

# A Hybrid Deep-Learning Framework for Detecting AI-Generated Phishing Attacks in Cloud-Based Communication Systems

**Santanu Kumar Jana**

Assistant Professor, Tamralipta Institute of Management and Technology Teaching and administrative experience: 12 years

## Abstract

Generative artificial intelligence has altered the operational character of phishing. Large language models can produce fluent, context-sensitive and highly personalised messages that resemble routine organisational correspondence, weakening filters that depend heavily on spelling errors, suspicious keywords or previously observed templates. This conceptual study develops a hybrid deep-learning framework for identifying both conventional and AI-assisted phishing in cloud-based communication systems. Following a design-science approach, the paper synthesises current standards, threat reports and recent research on transformer-based classification, adversarially modified phishing and explainable artificial intelligence. The proposed architecture combines six evidence streams: contextual semantics, stylometry, URL and domain properties, authentication headers, sender behaviour and attachment risk. An attention-based fusion layer produces a calibrated risk estimate, while an explanation layer communicates the principal factors behind each decision. The paper also specifies a reproducible validation protocol based on temporally ordered data, class-aware metrics, adversarial perturbations, ablation studies and paired statistical tests. The framework does not assume that AI-generated text is inherently malicious; instead, it assesses deceptive intent, technical inconsistency and contextual abnormality. The study argues that robust protection requires multi-source evidence, continuous monitoring, human oversight and privacy-sensitive deployment. Because no prototype is implemented in the present work, claimed benefits are framed as testable design propositions rather than empirical performance results.

**Keywords:** artificial intelligence; phishing detection; cloud security; transformers; behavioural analytics; explainable AI; adversarial robustness

## 1. Introduction

Cloud email, collaboration suites, shared-document platforms and enterprise messaging have become basic organisational infrastructure. Their convenience also expands the channels through which social-engineering attacks can reach employees. Phishing succeeds by persuading a recipient to disclose credentials, approve a transaction, open a malicious file or visit a fraudulent service. In cloud environments, the apparent legitimacy of the delivery channel can be misleading: a message may originate from a compromised account, a reputable hosting domain or a familiar collaboration platform. Earlier campaigns often contained conspicuous errors, generic salutations and repeated templates. These

signals supported rules, blocklists and conventional spam filters. Generative AI changes the balance. It can produce polished messages in multiple languages, adapt tone to a target, imitate institutional conventions and rapidly paraphrase the same malicious request. ENISA (2024, 2025) identifies phishing and social engineering as persistent initiation vectors and notes that generative AI can improve the quality and scale of malicious content. NIST (n.d.) similarly warns that increasingly convincing phishing messages require layered safeguards.

The central difficulty is therefore not simply to identify machine-written prose. Legitimate users routinely employ writing assistants, while human attackers can write fluent messages. The security task is to determine whether a communication is deceptive or harmful. That determination requires the joint interpretation of message meaning, link and domain evidence, authentication results, sender-recipient relationships, account behaviour and attachment properties.

### **1.1 Problem statement**

Many published detectors are trained and tested on historical email corpora using random splits. Such evaluations can reward memorisation of recurring campaigns and may overstate performance against future, paraphrased or previously unseen attacks. Content-only transformers improve semantic representation, but a carefully written payment request may remain indistinguishable from legitimate business correspondence unless the system also observes that the reply-to domain has changed, the sender-recipient relationship is new or the requested action conflicts with normal practice. Conversely, an urgency word alone may create unnecessary alerts. The problem is to design a detector that combines heterogeneous evidence, remains interpretable to reviewers and can be evaluated under realistic temporal and adversarial change.

### **1.2 Contribution of the paper**

- A threat-oriented feature taxonomy for AI-assisted phishing in cloud communication.
- A six-module hybrid architecture with attention-based evidence fusion and calibrated risk output.
- A validation plan that separates design propositions from future empirical findings.
- Operational, privacy and explainability safeguards suitable for institutional deployment.

