Detection Of AI: Generated Phishing Emails Using Cross Model Generalization

A Study of Cross-Model Detection and Generalization In AI-Generated Phishing Emails

Alokananda Ghosh[1], Shamsher Singh Hada[2], Shruti Dhotre[3], Guna Dhondwad[4]
doi.org/10.64643/ijirt.208482-459

***Abstract*—Large language models (LLMs) have changed the practical characteristics of phishing email. A message that once required deliberate manual drafting can now be produced in seconds, adapted to a target role, and rewritten repeatedly while maintaining fluent grammar and a credible business tone. This development weakens the assumption behind many older filters: that phishing is recognizable because it is poorly written, repetitive, or linguistically unusual. The central question of this paper is therefore not whether artificial intelligence can detect phishing, but whether current AI-assisted detection approaches remain reliable when the generator, wording, context, and delivery conditions change.**

This paper presents a structured gap analysis of recent approaches to detecting AI- and LLM-generated phishing emails. The review focuses on four recurring dimensions: detection capability, cross-model generalization, operational latency, and interpretability. Recent evidence shows that stylometric systems can perform strongly when the evaluation distribution is close to the training distribution, while broader reviews identify over-reliance on manually assembled datasets and disproportionate attention to GPT-family models. A 2026 cross-model study further demonstrates that a detector trained on one generator can lose substantial transfer performance when evaluated on another, although threshold recalibration and multi-generator training can reduce the observed gap [1], [2], [3].

The analysis argues that the strongest and most defensible research gap is cross-model robustness under distribution shift. Latency and explainability remain important engineering constraints, but generalization is the issue that most directly challenges the scientific validity of a detector if its benchmark assumes a single or narrow generator family. The paper therefore treats generalization as the primary gap and positions latency, false-positive control, privacy, and explanation quality as secondary deployment requirements. Rather than claiming completed experimental results, the paper proposes an evaluation framework that can test detectors against known generators, held-out generators, paraphrased variants, human-written phishing, and legitimate business email.

The intended contribution is a human-readable research map: what existing approaches actually detect, where their evidence is strong, where their conclusions are narrow, and what a future lightweight detector should prove before being considered robust. This distinction is important because high accuracy on an internal test split is not equivalent to operational reliability. The paper concludes with a research direction based on multi-generator training, explicit held-out-generator testing, compact feature extraction, selective escalation, and structured explanations.

***Index Terms*—AI-generated phishing; LLM-generated email; phishing detection; stylometry; cross-model generalization; robustness; distribution shift; explainable cybersecurity; email security.**

The threat model also matters. An attacker does not need to use one model consistently. Different messages in the same campaign may be produced by different models, edited by humans, translated, or passed through paraphrasing systems. This makes generator identity an unstable feature. A research design that depends on identifying the exact source model may therefore have limited operational value. The stronger objective is to learn properties associated with malicious communication that survive changes in the generation process. This is one reason the paper emphasizes cross-model evaluation rather than model attribution alone.

A useful benchmark should also distinguish between message-level and campaign-level generalization. A model may detect individual phishing emails well

while missing a coordinated campaign whose messages vary in wording but share infrastructure or behavioral signals. Conversely, infrastructure indicators can become stale quickly. Combining independent evidence sources may therefore improve resilience, but the experiment must measure which source contributes what. This reinforces the value of ablation and error analysis in addition to aggregate performance metrics. The same principle applies to model choice. A model family should not be selected only because it produces the highest benchmark number. Its behavior under generator shift, its calibration, its computational footprint, and its explanation quality should all be considered. This creates a more realistic decision criterion for future implementation and prevents the research from reproducing the common pattern of optimizing one metric while leaving the underlying robustness problem unresolved.

I. INTRODUCTION

1.1. Background

Phishing is a social-engineering problem in which the attacker attempts to make a recipient perform an action that benefits the attacker: opening a document, following a link, disclosing credentials, changing payment details, or transferring funds.

Email defenses have traditionally combined sender authentication, reputation systems, URL analysis, lexical heuristics, content classification, and user reporting. These mechanisms remain useful, but generative AI changes the linguistic component of the threat. The attacker no longer has to write the message personally or repeatedly edit it to achieve a professional tone.

The shift matters because language quality is not the same as malicious intent.

A legitimate message can be urgent, persuasive, concise, or highly polished, while a phishing message can be grammatically perfect and tailored to the recipient. Consequently, a detector that treats grammar errors as strong evidence of phishing may become less useful as attackers adopt language models. The more important signals increasingly involve relationships among sender identity, request type, recipient history, links, authentication state, organizational context, and subtle linguistic structure.

1.2. Why the Detection Problem Is Changing

The recent literature does not support the simplistic claim that LLM-generated phishing is impossible to detect. In fact, several studies report strong performance under controlled conditions. The more precise concern is transferability: what happens when the model that generated the evaluation email is different from the model that generated the training examples? A 2026 cross-model evaluation explicitly frames this as a central question and reports that stylometric detectors can show a meaningful transfer gap across generators [3].

