IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, March 2026
Impact Factor: 8.2

CADET-Phish: A Cost-Aware Continual Adversarial Framework for Robust Phishing Email Detection under Concept Drift

Karedia Uzair, Chougle Talha and Shaikh Amr
Department of Computer Science
Royal College of Arts, Science and Commerce, Mumbai

Abstract: *Research into phishing email detection has benefited substantially from advances in deep learning and language models, with reported accuracies often exceeding 98% on benchmark datasets. However, these results largely reflect static evaluation conditions. In operational environments, phishing campaigns evolve rapidly; attackers increasingly leverage AI to generate linguistically sophisticated content, introducing both adversarial perturbations and concept drift. Consequently, existing systems still exhibit three persistent weaknesses: limited robustness to adversarial rephrasing, brittle performance under distributional shifts, and prohibitive costs when naïvely integrating large language models (LLMs) into the detection pipeline.*

To address these gaps, this paper proposes CADET-Phish, a cost-aware continual adversarial framework for phishing email detection. CADET-Phish combines four key components: (i) an efficient deep learning base classifier for high-throughput filtering, (ii) an adaptive drift and uncertainty monitor, (iii) an LLM-assisted semantic analyser invoked only on selected borderline and drift-suspect samples, and (iv) an adversarial augmentation module that generates realistic phishing paraphrases to harden the detector against evolving attacks. Concept drift is managed using streaming drift detectors and confidence trends, while adversarial robustness is improved through iterative training on constrained, LLM generated paraphrases and structural perturbations inspired by realistic evasion strategies.

The framework is designed to be implementable with publicly available datasets and open-source tooling, enabling empirical evaluation of robustness, drift adaptation, and cost–performance trade-offs. We outline experimental scenarios to compare CADET-Phish against conventional deep learning baselines and pure LLM classifiers, emphasizing operationally relevant metrics such as false positives on benign emails, robustness to AI-generated phishing, and LLM invocation cost.

Keywords: Phishing detection, adversarial machine learning, concept drift, large language models, email security, cost-aware hybrid architecture.

I. INTRODUCTION

Phishing remains one of the most pervasive and damaging cyber threats, serving as a primary vector for credential theft, account compromise, and downstream attacks such as ransomware. Recent advances in deep learning—particularly models leveraging contextual embeddings, recurrent architectures, and attention mechanisms—have reported impressive detection accuracies on benchmark datasets. However, these results are largely derived from static and curated corpora, which fail to reflect the adaptive nature of real-world phishing campaigns. In operational settings, attackers continuously modify linguistic patterns, delivery strategies, and contextual cues, causing model performance to degrade over time.

The rapid adoption of generative artificial intelligence has further intensified this challenge. Modern phishing emails, increasingly authored or refined by large language models, exhibit high linguistic fluency and contextual coherence, enabling them to evade traditional filters that rely on surface-level patterns. While Large Language Models (LLMs) on

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, March 2026
Impact Factor: 8.2

the defensive side demonstrate strong semantic reasoning and explainability, their computational cost and latency make them impractical as standalone solutions for high-throughput email filtering systems.These dynamics highlight a critical gap in existing phishing detection approaches: the lack of systems that simultaneously address adversarial robustness, concept drift, and operational cost constraints. To bridge this gap, we propose **CADET-Phish**, a cost-aware continual adversarial framework for phishing email detection. The framework combines an efficient deep learning base classifier with real-time drift and uncertainty monitoring, selectively invokes LLM-based semantic analysis only for high-risk or ambiguous samples, and incorporates realistic adversarial data augmentation to improve resilience against evolving attacks. By integrating these components into a unified pipeline, CADET-Phish aims to move phishing detection beyond static benchmarks toward reliable and economically viable deployment in dynamic environments.

II. LITERATURE REVIEW

2.1 Deep learning for phishing detection

Deep learning models have demonstrated strong performance in phishing email detection, consistently outperforming traditional machine learning approaches on benchmark datasets. Architectures leveraging contextual embeddings and attention mechanisms are particularly effective at capturing semantic patterns in email content, resulting in high reported accuracy and F1 scores.