## **2. Background and Threat Context**

### **2.1 From bulk phishing to adaptive social engineering**

Bulk phishing distributes nearly identical messages to large populations. Spear phishing narrows the target and incorporates personal or organisational details, while business email compromise imitates executives, employees, suppliers or professional advisers. Generative models lower the cost of moving from the first pattern to the second: attackers can create numerous stylistic variants, translate them, sustain conversations and tailor requests to public information. Synthetic voice, images and video may further reinforce a false identity, although the present framework focuses on textual communication and its associated metadata.

### **2.2 Operational definition**

In this paper, AI-generated phishing means a deceptive communication whose wording, structure, translation, personalisation or conversational response has been generated or materially revised with a generative model. This definition concerns attack construction, not the detector's target label. The target remains phishing risk. AI authorship is treated as uncertain contextual evidence because reliable authorship attribution is difficult and because machine assistance is common in legitimate writing.

### 2.3 Cloud-specific risks

| Cloud condition          | Security implication                                                       |
|--------------------------|----------------------------------------------------------------------------|
| Account compromise       | A malicious message can pass sender-address and authentication checks.     |
| Shared cloud links       | Reputable hosting domains may conceal fraudulent or weaponised content.    |
| Distributed access       | New devices, locations and work patterns complicate behavioural baselines. |
| Multiple channels        | Email, chat and shared documents fragment security visibility.             |
| High message volume      | Manual verification cannot serve as the primary control.                   |
| Third-party integrations | Tokens, bots and applications create additional trust relationships.       |

### 2.4 Threat assumptions

The proposed system assumes that attackers may control message wording, display names, embedded links and attachments; may obtain public information about a target; and may occasionally use a compromised legitimate account. It does not assume unrestricted access to protected historical communications. Behavioural features must be derived under organisational authority, data-minimisation rules and access controls. The system is a decision-support component, not a substitute for multifactor authentication, secure payment verification, domain protection, user education or incident response.

### 2.5 Design requirements

- Detect malicious intent even when language is fluent and professionally formatted.
- Use independent technical and contextual evidence when text is ambiguous.
- Remain robust to paraphrasing, obfuscation and previously unseen campaigns.
- Provide concise reasons and calibrated uncertainty for human review.
- Limit collection, retention and exposure of personal communication data.

## 3. Review of Related Literature

### 3.1 Conventional machine learning and deep learning

Earlier email and URL detectors commonly employed logistic regression, naïve Bayes, support-vector machines, decision trees, random forests and gradient boosting. Typical features included token frequency, URL length, domain properties, HTML structure and header fields. These models are computationally economical and can perform well when training and deployment data are similar. Their dependence on engineered or historically stable features, however, makes them vulnerable to vocabulary shifts, new brands and changing delivery patterns.

Convolutional and recurrent neural networks reduced reliance on manual feature design by learning representations from sequences. Transformers advanced this approach through contextual attention. BERT and related models represent a token in relation to its surrounding words, enabling better recognition of requests, impersonation and persuasive strategies (Devlin et al., 2019; Vaswani et al., 2017). Recent phishing studies report strong results for transformer-based classifiers, but performance estimates remain tied to particular datasets and evaluation conditions.

### 3.2 Large language models and explanation

Koide et al. (2024) investigated large language models for phishing analysis and generated reasons for

model decisions. Uddin, Mahiuddin and Sarker (2026) reported an explainable transformer-based detector and combined attribution methods to improve interpretability. These studies show the value of contextual reasoning and explanation, but they do not remove the need for independent technical evidence. Explanations can also be plausible without being faithful; consequently, a production system should ground explanations in measured features rather than unrestricted narrative generation.

### 3.3 AI-generated and adversarial phishing

The emerging literature explicitly tests human-written, LLM-generated and adversarially modified messages. Kulal et al. (2025) show that robust preprocessing and adversarial evaluation matter when malicious wording is modified while intent is preserved. Tóth, Bisztray and Dubniczky (2025) contribute labelled distinctions among phishing, spam and legitimate messages and examine rephrased samples. Together, these works indicate that grammatical quality is no longer a dependable discriminator and that evaluation datasets must reflect current attack construction.