A second concern is dataset composition. The 2026 systematic literature review by Sivaneswaran and colleagues screened 142 records and retained 36 eligible studies. Its analysis reports that many studies rely on manually generated datasets and that GPT-family models receive considerably more attention than other LLM families [1]. This creates a risk that reported accuracy reflects familiarity with the data-generation process rather than a general property of AI-generated phishing.

1.3. Research Aim

The aim of this paper is to identify the most consistent limitations of existing AI/LLM-generated phishing-email detection approaches and translate those limitations into a testable research agenda. The paper is therefore deliberately a gap-analysis study rather than a claim of a completed production detector.

A further issue is the difference between benchmark difficulty and real-world difficulty. Public datasets often contain labels and collection artifacts that are invisible to a deployed system. If phishing and benign messages come from different sources, a classifier may learn source-specific vocabulary or formatting instead of malicious intent. Generator-held-out testing reduces one form of leakage, but source diversity and temporal separation are also important. A credible evaluation should document where messages came from, how they were labeled, and which preprocessing operations were applied before training.

Another practical consideration is class balance. A highly imbalanced deployment environment can make accuracy misleading because a detector may achieve a high score by favoring the majority class. Precision, recall, false-positive rate, and precision-recall curves are more informative for rare-event detection. The evaluation should also report counts, not only

percentages, so that readers can understand how many messages produced the reported errors. This is especially important when comparing studies with very different dataset sizes.

The review also highlights why future benchmark releases should preserve provenance. For each sample, researchers should know whether the message was human-written, AI-generated, AI-assisted, or transformed after generation, and which collection process produced the label.

Provenance makes it possible to design controlled experiments rather than treating every sample as interchangeable. It also allows later researchers to test whether a detector is sensitive to a particular collection artifact.

II. RESEARCH QUESTIONS AND REVIEW SCOPE

2.1. Research Questions

RQ1. What methodological families are currently used to detect AI/LLM-generated phishing emails?

RQ2. What datasets, generators, and evaluation settings are used, and how representative are they of changing real-world conditions?

RQ3. Which limitations recur across stylometric, machine-learning, transformer, LLM, retrieval-augmented, and hybrid approaches?

RQ4. Is cross-model generalization the most persistent gap, or is another limitation more consistently supported by the literature?

RQ5. What evaluation protocol would provide stronger evidence of robustness than a conventional random train/test split?

2.2. Scope

The analysis focuses on email-based phishing and closely related text-classification work in which artificial intelligence is used either to generate phishing content or to identify AI-generated characteristics. The scope includes stylometric features, classical machine learning, transformer classifiers, direct LLM classification, retrieval/context-based methods, and hybrid systems. General phishing detection research is used as background, but the central analytical lens is the additional challenge created by LLM-generated content.

2.3. Evidence Base

The uploaded draft supplied the initial conceptual structure and a set of candidate studies. For the final version, the discussion is anchored to verifiable recent sources available through academic or institutional repositories. Particularly relevant evidence includes the 2026 systematic review of 36 studies [1], the 2026 cross-model evaluation [3], the 2025 Opara et al. stylometric study [4], and a 2025 bibliometric review covering 1,096 AI/ML/DL/NLP phishing-detection documents [5].

This approach also resolves an important academic-writing issue: a literature gap should not be defined merely because one proposed system appears more sophisticated than an older one. A gap is stronger when multiple independent studies reveal a recurring weakness, when the weakness affects validity or deployment, and when a concrete evaluation can determine whether the weakness has been reduced.

2.4. Analytical Criteria

Each approach is assessed using five criteria: (i) detection objective, (ii) generator diversity, (iii) evaluation design, (iv) operational cost, and (v) interpretability. These criteria allow a high-accuracy result to be separated from evidence of robustness. The distinction is central to the paper's conclusion.

The literature also suggests that a detector should be evaluated at more than one decision threshold. Security operations rarely use a universal threshold independent of business risk. A financial institution may accept more analyst reviews to reduce false negatives, whereas a high-volume mailbox service may prioritize low latency and low false-positive friction. Reporting precision-recall behavior and calibrated probabilities therefore provides more useful evidence than a single accuracy value. Threshold selection should be performed on validation data and frozen before the held-out test is examined.

The research should also separate phishing detection from AI-text detection in the annotation scheme. Each message can be labeled along at least two dimensions: malicious versus legitimate, and human-written versus AI-assisted where that label is reliable. This enables a four-cell analysis and prevents the detector from being rewarded for solving the wrong task. For example, legitimate AI-assisted business email should not automatically be treated as malicious, while human-written phishing should remain in the phishing class.

A robust benchmark should include both difficult phishing and difficult legitimate email. The latter is especially important because legitimate communication often contains urgency, authority, payment instructions, links, or technical terminology. If benign examples are too generic, a detector can appear strong while simply learning that unusual business language is suspicious. Hard-negative design is therefore a central part of a meaningful gap analysis.