However, existing studies largely evaluate these models under static experimental conditions, relying on public corpora such as Enron, SpamAssassin, and PhishTank. While issues like class imbalance and interpretability are often addressed, the assumption of a fixed data distribution remains prevalent. As a result, these models are vulnerable to performance degradation when deployed in real-world environments characterized by adversarial manipulation and evolving phishing strategies.

2.2 Adversarial machine learning in phishing and spam

Prior research in adversarial machine learning has demonstrated that phishing and spam detection systems are susceptible to evasion through carefully crafted perturbations. Techniques such as synthetic data generation and adversarial augmentation have been shown to degrade classifier performance, confirming that low-cost modifications can significantly impact detection rates.

Despite these insights, much of the existing work focuses on feature-level manipulation or phishing websites rather than raw email content. Moreover, the adversarial robustness of LLM-based phishing detectors remains largely unexplored, with limited systematic investigation into red-teaming or continual adversarial training. Consequently, while vulnerability is well established, the literature lacks operationally grounded, adaptive defense architectures tailored to live email streams.

2.3 Concept drift and streaming detection

Concept drift, referring to changes in the relationship between input data and target labels over time, is a well-documented cause of performance degradation in dynamic domains such as fraud and spam detection. Established drift detection algorithms, including DDM, EDDM, and ADWIN, are commonly used to monitor error rates or distributional changes and trigger model updates.

However, despite their proven effectiveness in other streaming applications, drift detection mechanisms are rarely systematically integrated into phishing email detection pipelines. Most existing systems continue to rely on batch training or infrequent retraining cycles, limiting their ability to adapt to rapidly evolving phishing tactics in long-term deployments.

2.4 Identified research gaps

Synthesizing the existing body of work reveals four critical gaps that necessitate further investigation:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, March 2026
Impact Factor: 8.2

- **Adversarial Robustness:** There is insufficient protection against email-level perturbations, particularly those generated by LLMs to paraphrase attacks.
- **Concept Drift Adaptation:** Despite its proven importance in spam filtering, drift adaptation is often treated cursorily in phishing detection research.
- **Cost-Aware Hybrid Designs:** While hybrid ML/LLM architectures are theoretically promising, there is a lack of concrete, evaluated pipelines tailored to the economic constraints of email processing.

These unresolved challenges motivate the specific research objectives and architectural design proposed in this paper.

III. METHODOLOGY

3.1 Identified Research Gaps

A synthesis of existing literature reveals several critical gaps that limit the effectiveness of current phishing email detection systems in real-world deployments.

First, there is a lack of empirically validated frameworks that simultaneously address adversarial robustness against AI-generated phishing, adaptation to concept drift in streaming email traffic, and the cost-aware integration of Large Language Models (LLMs). Most existing approaches either optimize for static accuracy or attempt to deploy monolithic LLM-based solutions without a viable strategy for managing computational and latency constraints.

Second, realistic adversarial augmentation at the email content level remains underexplored. Prior work largely focuses on URL manipulation or abstract feature-space perturbations, paying limited attention to generating linguistically plausible phishing emails that reflect realistic attacker behavior and operational constraints.

Third, although drift detection algorithms are well established in streaming domains, their application to phishing email pipelines is limited. Moreover, there is insufficient guidance on how to combine statistical drift signals with model uncertainty or user feedback to support timely and reliable adaptation.

Finally, evaluation protocols frequently rely on legacy, human-authored phishing datasets, offering limited insight into robustness against modern AI-generated attacks. The scarcity of contemporary benchmarks restricts the assessment of long-term resilience under evolving threat landscapes.

These gaps collectively motivate the need for a cost-aware, adaptive phishing detection framework capable of maintaining robustness and efficiency in dynamic, adversarial environments.

3.2 Research objectives

To address these challenges, this study pursues the following objectives:

- **Design CADET-Phish:** To develop a modular framework that integrates a deep learning base classifier, a drift and uncertainty monitor, an LLM-assisted analyser, and an adversarial augmentation module.
- **Define Cost-Aware Constraints:** To model both attacker and defender costs, using these operational constraints to guide the architectural design and decision policies.
- **Establish Experimental Protocols:** To specify rigorous testing scenarios using public datasets and synthetic AI-phishing campaigns, allowing for the empirical evaluation of robustness, drift adaptation, and cost–performance trade-offs.
- **Provide Implementation Guidelines:** To offer practical implementation strategies suitable for researchers and practitioners, prioritizing the use of open-source components and widely accessible platforms.