### 3.4 Behavioural and technical evidence

Authentication protocols and reputation services can detect spoofing and known malicious infrastructure, but they are not conclusive when a legitimate account or cloud domain is abused. Behavioural analytics adds relationship and temporal context: normal sending hours, typical recipients, attachment frequency, location, message length and transaction patterns. Its value is greatest when address-level trust fails, though it introduces privacy, cold-start and role-change concerns.

### 3.5 Explainability and lifecycle security

Local explanation methods such as LIME estimate which features influenced a prediction (Ribeiro et al., 2016). In cybersecurity, useful explanation must be brief, evidence-based and resistant to manipulation. NIST's adversarial machine-learning taxonomy emphasises testing and mitigation throughout the lifecycle rather than treating a deployed model as static (Vassilev et al., 2024).

## 4. Research Gap, Objectives and Design Propositions

### 4.1 Research gap

The literature reveals a fragmented evidence base. First, many detectors emphasise message content and underuse sender behaviour, communication history and attachment context. Second, random train-test division can leak campaign-specific patterns and does not approximate future deployment. Third, conventional corpora underrepresent multilingual, AI-assisted and adversarially paraphrased messages. Fourth, some high-performing models provide limited operational explanations. Fifth, static evaluation overlooks drift in both attack tactics and legitimate organisational behaviour. Finally, common accuracy measures do not reflect the unequal consequences of false negatives and false positives. A unified design is needed that addresses these limitations together and states how its added complexity will be tested.

### 4.2 Objectives

1. To critically examine why conventional and content-only phishing detectors are vulnerable to fluent, personalised and paraphrased attacks.
2. To identify complementary semantic, stylometric, technical, behavioural and attachment indicators relevant to AI-assisted phishing.
3. To design an explainable hybrid deep-learning architecture that fuses these evidence streams into a calibrated risk estimate.
4. To formulate a reproducible experimental protocol for temporal generalisation, adversarial robustness and component-level contribution.

5. To specify ethical, privacy and operational safeguards for deployment in cloud communication environments.

#### 4.3 Research questions

6. Which failure modes of conventional filtering become more pronounced when phishing text is generated or revised by an LLM?
7. Which combination of content, technical and behavioural evidence offers the most defensible detection design?
8. Does hybrid fusion improve detection and calibration over a transformer-only baseline under temporal and adversarial testing?
9. How can the system communicate risk factors without creating false certainty or disclosing exploitable detail?

#### 4.4 Testable design propositions

Because the present study does not implement a prototype, formal hypotheses are reserved for the subsequent experiment. The following propositions establish testable expectations without presenting unobserved results:

- P1: The full hybrid model will achieve a higher macro-F1 and area under the precision–recall curve than a transformer-only baseline on temporally held-out data.
- P2: Adding sender-behaviour features will reduce false negatives in compromised-account scenarios.
- P3: Content-only models will experience a larger performance decline under intent-preserving paraphrasing than the hybrid model.
- P4: Removing any major evidence module will produce a measurable loss for at least one attack subtype, demonstrating complementary rather than redundant value.

### 5. Research Methodology

#### 5.1 Research design

The paper adopts design-science research supported by structured secondary analysis. The design-science cycle comprises problem identification, definition of objectives, architecture development, specification of evaluation procedures and communication of limitations. Scholarly articles, standards and threat reports were examined for recurring attack features, model weaknesses, evaluation practices and governance requirements. The output is an artefact specification and validation plan, not a completed benchmark.