III. EVOLUTION OF AI-ASSISTED PHISHING DETECTION

3.1. Rule-Based and Reputation-Oriented Defenses

Early email defenses relied heavily on known malicious domains, URL reputation, sender authentication, blacklists, and deterministic rules. Their major advantage is speed and operational transparency. Their weakness is that a new attack can appear before a corresponding rule exists. These methods therefore remain valuable as a first line of defense but cannot by themselves characterize a previously unseen social-engineering message.

3.2. Classical Machine Learning

Machine-learning systems broadened the feature space by incorporating word frequencies, punctuation, message length, sender metadata, URL characteristics, lexical diversity, and other structured signals. Tree ensembles and linear models are attractive because inference is inexpensive and feature importance can be inspected. However, the same feature engineering can create a hidden dependency on the training distribution. If a generator changes its wording patterns, a feature that once separated malicious and benign messages may lose discriminative value.

3.3. Stylometric Detection

Stylometry attempts to measure writing style rather than relying only on obvious keywords. Opara, Modesti, and Golightly reported a 2025 study that tested major spam filters against AI-generated phishing and introduced 47 stylometric features; their XGBoost classifier achieved 96% accuracy under the reported evaluation setting [4].

The result is important because it demonstrates that AI-generated phishing can leave measurable stylistic traces. It is not, however, proof that the same traces remain stable across generators or future model versions.

3.4. Transformer and LLM-Based Detection

Transformer encoders can model longer semantic dependencies and may outperform shallow classifiers when training and test distributions are aligned. Direct use of general-purpose LLMs can add contextual reasoning but introduces cost, latency, privacy, and reproducibility concerns. More importantly, a powerful classifier does not automatically solve distribution shift. A detector can understand an email extremely well and still be systematically wrong when the attack distribution changes.

3.5. Hybrid and Context-Aware Systems

Hybrid systems combine multiple signals: content, metadata, retrieval, threat intelligence, and model reasoning. Their promise is not simply higher accuracy but better decision-making under ambiguity. Their cost is architectural complexity. Retrieval stores require governance, context features require privacy controls, and multiple model stages require careful latency accounting.

There is also a practical argument for lightweight first-stage models. Most incoming messages do not require expensive reasoning. Authentication failures, suspicious domains, obvious impersonation patterns, or strong stylometric anomalies can often be identified with compact features. The difficult cases are the minority in which evidence conflicts. Selective escalation can reserve deeper analysis for those messages. This architecture does not assume that small models are inherently more accurate; it assumes that they can be more efficient when used as a screening layer. The research question is whether the resulting cascade preserves robustness while lowering average cost.

A further opportunity is calibrated selective prediction. Instead of forcing every ambiguous email into a binary decision, the system can expose an uncertainty score and route only a controlled fraction of messages to deeper analysis. This makes the relationship between detection quality and compute cost measurable. The study can then report coverage versus risk: how many messages can be handled by the fast path while maintaining a chosen upper bound on false negatives or false positives.

The proposed architecture should also be evaluated for maintenance cost. A detector that requires constant retraining may be accurate but operationally expensive. A detector that can be updated through threshold calibration or a small adapter may be easier to maintain. Measuring update effort, not just initial training performance, would extend the research from a static benchmark toward a realistic security lifecycle.

IV. COMPARATIVE GAP ANALYSIS

The following matrix summarizes the principal evidence relevant to the research questions. Values are reported qualitatively unless a cited study provides a specific figure. The table intentionally avoids treating unrelated datasets or evaluation objectives as directly comparable.

Approach / Evidence	Primary Signal	Typical Strength	Main Limitation	Gap Relevance
Rule / reputation	URLs, sender, known indicators	Very fast; transparent	Reactive; weak on novel content	Low
Classical ML	Lexical + metadata features	Low compute cost	Distribution sensitivity	Moderate
Stylometry	Writing-style statistics	Strong AI-text signal in controlled tests	Generator-specific features may shift	High
Transformer classifier	Contextual text representations	Strong in-domain discrimination	Needs broad training coverage	High
Direct LLM judge	Semantic/contextual reasoning	Flexible interpretation	Latency, privacy, reproducibility	Moderate–High
RAG/context systems	History + threat intelligence	Better behavioral context	Retrieval cost and privacy	High
Hybrid cascade	Multiple tiers + escalation	Can balance cost and depth	Engineering complexity	High
Cross-model evaluation [3]	Held-out generator transfer	Tests robustness directly	Requires multi-generator corpus	Very High
Systematic review [1]	Evidence synthesis	Reveals field-wide patterns	Depends on quality of included studies	Very High

The strongest pattern is not that every existing detector fails. Rather, the literature shows that performance is highly conditional on the relationship between training data and evaluation data. The 2026 cross-model study is especially useful because it directly measures transfer between generators rather than inferring robustness from an in-distribution score [3].