IV. PROPOSED MODEL / SYSTEM: CADET-PHISH

4.1 High-level description

We propose CADET-Phish, a multi-tiered framework engineered for the real-time analysis of high-velocity email streams. The system is predicated on the operational reality that the vast majority of email traffic can be accurately processed by efficient, lightweight models, while only a minority of high-risk or ambiguous messages necessitate deep semantic scrutiny. Accordingly, the framework comprises four core components:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, March 2026
Impact Factor: 8.2

1. **Base Classifier:** A lightweight deep learning model (utilizing architectures such as Bi-GRU or hybrid CNN–GRU) operating on token or subword embeddings. Its primary function is high-throughput filtering of benign and obvious phishing emails.
2. **Drift and Uncertainty Monitor:** A streaming surveillance module that tracks model confidence, output distributions, and key feature statistics. It employs drift detection algorithms, such as ADWIN or DDM, to identify potential concept drift in real-time.
3. **LLM-Assisted Analyzer:** A compact Large Language Model (LLM), fine-tuned or prompt-engineered specifically for security contexts. This component acts as a secondary auditor, invoked strictly for samples flagged by uncertainty or drift criteria.
4. **Adversarial Augmentation Engine:** A generative pipeline that produces constrained paraphrases and perturbations. Guided by specific threat models, this engine generates adversarial variants of phishing emails to fortify the model against evasion attempts during training.

4.2 Architecture explanation

The CADET-Phish framework is structured as a modular, multi-layer architecture designed for efficient and adaptive phishing email detection. The core functional layers are as follows:

- **Input and Preprocessing Layer:**
  This layer ingests raw email data, including headers, subjects, bodies, and selected metadata. It performs text normalization, tokenization, and structural feature extraction to generate representations suitable for downstream processing.
- **Base Classification Layer:**
  A lightweight deep learning model processes the preprocessed inputs and outputs a phishing probability score along with a calibrated confidence estimate, enabling high-throughput initial filtering.
- **Monitoring Layer:**
  This layer continuously tracks prediction confidence and output distributions. Drift detection mechanisms are applied over sliding windows to identify potential distributional shifts and flag uncertain or drift-affected samples.
- **LLM Analysis Layer:**
  Emails flagged by uncertainty or drift criteria are selectively routed to a compact LLM-based semantic analyzer, which provides an auxiliary classification decision and supporting rationale.
- **Decision Fusion Layer:**
  Outputs from the base classifier and the LLM are combined using a predefined fusion policy, such as LLM override under high-confidence disagreement, to produce the final classification.
- **Training and Augmentation Layer:**
  Upon detected drift, this layer initiates retraining using recently labeled samples augmented with adversarially generated email variants to improve robustness against evolving attacks.

4.3 Workflow explanation

The operational workflow of CADET-Phish is organized into four sequential phases that support continual and cost-aware phishing detection.

Phase I: Initialization and Training.

An initial training dataset is constructed from diverse phishing and benign email sources. The base classifier is trained using supervised learning with an emphasis on calibrated confidence estimation, while the LLM component is aligned using a subset of high-uncertainty samples.

Phase II: Real-Time Deployment. Incoming emails are first processed by the base classifier. Predictions with high confidence under stable conditions are finalized immediately, whereas samples flagged by uncertainty thresholds or

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 10, March 2026 Impact Factor: 8.2

drift indicators are selectively escalated to the LLM for semantic analysis. Final decisions are produced through a fusion of base and LLM outputs.

Phase III: Drift Adaptation and Retraining. The monitoring layer continuously evaluates prediction distributions to detect concept drift. When drift is identified, a retraining cycle is triggered using recently labeled samples augmented with adversarially generated variants to maintain robustness against evolving phishing strategies.

Phase IV: Cost-Aware Optimization.LLM usage is governed by predefined budget constraints. Uncertainty thresholds and drift sensitivity parameters are dynamically adjusted to balance detection performance with latency and computational cost.