#### 5.2 Proposed data sources and labels

A future implementation should combine lawful public corpora such as SpamAssassin and Enron-derived legitimate mail, verified phishing collections, publicly reported malicious URLs, expert-reviewed AI-generated samples and controlled adversarial variants. Enterprise data may be used only with institutional approval, lawful authority, minimisation and de-identification. Each record should include, where available, the body, subject, safe header representation, extracted URL features, attachment metadata and a bounded behavioural feature vector.

| Class                 | Operational description                                                |
|-----------------------|------------------------------------------------------------------------|
| Legitimate            | Normal personal or organisational communication.                       |
| Spam                  | Unsolicited content without verified credential theft or fraud intent. |
| Conventional phishing | Human-authored malicious or deceptive communication.                   |

|                           |                                                                   |
|---------------------------|-------------------------------------------------------------------|
| AI-assisted phishing      | Phishing generated or materially rewritten by a generative model. |
| Adversarial phishing      | Phishing deliberately transformed to evade detection.             |
| Business email compromise | Fraudulent communication imitating an authorised business actor.  |

### 5.3 Inclusion, exclusion and leakage control

Included messages must have defensible labels, usable content and documented lawful provenance. Records with uncertain labels, corruption or personal information that cannot be appropriately protected should be excluded. Exact and near duplicates must be grouped before splitting so that variants of one campaign do not appear across training and test sets. Synthetic samples that reproduce operationally useful attack instructions should not be released. Dataset documentation should record collection dates, language, class origin, generator family where known and transformation history.

### 5.4 Pre-processing

The pipeline separates subject, body, headers, URLs and attachments; converts HTML to safe textual and structural features; normalises Unicode while retaining suspicious homoglyph information; detects language; masks unnecessary identifiers; tokenises text; and standardises numerical features using training-set parameters only. Cleaning must not erase security signals such as unusual punctuation, hidden text or deliberate misspellings. Potentially malicious links and files are never opened in the ordinary feature-extraction environment.

### 5.5 Reproducibility

The experimental report should publish dataset versions, split dates, label rules, random seeds, software versions, hyperparameters and decision thresholds. Where data cannot be shared, reproducible feature schemas and privacy-preserving aggregate statistics should be supplied.

## 6. Proposed Hybrid Architecture

The framework treats each message  $e$  as a collection of complementary views rather than a single text string. Specialised modules generate representations that are fused only after independent processing. This design reduces reliance on any one feature family and supports module-level explanations.

| Module                | Primary input               | Purpose                                                      |
|-----------------------|-----------------------------|--------------------------------------------------------------|
| Semantic              | Subject and body            | Detect deceptive intent, impersonation and requested action. |
| Stylometric           | Writing characteristics     | Measure style, tone and claimed-sender consistency.          |
| URL/domain            | Links and domain data       | Identify obfuscation, redirection and reputation anomalies.  |
| Header/authentication | Sender and routing fields   | Detect spoofing and identity inconsistencies.                |
| Behavioural           | Historical account activity | Identify departures from normal relationships and actions.   |
| Attachment            | Metadata and sandbox output | Estimate file-related risk without unsafe execution.         |

### 6.1 Processing sequence

1. Ingest the message through a protected gateway and assign a non-identifying record key.
2. Parse the body, subject, headers, URLs and attachments in an isolated service.

3. Apply data minimisation and create module-specific feature representations.
4. Estimate module embeddings or risk evidence, together with availability masks.
5. Fuse available evidence and generate a calibrated phishing probability.
6. Apply an organisation-specific decision threshold and response policy.
7. Present evidence-grounded reasons and uncertainty to the authorised reviewer.
8. Record feedback, monitor drift and trigger controlled retraining when required.

### 6.2 Semantic module

A pretrained transformer is fine-tuned to identify credential requests, payment instructions, authority claims, confidentiality demands, urgency, fear and attempts to bypass normal procedure. For token sequence  $x = (x_1, \dots, x_n)$ , the contextual representation is  $h_{\text{text}} = T(x)$ , where  $T$  denotes the fine-tuned encoder. The module should learn attack intent rather than simple presence of urgency words. Class weighting, focal loss or balanced sampling may be compared to address skewed class distributions.