The 2026 systematic review strengthens this observation from a field-level perspective: GPT-family models are disproportionately represented, while many studies rely on manually created data rather than shared benchmark corpora [1]. Together, these findings support a research gap centered on evaluation design. The problem is partly a detector problem, but it is also a benchmarking problem: without held-out generators and realistic benign email, robustness cannot be established convincingly.

Human factors should remain visible in the evaluation. A detector that blocks a legitimate invoice or executive request can create operational harm even when its overall false-positive rate appears small. Conversely, a detector that produces a high-confidence warning without a useful rationale may cause analysts to ignore future alerts. The paper therefore treats explanation quality and false-positive friction as deployment requirements rather than cosmetic interface features. Future experiments should include representative legitimate messages that are unusual, urgent, or outside a recipient's normal communication pattern.

Operational evaluation should include failure recovery as well. External threat-intelligence services may be unavailable, local models may be under load, or retrieval stores may temporarily fail. A production-oriented design should specify safe fallback behavior and measure the effect of degraded components. This does not require implementing a full enterprise mail system for the research paper; it requires

acknowledging that robustness includes the system around the classifier, not only the classifier itself.

In addition, the paper's emphasis on held-out generators does not imply that generator identity should be ignored completely. Knowing which generator produced an attack can be useful for forensic analysis, threat intelligence, and research diagnostics. The distinction is that generator identity should not be a prerequisite for basic phishing detection. The detector should remain useful even when the source model is unknown or changes unexpectedly.

V. THE PRIMARY GAP: CROSS-MODEL GENERALIZATION

5.1. Definition of the Gap

Cross-model generalization refers to the ability of a detector trained or tuned using phishing generated by one set of language models to retain useful performance when the attack messages are generated by different, unseen models. The distinction is subtle but fundamental. A detector may correctly distinguish GPT-generated phishing from human-written phishing because it learns generator-specific regularities. If the generator changes, those regularities may disappear.

5.2. Evidence from Recent Work

Gutierrez, Villegas-Ch, and Govea provide a direct experimental example. Their released corpus contains human-written phishing and LLM-generated phishing from GPT-4.1, DeepSeek 3.2, and Llama 3.3 70B. The authors report strong intra-model F1 values while observing a 28.1-percentage-point default-threshold cross-model transfer gap.

They also report that threshold recalibration reduced the gap substantially, and that an aggregated multi-generator detector performed strongly on the evaluated models [3].

This result is important for two reasons. First, it shows that generalization can be measured rather than assumed. Second, it suggests that the gap may contain multiple causes: genuine feature instability, calibration mismatch, or both. A useful research program therefore should not simply add more model capacity. It should identify which features remain stable across generators and which are artifacts of a particular generator family.

5.3. Why This Gap Is Stronger Than a Generic "Accuracy Gap"

Accuracy is usually a property of a specified dataset and split. Generalization is a property of the detector under change. In a security environment, the second property is often more consequential because attackers are not required to preserve the distribution used by the researcher. If a new model, prompt strategy, paraphraser, or translation pipeline changes the surface characteristics of phishing, the detector must continue to provide useful separation.

5.4. Researchable Form of the Gap

The gap can be converted into a falsifiable hypothesis: a detector trained on a diverse set of generators and evaluated on a held-out generator will show a smaller performance drop than a detector trained on a single generator. A second hypothesis is that a compact combination of stable stylometric, semantic, and metadata signals can approach the robustness of heavier models while reducing inference cost. These hypotheses can be tested without claiming results in advance.

The review further indicates that the field is moving toward hybridization. This is understandable because phishing contains multiple evidence types. Text can reveal social-engineering intent, while headers and authentication records provide infrastructure evidence. Threat-intelligence retrieval can add information unavailable in the message itself. A single model may be asked to infer all of these signals, but a modular design makes the sources of evidence easier to test independently. Modularity also supports safer updates: a threat-intelligence component can be refreshed without retraining the entire language model.

The explanation component can be made more rigorous by separating evidence extraction from natural-language summarization. The underlying decision should reference measurable signals, while the language model can translate those signals into analyst-friendly prose. This reduces the risk that a generated explanation invents a reason that was not actually used by the classifier. Future experiments should test explanation faithfulness by comparing the stated rationale with feature importance, retrieved evidence, and controlled perturbations.

A useful evaluation report should therefore contain three layers of evidence: aggregate metrics, per-

generator metrics, and qualitative error analysis. Aggregate results show overall performance; per-generator results reveal transfer behavior; and error analysis explains why failures occurred. Together these layers make it possible to distinguish a broad improvement from a narrow benchmark effect.

VI. SECONDARY GAPS: LATENCY, EXPLAINABILITY, DATA AND PRIVACY

6.1. Latency and Throughput

A detector for an enterprise mail gateway is not evaluated only by classification quality. It must also process messages at operational volume. A cloud LLM call can provide sophisticated reasoning, but network transfer, queueing, API response time, rate limits, and service availability can make the total decision path unsuitable for every message. The relevant metric is end-to-end latency, not only model inference time.

