFENCE RESEARCH AGENDA

The gap identified here is not the absence of phishing classifiers or the complete ineffectiveness of existing methods. Rather, the literature does not consistently establish whether reported performance transfers across time, sources, languages, and changing attack strategies. In addition, studies may evaluate different inputs and decision stages, making direct comparison difficult. The resulting research problem is methodological as well as technical: how should a phishing defence be evaluated so that its apparent performance reflects the conditions under which it would actually be used? The paper's contribution is to organise this problem around four linked dimensions—attack evolution, detection evidence, evaluation validity, and operational consequences—and to derive a set of testable priorities from that synthesis.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 2, November 2026 Impact Factor: 8.2

Image 2 presents a conceptual adaptive-defence framework intended to guide future research; it is not a validated architecture or a claim of measured improvement. The framework links privacy-governed signal collection, complementary detection features, risk scoring, explanations, human review, verified incident feedback, adversarial evaluation, and controlled model updates. Its value lies in making the relationships among these components explicit and identifying where empirical tests are needed. Each component should be evaluated against suitable baselines and under stated threat assumptions before the framework is treated as operationally effective.

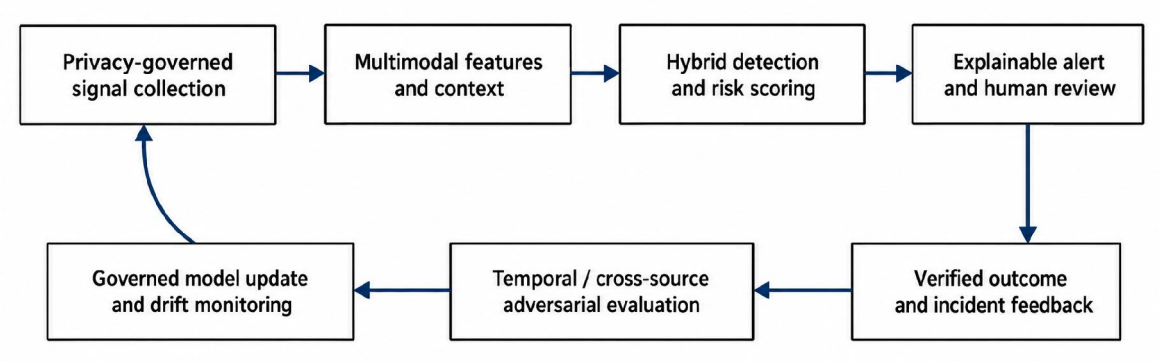


Image 2. Proposed adaptive phishing-defence research framework.

Adaptation must be governed carefully. Incident reports can improve detection, but they may be incomplete, delayed, inconsistent, or deliberately poisoned. Model updates should therefore require provenance checks, access controls, audit trails, rollback procedures, and periodic independent evaluation. Privacy-preserving techniques and local processing may reduce unnecessary exposure of message content, although their benefits should be measured against detection performance and implementation cost. In this framework, an adaptive system is not simply one that updates frequently. It is one whose response to change can be measured, audited, and reversed when an update causes harm.

A practical evaluation matrix could combine three dimensions: time (random split versus future-period testing), source (same-source versus external-source testing), and modality (text-only versus combined contextual signals). Results should be reported separately for known and previously unseen campaigns. Such a matrix would not remove every source of bias, but it would make claims about generalisation more testable and would help identify the conditions under which a method fails. Researchers should also compare complex models with transparent baselines, including simple rules and reputation systems. Without these comparisons, improvements may be attributed to model sophistication when they arise from data leakage, feature availability, or an unusually favourable test set.

Model updating should be risk-based rather than continuous by default. Frequent updates may help a detector respond to new campaigns, but they can also introduce regressions, drift, and feedback loops. A controlled update process should use verified examples, a hold-out evaluation set, checks across relevant languages and user groups, a documented approval decision, and a clear rollback path. Where incident labels remain uncertain, records should be treated as provisional rather than immediately incorporated as ground truth. The effectiveness of the framework should ultimately be assessed through longitudinal testing and operational outcomes, not inferred from its conceptual design alone.

VII. DISCUSSION AND IMPLICATIONS

The reviewed evidence does not support a simple conclusion that LLMs are either inherently superior detectors or an uncontrollable source of phishing. Their effects depend on the task, the information available, and the surrounding security controls. GenAI can make deceptive messages more fluent and variable, while the same family of technologies can assist semantic analysis and contextual triage. A model may be useful for prioritising messages yet unsuitable as the sole authority for blocking them. A more defensible approach is layered: automated systems identify and prioritise

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

risk, complementary controls verify identity and restrict harmful actions, and human reviewers handle uncertain or high-impact cases.

For organisations, high-risk requests involving payments, credential resets, account recovery, or confidential information should be verified independently of the message that initiated them. Email authentication, phishing-resistant multifactor authentication, least-privilege access, clear reporting channels, and timely incident response remain important even where AI-based detection is available. In India, these principles are relevant to digital-payment services and enterprise communication, but general cybercrime complaint figures should not be treated as direct measures of phishing prevalence. Claims about national trends require sources that specifically classify phishing incidents and explain the relevant reporting and measurement methods.

For researchers, progress depends on transparent methods as much as on new architectures. Studies should describe dataset provenance, define the threat model, separate development data from later test periods, report false-positive and false-negative costs, and provide reproducible protocols where doing so does not create additional security risks. Comparative claims should be based on robust decision quality across meaningful test conditions, not benchmark accuracy alone. These practices would make it easier to distinguish genuine methodological progress from improvements that depend on a particular dataset or evaluation setup.