### 6.3 Stylometric module

Stylometric features include sentence length, lexical diversity, punctuation, formality, greetings, signatures, emotional intensity and similarity to the claimed sender's prior style. The output  $h_{\text{style}}$  is a risk component, not proof of AI generation. Sender-style comparison should be attempted only when sufficient authorised history exists; otherwise the module must expose a missing-data indicator rather than infer certainty from a weak baseline.

### 6.4 Missing-modality handling

Not every platform supplies full headers, behavioural history or attachments. Training should simulate missing modalities, and the fusion layer should receive explicit availability masks. This avoids treating absence of evidence as evidence of legitimacy and permits consistent deployment across email and collaboration channels.

## 7. Technical and Behavioural Modules

### 7.1 URL and domain analysis

The URL module examines length, subdomain count, IP-address use, shortening services, Unicode look-alikes, displayed-versus-actual link differences, domain reputation, lawful domain-age information, redirects and suspicious parameters. Features are extracted without browsing to the destination in the ordinary pipeline. Dynamic inspection, when authorised, occurs in a segregated sandbox. Cloud-hosted links require careful interpretation because a reputable parent domain does not guarantee benign content.

### 7.2 Header and authentication analysis

The header module processes SPF, DKIM and DMARC outcomes; reply-to and return-path consistency; display-name impersonation; routing anomalies; and unusual sending infrastructure. Authentication success raises confidence in domain alignment but cannot establish benign intent: an attacker may control an authorised domain or compromised account. The model therefore uses authentication as one evidence stream rather than a hard allow rule.

### 7.3 Sender-behaviour analysis

For sender  $s$ , let  $b_s$  represent an authorised baseline containing aggregated characteristics such as typical sending time, recipient set, message length, attachment rate, geographic pattern and transaction behaviour. For new event  $e$ , an anomaly score can be expressed as  $A(e,s) = D(b_e, b_s)$ , where  $D$  is a learned or statistical distance. Relationship graphs may supplement individual baselines by modelling the frequency and direction of sender-recipient interaction.

Behavioural analysis requires safeguards. New employees, changed roles, travel and seasonal workflows can generate legitimate anomalies. Baselines should therefore decay over time, distinguish stable from volatile features and expose confidence based on history volume. High behavioural anomaly should trigger corroboration, not automatic accusation.

## 7.4 Attachment-risk analysis

The attachment module inspects file type, extension mismatch, macros, embedded scripts, password-protected archives, executable content, metadata, reputation and authorised sandbox results. Attachments must never execute inside the feature pipeline. Content disarm and reconstruction, malware scanning and quarantine remain separate controls even if their outputs contribute to the model.

## 7.5 Security of the detector

- Validate and isolate parsers because malformed documents can attack the inspection system itself.
- Restrict access to raw messages, behavioural profiles, model artefacts and explanation logs.
- Monitor poisoning, evasion, model extraction and privacy leakage across the lifecycle.
- Version models and features so each alert can be reproduced during audit.
- Use rate limits and response abstraction to reduce attacker feedback about internal thresholds.

## 7.6 Data protection principles

Feature collection should follow purpose limitation, minimisation, retention schedules and role-based access. Whenever possible, retain derived aggregates rather than raw communication content. Human review procedures should prohibit unrelated employee monitoring and provide an appeal or correction route for persistent false alerts.

# 8. Feature Fusion, Risk Classification and Explanation

## 8.1 Fusion model

Let  $h_i$  denote the representation from module  $i$  and  $m_i$  its availability mask. The available module outputs are projected to a common dimension. A gated attention mechanism assigns context-dependent

weights:  $\alpha_i = \frac{\exp(q_i)}{\sum_j \exp(q_j)}$ , calculated only for available modules. The fused representation is  $z^* = \sum \alpha_i h_i$ .

classification head estimates  $P(y = 1 | e) = \sigma(Wz * + b)$ , where  $y = 1$  denotes phishing risk and  $\sigma$  is the sigmoid function. Concatenation, late fusion and gradient-boosted meta-classification should be evaluated as alternatives so that the selected fusion method is justified empirically rather than assumed superior.