6.2. Explainability

A binary label is often insufficient for security operations. Analysts need to know whether the decision was driven by sender inconsistency, a suspicious request, unusual recipient context, a stylometric anomaly, or a combination of evidence. Explainability also matters when a legitimate business email is incorrectly quarantined. A structured rationale can help the analyst distinguish a model error from a genuinely unusual message.

6.3. Dataset Quality and Benchmark Fragmentation

The 2026 systematic review reports that many studies use manually generated datasets and that GPT-family models dominate the evaluated model distribution [1]. This creates a reproducibility problem. Two papers can both report high accuracy while using different definitions of phishing, different benign-message sources, different prompt templates, and different generators. Without shared protocols, a numerical leaderboard can conceal rather than resolve the robustness question.

6.4. Privacy and Context

Context-aware detection can improve behavioral analysis because it allows the system to compare a new message with prior communication patterns. However, storing and retrieving communication history introduces data-protection obligations. A practical architecture should minimize retained content, restrict access, document retention periods, and avoid unnecessary transmission of sensitive email to external services. Privacy is therefore not an optional feature; it is part of deployability.

6.5. Relationship Among the Gaps

These gaps interact. Adding a larger model may improve semantic understanding while increasing latency. Adding recipient history may reduce false positives while increasing privacy risk. Adding more training data may improve generalization while increasing collection and labeling cost. The objective is consequently not to maximize one metric in isolation but to identify a defensible operating point under multiple constraints.

Another research consideration is reproducibility. Model APIs change, safety filters change, and proprietary generators may not produce identical outputs from the same prompt over time. For this reason, future work should record model family, version when available, generation settings, prompt templates, sampling assumptions, and collection dates. These details are not administrative trivia. They define the distribution on which the detector was trained and determine whether a later researcher can reproduce the observed transfer gap.

Privacy-preserving design can also be evaluated experimentally. Instead of storing full message histories, a system could retain compact behavioral summaries such as sender frequency, normal request categories, or relationship patterns. The research question is whether these summaries preserve the false-positive benefit of contextual retrieval while reducing exposure of message content. This would turn privacy from a general ethical statement into a measurable architectural trade-off.

The research can also contribute to safer deployment by defining clear escalation criteria. Messages with conflicting evidence should not be forced into an automated allow or block decision when the cost of error is high. A review queue can be treated as a legitimate output class. This design acknowledges that uncertainty is part of cybersecurity and that automation is most useful when it improves human decision-making rather than pretending to eliminate it.

VII. PROPOSED RESEARCH DIRECTION

7.1. Conceptual Architecture

Based on the gap analysis, the recommended research direction is a lightweight, evidence-fusing detector rather than a single monolithic classifier. The architecture is proposed as a research design, not as a claimed completed result. Its purpose is to make cross-model robustness measurable while keeping routine inference inexpensive.

Stage	Role	Representative Evidence
Stage 1	Fast screening	Header/authentication checks, lexical statistics, stylometric features, URL indicators, and a compact tree-based classifier.
Stage 2	Context verification	Recipient-aware history, sender consistency, threat-intelligence indicators, and behavioral anomaly checks for ambiguous cases.
Stage 3	Selective reasoning	A locally hosted small language model or compact transformer produces a structured rationale only when earlier evidence is inconclusive.
Decision	Action	Allow, quarantine, or analyst-review outcome with evidence fields rather than a bare probability.

7.2. Why a Cascade Fits the Gap

A cascade makes the research question clearer. The fast stage can be evaluated for generator invariance; the context stage can be evaluated for false-positive reduction; and the reasoning stage can be evaluated for explanation quality. Because deeper analysis is reserved for ambiguous cases, the experiment can report both average latency and the percentage of messages escalated. This is more informative than reporting a single model inference time.

7.3. Expected Contribution

The contribution should be framed as a methodology for robust detection rather than a guaranteed numerical improvement. The research should demonstrate whether multi-generator training, held-out-generator testing, compact feature selection, and selective escalation together produce a detector that is more stable than a single-generator baseline. If the results do not support that claim, the study should report the failure and identify which assumption broke.

The strongest research contribution may therefore come from measurement rather than architectural novelty. A carefully controlled benchmark can reveal whether a simple detector already performs adequately when trained on diverse generators, whether additional context materially improves results, and whether a local reasoning stage is worth its computational cost. Such findings can be valuable even if the proposed architecture does not outperform every baseline. Negative or mixed results can identify which assumptions about AI-generated phishing are incorrect and prevent future systems from repeating the same evaluation mistake.

The literature also motivates longitudinal evaluation. A model that performs well immediately after training may degrade as language models, attacker prompts, and business communication patterns evolve. Periodic re-evaluation should therefore be part of the proposed lifecycle. The detector can be monitored using drift indicators, held-out samples, and newly observed phishing families. Retraining should be triggered by evidence of degradation rather than by an arbitrary calendar schedule alone.