The framework also raises questions of research governance. Email and messaging datasets may contain personal information, financial details, or commercially sensitive material. Additional context can improve detection while increasing privacy risks. Data minimisation, restricted access, clear retention periods, redaction, and secure evaluation environments should therefore be built into study design. Researchers should document the lawful basis or consent arrangements where applicable and explain how sensitive records are protected. The objective is not to collect the maximum possible context, but to identify which signals add demonstrable security value and can be used proportionately.

VIII. LIMITATIONS

This article is a selective critical review, not a systematic review or meta-analysis. The source selection was based on relevance to the paper's research questions and on publicly discoverable scholarly and publisher records; exhaustive coverage of all databases and venues was not attempted. The synthesis is therefore vulnerable to selection bias and may omit relevant studies. Because the article does not conduct original experiments, its proposed adaptive-defence framework remains conceptual. Reported performance values are discussed as findings from the cited studies rather than as independently replicated results. These limitations constrain the breadth of the claims but do not remove the value of a critical comparison of evaluation practices and research priorities.

IX. CONCLUSION

Generative AI may change phishing by reducing the effort required to produce fluent, varied, and context-sensitive deceptive communications, while multimodal systems extend the possibilities for impersonation. These developments also create opportunities for context-aware detection and analyst support. The literature reviewed here does not establish that model sophistication or a high score on a single dataset is sufficient to demonstrate operational reliability. Dataset quality, temporal and cross-source generalisation, error calibration, explainability, privacy, and deployment costs remain central to assessing whether a detection method is useful in practice.

Phishing defence should be evaluated as a continuing socio-technical capability rather than as a one-time model-selection exercise. Future research should prioritise reproducible time-separated benchmarks, external-source testing, transparent reporting of false-positive and false-negative costs, complementary technical signals, uncertainty-aware escalation, and controlled feedback-driven updates. These priorities should be tested against realistic threat models and established security controls. The proposed agenda is intended to make claims about adaptive phishing defence more comparable and falsifiable; its effectiveness must be established through subsequent empirical evaluation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

REFERENCES

1. Marc Schmitt & Ivan Flechais, Digital Deception: Generative Artificial Intelligence in Social Engineering and Phishing, 57 Artificial Intelligence Review 324 (2024), https://doi.org/10.1007/s10462-024-10973-2.
2. Raja Jabir, John Le & Chau Nguyen, Phishing Attacks in the Age of Generative Artificial Intelligence: A Systematic Review of Human Factors, 6(8) AI 174 (2025), https://doi.org/10.3390/ai6080174.
3. Dinushan Sivaneswaran et al., A Systematic Literature Review of Large Language Models in Phishing Attack Generation and Detection, 30 Array 100775 (2026), https://doi.org/10.1016/j.array.2026.100775.
4. Daniela Popescul & Laura Diana Radu, AI in Phishing Detection: A Bibliometric Review, 8 Frontiers in Artificial Intelligence (2025), https://doi.org/10.3389/frai.2025.1496580.
7. A Systematic Review on Deep-Learning-Based Phishing Email Detection, 12(21) Electronics 4545 (2023), https://doi.org/10.3390/electronics12214545.
8. Bilal Naqvi et al., Mitigation Strategies Against the Phishing Attacks: A Systematic Literature Review, 132 Computers & Security 103387 (2023), https://doi.org/10.1016/j.cose.2023.103387.
9. Fariza Rashid, Ben Doyle, Soyeon Caren Han & Suranga Seneviratne, Phishing URL Detection Generalisation Using Unsupervised Domain Adaptation, 245 Computer Networks 110398 (2024), https://doi.org/10.1016/j.comnet.2024.110398.
10. Fabian Nicklas, Nicolas Ventulett & Jan Conrad, Enhancing Phishing Email Detection with Context-Augmented Open Large Language Models, Proceedings of the Upper-Rhine Artificial Intelligence Symposium (2024), https://doi.org/10.60643/urai.v2024p159.
11. Suhaima Jamal, Hayden Wimmer & Iqbal H. Sarker, An Improved Transformer-Based Model for Detecting Phishing, Spam and Ham Emails: A Large Language Model Approach, 7(5) Security and Privacy e402 (2024), https://doi.org/10.1002/spy2.402.
12. Het Patel, Umair Rehman & Farkhund Iqbal, Evaluating the Efficacy of Large Language Models in Identifying Phishing Attempts, 2024 16th International Conference on Human System Interaction (HSI) (2024), https://doi.org/10.1109/HSI61632.2024.10613528.
13. Chidimma Opara, Paolo Modesti & Lewis Golightly, Evaluating Spam Filters and Stylometric Detection of AI-Generated Phishing Emails, 276 Expert Systems with Applications 127044 (2025), https://doi.org/10.1016/j.eswa.2025.127044.
14. Novel Interpretable and Robust Web-Based AI Platform for Phishing Email Detection, 120 Computers and Electrical Engineering 109625 (2024), https://doi.org/10.1016/j.compeleceng.2024.109625.
15. Machine Learning and Neural Networks for Phishing Detection: A Systematic Review (2017–2024), 14(18) Electronics 3744 (2025), https://www.mdpi.com/2079-9292/14/18/3744.
17. Takashi Koide et al., ChatSpamDetector: Leveraging Large Language Models for Effective Phishing Email Detection, SecureComm 297–319 (2024), https://doi.org/10.1007/978-3-031-94455-0_14.
18. A Feature-Engineered Dataset of Benign and Phishing URLs for Machine Learning and Large Language Models Evaluation, 63 Data in Brief 112162 (2025), https://doi.org/10.1016/j.dib.2025.112162.