## 8.2 Calibration and cost-sensitive decisions

A score is useful only when its probabilities are calibrated for the deployment population. Temperature scaling or isotonic regression may be fitted on the validation set and evaluated with reliability curves, Brier score and expected calibration error. Decision thresholds should reflect organisational costs: a missed credential-theft attempt may be more damaging than a short verification delay, while excessive quarantining can interrupt operations and create warning fatigue.

| Illustrative score | Operational band | Illustrative response                                       |
|--------------------|------------------|-------------------------------------------------------------|
| 0.00–0.29          | Low              | Deliver; retain routine monitoring.                         |
| 0.30–0.59          | Moderate         | Display contextual warning or request confirmation.         |
| 0.60–0.79          | High             | Quarantine pending independent verification.                |
| 0.80–1.00          | Critical         | Block, isolate related artefacts and notify security staff. |

Note. These bands are design examples, not validated thresholds. Each organisation must calibrate them on representative data, risk appetite and response capacity.

## 8.3 Explainable decision module

The explanation layer should translate measured evidence into short, reviewable statements: a reply-to mismatch, a newly observed sender-recipient relationship, a suspicious destination domain, an unusual payment request or risky attachment property. It may highlight influential phrases but should avoid presenting attention weights as causal proof. Explanations should include uncertainty and identify missing evidence.

An appropriate alert might state: “High risk: the reply-to domain differs from the authenticated sender domain; this account has not previously contacted the recipient; and the message requests urgent credential verification.” The system should not disclose exact weights, thresholds or bypass conditions that would help an attacker optimise evasion.

## 8.4 Human oversight

Moderate and high-risk decisions should support review, feedback and reversal. Reviewer outcomes become labels only after quality control, because hurried decisions can propagate bias. Automation may quarantine clear threats, but consequential actions such as disciplinary attribution should never rely solely on the classifier.

# 9. Experimental Evaluation Protocol

## 9.1 Temporal and external validation

Messages should be ordered by collection time. The oldest 70% form the training set, the next 15% the validation set and the newest 15% the test set. Campaign groups and near duplicates must remain within one partition. Temporal separation tests whether the detector recognises attacks that occur after training. A second dataset from a different source or organisation should be reserved for external validation where lawful and feasible.

## 9.2 Baselines

The full framework should be compared with a rule-based filter, logistic regression, support-vector machine, random forest, recurrent network and transformer-only classifier. Baselines must receive appropriate tuning budgets, identical train-validation-test boundaries and equivalent access to permitted features. A comparison is otherwise unfair and provides little evidence about the value of hybridisation.

## 9.3 Metrics

Let TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives. The

principal measures are  $\text{Precision} = \frac{TP}{TP+FP}$ ,  $\text{Recall} = \frac{TP}{TP+FN}$ , and  $F1 = \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$

reported but not treated as sufficient under class imbalance. Macro-F1, area under the precision–recall curve, area under the receiver-operating-characteristic curve, Matthews correlation coefficient, false-negative rate and false-positive rate should be included. Performance must also be stratified by attack type, language, data source and message age.

#### 9.4 Efficiency and calibration

Operational evaluation should record median and tail inference latency, throughput, memory use, cost per message and the rate at which messages require human review. Calibration is assessed with Brier score, expected calibration error and reliability plots. These measures reveal whether a model that classifies correctly also expresses trustworthy confidence.

#### 9.5 Error analysis

A blinded sample of false positives and false negatives should be independently reviewed by security and language experts. Errors are coded by likely cause: ambiguous intent, label uncertainty, new brand, missing header, compromised account, multilingual content, attachment dependence or adversarial transformation. Disagreements are resolved through a documented adjudication process. This qualitative analysis identifies design weaknesses that aggregate metrics conceal.