Another benefit of the proposed framing is comparability. If future studies adopt the same held-out-generator protocol, researchers can compare detectors even when their internal architectures differ. This would shift the field toward evaluating robustness under controlled change rather than comparing isolated accuracy values. Such a benchmark could become more valuable than another single-model classifier because it measures the property that deployment most needs.

VIII. EXPERIMENTAL DESIGN AND EVALUATION PROTOCOL

8.1. Dataset Construction

The evaluation corpus should contain four primary groups: human-written phishing, LLM-generated phishing, legitimate business email, and legitimate but unusual or urgent business email. LLM-generated samples should come from multiple generator families, with at least one family held out from training. The generator identity should be recorded for analysis but withheld from the classifier during the

core test so that the experiment measures transfer rather than memorization.

8.2. Generator-Split Protocol

The central split should be by generator, not only by message. For example, training may use generators A, B, and C; validation may use a held-out portion from A, B, and C; and the primary robustness test may use generator D. A second experiment can reverse the arrangement so that each generator becomes the held-out family in turn. This leave-one-generator-out protocol provides a more stable estimate of cross-model robustness than a random split.

8.3. Metrics

Performance should be reported using precision, recall, F1, ROC-AUC, and false-positive rate. Because phishing detection is operational, the paper should also report calibration, confusion matrices, and performance on the urgent-legitimate subset. For latency, report median, 95th percentile, and 99th percentile end-to-end time, including feature extraction, retrieval, orchestration, and model inference. Average latency alone can hide queueing and tail behavior.

8.4. Baselines

At minimum, the proposed approach should be compared with: (a) a rule/reputation baseline, (b) a classical ML classifier, (c) a stylometric XGBoost model, (d) a transformer classifier, and (e) a direct LLM or compact local language-model baseline where feasible. Each baseline should use the same evaluation corpus and the same generator-held-out protocol. This prevents the proposed system from benefiting merely from a more favorable dataset.

8.5. Statistical and Error Analysis

Performance differences should be accompanied by confidence intervals or repeated-seed estimates where appropriate. More importantly, the study should inspect false negatives and false positives by category: impersonation, credential theft, payment request, delivery notification, IT support, HR communication, and other common pretexts. Error analysis can reveal whether a model is learning phishing behavior or merely learning a stylistic artifact of the corpus.

From a security perspective, robustness should be understood as graceful degradation. No detector can be expected to maintain identical performance against every future generator. A useful system should instead provide stable performance across known changes, calibrated uncertainty when evidence is weak, and an escalation path for ambiguous messages. This perspective shifts the objective from perfect classification toward resilient decision support. It also aligns the research more closely with operational security, where uncertainty is expected and analyst review is part of the control process.

In an academic setting, reproducibility is strengthened when the paper publishes not only a final score but also the exact split logic and preprocessing pipeline. Text normalization, removal of signatures, HTML stripping, URL replacement, and tokenization can all change the evidence available to a detector. These operations should be described explicitly. Otherwise, a future researcher may reproduce the model but not the experiment. A transparent preprocessing protocol is therefore part of the contribution.

The final paper therefore keeps the contribution deliberately evidence-led. It identifies what recent studies have demonstrated, distinguishes those findings from assumptions, and converts the remaining uncertainty into research questions. This approach makes the paper suitable as a foundation for an experimental implementation while avoiding unsupported claims about performance that has not yet been measured.

IX. ABLATION, ROBUSTNESS AND ADVERSARIAL TESTING

9.1. Ablation Study

A strong paper should show which components actually matter. The ablation plan should therefore remove one component at a time: stylometric features, metadata/authentication signals, contextual retrieval, and local reasoning. If removing a component produces little change, its inclusion should be reconsidered. If a component improves in-domain F1 but harms cross-model transfer, that trade-off should be reported rather than hidden.

9.2. Prompt and Paraphrase Variation

LLM-generated phishing is not a single fixed distribution. The same malicious intent can be expressed using different prompts, temperatures, languages, or levels of formality. Robustness testing

should therefore include paraphrased versions, changes in message length, subject-line variation, and controlled modifications of tone. The objective is not to teach evasion tactics but to test whether the detector depends too heavily on superficial wording.

9.3. Generator Evolution

A particularly valuable experiment is temporal or version-based evaluation. A detector trained using an earlier model version can be tested on output from a later version or a different model family. This approximates a realistic deployment condition in which the attacker changes tools without notifying the defender. A robust detector should degrade gracefully rather than collapse when the generator changes.

9.4. Calibration and Threshold Transfer

The cross-model study by Gutierrez et al. shows that some observed transfer loss can be reduced through threshold recalibration [3]. Therefore, evaluation should separate feature-level generalization from calibration effects. A detector that preserves ranking quality but requires a modest threshold adjustment has a different failure mode from one whose representations cease to separate the classes. This distinction can guide practical maintenance strategies.

9.5. Human Analyst Evaluation

If explainability is part of the contribution, analysts should rate the usefulness of the generated rationale. The evaluation can ask whether the explanation identifies the relevant evidence, whether it is consistent with the final decision, and whether it helps the analyst reach a decision more quickly. Explanation quality should not be inferred from the presence of a paragraph of generated text; it should be evaluated as a security-operations artifact.