#### 9.6 Repeated evaluation

Where stochastic training is used, experiments should be repeated across multiple seeds and reported as distributions rather than a single best run. Hyperparameter selection must use the validation set only. The test set remains sealed until the analysis plan and thresholds are fixed.

### 10. Robustness, Ablation and Statistical Analysis

#### 10.1 Robustness tests

The held-out evaluation should include intent-preserving paraphrasing, translation and back-translation, synonym substitution, Unicode homoglyphs, character insertion, hidden HTML, shortened URLs, look-alike domains, unseen brands and simulated legitimate-account compromise. Transformations must be reviewed to confirm that the label and malicious objective remain unchanged. Robustness is reported as both absolute performance and degradation relative to unmodified samples.

#### 10.2 Ablation study

| Removed component       | Question answered                                           |
|-------------------------|-------------------------------------------------------------|
| Semantic module         | How much does meaning contribute across attack types?       |
| Stylometric module      | Does writing-pattern evidence add value beyond semantics?   |
| URL/domain module       | How much link evidence supports textually fluent attacks?   |
| Header module           | What is gained from authentication and routing information? |
| Behavioural module      | Does context improve compromised-account detection?         |
| Attachment module       | Which file-dependent attacks require attachment evidence?   |
| Calibration/explanation | How do probability quality and reviewer utility change?     |

#### 10.3 Statistical analysis

Paired predictions from two classifiers on the same test messages may be compared with McNemar's test. Bootstrap confidence intervals should be reported for macro-F1, precision–recall AUC and calibration measures. When several classifiers are evaluated across multiple datasets or time windows, the Friedman test may be followed by multiplicity-adjusted Wilcoxon signed-rank comparisons. Effect sizes and confidence intervals should accompany p-values. A two-sided significance level of .05 may be

specified in advance, but practical importance should be judged against operational cost and uncertainty rather than statistical significance alone.

#### 10.4 Hypothesis decisions

P1 is supported only if the full model improves pre-specified primary metrics on sealed temporal and external tests, not merely on validation data. P2 requires a statistically and operationally meaningful false-negative reduction within compromised-account scenarios. P3 is assessed through comparative degradation under matched paraphrases. P4 requires ablation losses that are reproducible across seeds or datasets. Null or mixed results must be reported; module complexity should be removed when it offers no defensible benefit.

#### 10.5 Threats to validity

Internal validity may be affected by label error, duplicate leakage and unequal tuning. Construct validity is threatened if “AI-generated” is inferred from style rather than documented provenance. External validity is limited by organisation, language, time and channel. Statistical conclusion validity is weakened by repeated uncorrected comparisons or selection of the best run. The proposed protocol addresses these threats but cannot eliminate them.

### 11. Discussion and Practical Implications

The framework reframes phishing detection as evidence integration. A polished message may evade spelling rules yet still conflict with authentication, relationship history, normal transaction behaviour or destination-domain evidence. Conversely, a legitimate urgent request need not be blocked when technical and contextual indicators support authenticity. The design therefore reduces dependence on a fragile assumption: that malicious intent is visible in poor language.

Hybridisation is not automatically beneficial. More modules increase engineering cost, latency, privacy exposure and failure modes. The evaluation plan is consequently central to the contribution. Temporal tests establish whether gains survive change; ablations test whether each module earns its complexity; calibration links scores to action; and error analysis reveals who or what is disproportionately affected. A simpler model should be preferred when it achieves equivalent risk-adjusted performance.

#### 11.1 Institutional and policy implications

- Universities and organisations should treat AI-assisted phishing as an evolving social-engineering risk, not solely as an AI-authorship problem.
- Procurement standards should require temporal validation, subgroup reporting, calibration evidence, audit logs and documented update procedures.
- Security policies should combine automated detection with multifactor authentication, out-of-band transaction verification and incident reporting.
- Employee behavioural data should be collected only for defined security purposes, with retention limits and oversight against function creep.
- Awareness programmes should teach users to verify context and requested action rather than search only for grammatical mistakes.