The gap analysis also exposes an important distinction between model capacity and evidence diversity. Increasing parameter count may improve a detector's ability to model language, but it does not automatically expose the detector to the variations it must handle. A smaller model trained on diverse generators may therefore outperform a larger model trained narrowly when evaluated under distribution shift. This hypothesis should be tested directly rather than assumed. The comparison should control for training data, evaluation protocol, and hardware so that capacity and data diversity are not confounded.

A final methodological point is that no single evaluation split can establish universal robustness. The proposed study should use multiple complementary tests: held-out generator, held-out time period, held-out campaign type, and controlled perturbation. Agreement across these tests would provide stronger evidence than success on any one benchmark. If the results disagree, that disagreement is itself informative because it identifies which type of shift remains most damaging.

From an engineering perspective, the gap analysis suggests that the fastest path to improvement may be better data and evaluation rather than a larger model. If a detector fails because it has never seen the relevant generator family, increasing model size does not necessarily solve the problem. A broader training distribution, better hard negatives, and a held-out evaluation may provide more useful information about what the detector has actually learned.

X. DISCUSSION: WHAT THE LITERATURE ACTUALLY SUPPORTS

10.1. Strong Claims Supported by the Evidence

The literature supports several careful conclusions. First, LLMs can generate phishing content that is coherent, contextual, and persuasive, increasing the difficulty of relying on crude language-quality heuristics [1]. Second, stylometric features can provide useful discrimination in controlled experiments; Opara et al. report 96% accuracy with XGBoost in their evaluation [4]. Third, cross-model transfer is a measurable problem rather than a theoretical possibility: the 2026 cross-model study reports a substantial default-threshold transfer gap [3]. Fourth, the field suffers from dataset and model-selection concentration, with GPT-family models receiving disproportionate attention in the systematic review [1].

10.2. Claims That Should Not Be Made Without Experiments

The paper should not claim that one detector will always generalize across GPT, Claude, Gemini, Llama, DeepSeek, or future models. It should not claim a universal latency figure without hardware and traffic measurements. It should not state a specific F1 score as a measured result if the experiment has not been completed. Finally, it should not imply that AI-

generated phishing is inherently detectable merely because some datasets show strong stylometric separation.

10.3. Why the Gap Analysis Matters

The practical value of the gap analysis is that it changes the research objective. Instead of asking, "Which model gets the highest accuracy?" the more meaningful question becomes, "Which detector retains useful performance when the attack generator and communication context change?" That question is closer to the conditions a deployed defense actually faces. It also creates a cleaner experimental contribution because the held-out-generator protocol can be repeated by other researchers.

10.4. Implications for Future Systems

Future phishing detectors should increasingly combine content evidence with behavioral and infrastructure evidence. A message should not be judged only by whether its prose resembles machine-generated text. Sender authentication, relationship history, request sensitivity, link reputation, and organizational workflow can provide independent evidence. Such evidence is harder to reproduce merely by changing the language model that generated the email.

10.5. Academic Contribution

The contribution of this paper is therefore a research framing: cross-model generalization is the primary scientific gap, while latency, explainability, privacy, and false-positive handling are deployment constraints that should be evaluated alongside it. The proposed cascade is a candidate design for testing these requirements, not a substitute for empirical validation. Finally, the proposed research direction is intentionally falsifiable. If a multi-generator compact detector does not improve held-out-generator performance, the hypothesis should be rejected or refined. If contextual retrieval reduces false positives but causes unacceptable latency, that trade-off should be reported. If local reasoning produces explanations that analysts do not find useful, the explanation component should be redesigned. A research paper becomes stronger when it defines conditions under which its own proposal could fail. The purpose of this gap analysis is to make those conditions explicit.

Taken together, these considerations support a research program that values robustness evidence over architectural complexity. The best detector for this problem may not be the largest model. It may be the system whose errors are understood, whose performance degrades gradually under change, whose latency is predictable, and whose decisions can be audited. That is the standard against which the proposed research direction should ultimately be judged.

The final design principle is therefore simple: train for diversity, test for change, measure operational cost, and explain errors. These four principles connect the literature gap to an implementable research plan. They also provide a clear standard against which future versions of the proposed system can be judged.

XI. LIMITATIONS, ETHICS AND FUTURE WORK

11.1. Limitations of This Study

This paper is a gap-analysis and research-design study. It does not claim that the proposed architecture has already been deployed at line rate, nor does it claim a validated 97.4% F1 or a fixed 41.8 ms latency. Those values appeared in the earlier working material as target-style figures, but they are not treated here as experimental findings. Removing unsupported performance claims is essential for academic integrity and makes the paper more defensible.

A second limitation is the heterogeneity of the literature. Studies differ in their definition of phishing, generator selection, benign-message sources, class balance, feature sets, and evaluation splits. Direct numerical comparison is therefore inappropriate unless the underlying experimental conditions are aligned. The present analysis emphasizes methodological patterns rather than constructing a synthetic leaderboard from incomparable numbers.

11.2. Ethical and Privacy Considerations

Research involving phishing email can create dual-use risk. Datasets should be handled defensively, operational thresholds should not be published when they would materially enable evasion, and generated examples should be stored and shared under responsible-use conditions. Recipient-history features require additional safeguards because they can expose sensitive communication patterns. Any

implementation should apply data minimization, access control, retention limits, and audit logging.

11.3. Future Work

- Build a multi-generator corpus with explicit generator metadata and a held-out-generator evaluation protocol.
- Reproduce published stylometric baselines and test whether their strongest features remain stable across generators.
- Compare compact classifiers with transformer and local SLM baselines under identical latency measurement conditions.
- Evaluate calibration, false positives, and analyst usefulness in addition to F1 and accuracy.
- Run longitudinal tests as new model versions and new prompt styles enter the corpus.
- Publish code, feature definitions, split logic, and evaluation scripts where licensing and safety allow reproducibility.

11.4. Final Research Position

The central proposition is deliberately modest: a detector should not be called robust merely because it performs well on messages produced by the same generator family represented in its training data. Robustness must be demonstrated under generator change. This is the most direct and testable gap identified by the evidence reviewed in this paper.

XII. CONCLUSION AND REFERENCES

12.1. Conclusion

AI-assisted phishing has not made detection impossible; it has made simplistic assumptions about phishing less reliable. Grammar quality, personalization, and fluency are no longer sufficient indicators of legitimacy. The research literature shows that stylometric, machine-learning, transformer, and LLM-based methods can achieve strong results under particular conditions, but the scientific question of robustness becomes harder when the generator changes. Recent cross-model evidence and the 2026 systematic review both point toward this issue as a meaningful gap [1], [3].

This paper therefore identifies cross-model generalization as the primary research gap and treats latency, false-positive handling, explainability, dataset quality, and privacy as closely related secondary requirements. The proposed research direction is a selective, evidence-fusing architecture whose main purpose is to test whether multi-generator training and held-out-generator evaluation can produce more stable detection without requiring every message to be processed by a large model. The framework remains a proposal until experiments validate it.

For the next stage of the project, the most valuable contribution is not another isolated accuracy number. It is a reproducible benchmark in which detectors are trained under one distribution and challenged under another, with realistic legitimate email, operational latency, and interpretable error analysis. Such an evaluation can reveal whether a detector has learned a durable property of phishing or merely learned the stylistic fingerprint of a particular generator.

REFERENCES

[1] D. Sivaneswaran, C. T. E. R. Hewage, H. M. K. K. M. B. Herath, R. S. Rathore, V. K. Singh, and W. Jiang, “A systematic literature review of large language models in phishing attack generation and detection,” *Array*, vol. 30, Art. no. 100775, 2026, doi: 10.1016/j.array.2026.100775.

[2] D. Sivaneswaran et al., “A systematic literature review of large language models in phishing attack generation and detection,” research repository record, 2026. Used here for bibliographic cross-checking of the published Array article.

[3] R. Gutierrez, W. Villegas-Ch, and J. Govea, “Cross-model evaluation of phishing detectors against LLM-generated emails,” Frontiers in Big Data, 2026. Dataset and experimental materials: Zenodo record 10.5281/zenodo.20250116.

[4] C. Opara, P. Modesti, and L. Golightly, “Evaluating spam filters and stylometric detection of AI-generated phishing emails,” Expert Systems with Applications, vol. 276, Art. no. 127044, 2025, doi: 10.1016/j.eswa.2025.127044.

[5] D. Popescul and L. D. Radu, “AI in phishing detection: a bibliometric review,” Frontiers in Artificial Intelligence, vol. 8, 2025, doi: 10.3389/frai.2025.1496580.

[6] R. Ghawa and A. Alhogail, “Hybrid AI-Based Detection of LLM-Generated Phishing Emails,”

Electronics, vol. 15, no. 15, Art. no. 3383, 2026, doi: 10.3390/electronics15153383.

[7] L. T. Manjate, “Robustness and Generalization of Artificial Intelligence Models for Text-Based Phishing Email Detection: A Critical State-of-the-Art Review and Research Agenda,” SSRN, Jul. 2026.

[8] I. Fette, N. Sadeh, and A. Tomasic, “Learning to detect phishing emails,” in Proc. 16th Int. Conf. World Wide Web (WWW), 2007.

[9] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, “Machine learning based phishing detection from URLs,” Expert Systems with Applications, vol. 117, pp. 345–357, 2019.

[10] National Institute of Standards and Technology, “Trustworthy Email,” NIST Special Publication 800-177 Rev. 1, 2019.

[11] MITRE, “Phishing (T1566),” MITRE ATT&CK Enterprise, accessed 2026.

[12] Cybersecurity and Infrastructure Security Agency, “Business Email Compromise,” CISA, accessed 2026.