#### 11.2 Limitations

The present paper does not implement or benchmark the architecture. Access to realistic behavioural data is constrained by privacy and confidentiality. Public corpora may contain stale attacks, uncertain labels and language imbalance. AI authorship may be unknown or wrongly inferred. Domain-reputation services introduce external dependency, and sandbox analysis adds delay and cost. Behavioural

baselines can penalise new staff, changed roles or travel. Attackers may adapt to visible warning patterns. These limitations prevent performance claims and require controlled, ethically reviewed implementation.

### 11.3 Future research

Future work should implement the framework in a controlled cloud environment; evaluate multilingual and code-mixed messages; investigate privacy-preserving or federated behavioural modelling; compare graph-based sender-recipient models; examine online learning with drift safeguards; and test explanation formats through authorised user studies. Extension to voice, video and cross-channel impersonation is important but requires distinct datasets and governance. Red-team testing should be authorised, isolated and designed to improve defence without producing reusable attack artefacts.

## 12. Conclusion

Generative AI enables attackers to produce fluent, personalised and rapidly varied phishing messages, reducing the value of familiar warning signs. The appropriate response is not a detector that merely guesses whether text was written by a machine. It is a context-aware security system that identifies deceptive intent and corroborates it with technical, behavioural and attachment evidence.

This paper has proposed a six-module hybrid architecture with masked attention fusion, calibrated risk scoring and evidence-grounded explanation. It has also defined a validation protocol based on temporal separation, external testing, adversarial transformations, component ablation, paired statistical analysis and operational metrics. The framework's advantages remain propositions until implementation. That boundary is essential to academic integrity and to credible cybersecurity research.

Effective protection will continue to depend on layered controls: secure identity, domain protection, transaction verification, protected analysis infrastructure, human review, privacy governance and continuous monitoring. A hybrid model can strengthen this system only when its incremental value is demonstrated transparently and its use remains proportionate to organisational risk.

## References

1. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT (pp. 4171–4186). <https://doi.org/10.18653/v1/N19-1423>
2. European Union Agency for Cybersecurity. (2024). ENISA threat landscape 2024. <https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024>
3. European Union Agency for Cybersecurity. (2025). ENISA threat landscape 2025. <https://www.enisa.europa.eu/publications/enisa-threat-landscape-2025>
4. Koide, T., Fukushi, N., Nakano, H., & Chiba, D. (2024). ChatSpamDetector: Leveraging large language models for effective phishing email detection. In SecureComm 2024 (pp. 297–319). Springer. [https://doi.org/10.1007/978-3-031-94455-0\\_14](https://doi.org/10.1007/978-3-031-94455-0_14)
5. Kulal, D. H., Arannou, C. P., Anwar, A., Rastogi, N., & Niyaz, Q. (2025). Robust ML-based detection of conventional, LLM-generated, and adversarial phishing emails using advanced text preprocessing. arXiv. <https://doi.org/10.48550/arXiv.2510.11915>
6. National Institute of Standards and Technology. (n.d.). Phishing. <https://www.nist.gov/itl/smallbusinesscyber/guidance-topic/phishing>
7. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the

- predictions of any classifier. In Proceedings of KDD (pp. 1135–1144). <https://doi.org/10.1145/2939672.2939778>
8. Tóth, R., Bisztray, T., & Dubniczky, R. (2025). Constructing and benchmarking: A labeled email dataset for text-based phishing and spam detection framework. arXiv. <https://doi.org/10.48550/arXiv.2511.21448>
  9. Uddin, M. A., Mahiuddin, M., & Sarker, I. H. (2026). An explainable transformer-based model for phishing email detection: A large language model approach. *Computer Networks*, 277, 112061. <https://doi.org/10.1016/j.comnet.2026.112061>
  10. Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2024). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations (NIST AI 100-2 E2023). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-2e2023>
  11. Vaswani, A., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, 30.