V. RESULTS AND DISCUSSION

5.1 Experimental design (conceptual)

To rigorously evaluate the efficacy of the proposed framework, we outline a comprehensive experimental study designed to reflect real-world operational conditions.

Datasets:

Historical Corpora: We utilize established benchmarks, including the Enron corpus (ham) and reputable public phishing datasets (e.g., PhishTank, SpamAssassin). Where permissible, institution-specific logs will be included to enhance ecological validity [5, 8, 11].

Synthetic AI-Generated Subset: To simulate modern threats, we will generate a dataset of AI-authored phishing emails. A text-generation model will be prompted to craft sophisticated spear-phishing attempts, credential harvesting schemes, and invoice fraud, mimicking the linguistic fluency of current large language models.

Baselines for Comparison:

Static Deep Learning Baseline: A standard Bi-GRU or CNN–GRU classifier trained once on the initial dataset, representing the status quo in many current deployments.

Periodic Retraining Baseline: A classifier that is retrained at fixed intervals without explicit drift detection or adversarial augmentation, representing a naive approach to model maintenance.

Pure LLM Classifier: A direct application of a Large Language Model to classify all incoming emails, providing a ceiling for semantic understanding but a floor for efficiency.

Evaluation Metrics:

Classification Performance: Standard precision, recall, and F1-scores for both phishing and benign classes.

Robustness: Measuring the degradation in phishing recall when exposed to AI-generated content and paraphrased attacks, as well as resilience to known evasion strategies.

Drift Adaptation: The "time to recovery"—quantifying how quickly the system restores baseline performance following an induced shift in phishing tactics.

Operational Cost: Average LLM calls per email, end-to-end latency, and the approximate computational cost per million emails processed.

5.2 Expected qualitative outcomes

Under static evaluation conditions, the base deep learning classifier is expected to achieve performance comparable to existing approaches, maintaining high precision and recall on conventional phishing datasets. In contrast, a pure LLM-based classifier is anticipated to provide stronger semantic reasoning and interpretability, albeit at the cost of increased latency and operational expense.

Performance differences are expected to become more pronounced under dynamic conditions. When exposed to AI-generated phishing and paraphrased attacks absent from the initial training data, static deep learning models are likely to exhibit a notable decline in recall, reflecting their sensitivity to adversarial rephrasing. Periodic retraining strategies may partially mitigate this degradation but are expected to respond slowly due to their reliance on fixed retraining schedules rather than data-driven triggers.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 10, March 2026 Impact Factor: 8.2

CADET-Phish is designed to demonstrate improved stability under such conditions. The integration of drift detection enables targeted retraining only in response to statistically significant distributional shifts, while adversarial augmentation strengthens the decision boundary against realistic rephrasing and minor structural perturbations. Selective LLM invocation provides an additional layer of semantic validation for uncertain or novel samples without incurring the cost of full LLM-based processing.

Overall, CADET-Phish is expected to outperform static and naive periodic baselines in robustness and recovery following distributional shifts, while maintaining comparable classification performance and a bounded, predictable computational cost.

5.3 Discussion

From a security engineering perspective, the primary contribution of CADET-Phish is not merely a marginal gain in static accuracy, but the structured, systemic management of adversarial pressure, distributional change, and resource constraints. By fusing drift detection, adversarial augmentation, and costaware LLM usage into a unified pipeline, this framework operationalizes key recommendations that have previously existed in isolation across the literature on phishing detection, adversarial machine learning, and concept drift.

- This architecture also opens fertile ground for future research. Key questions emerging from this design include:
- **Signal Integration:** How can drift signals best be combined with model uncertainty and LLM disagreement metrics to optimize the triggering of retraining cycles?
- **Budget Allocation:** What is the optimal strategy for allocating the LLM "budget" across different uncertainty bands and drift windows to maximize detection gain?
- **Feedback Loops:** How can user feedback (e.g., "report phishing" actions) be incorporated most effectively to refine the model while minimizing the impact of label noise?

These questions are ripe for systematic experimentation and will be critical in the continued evolution of adaptive email security.

VI. ADVANTAGES OVER EXISTING METHODS

When compared with existing phishing detection approaches, CADET-Phish offers several distinct architectural and operational advantages:

- **Explicit Adversarial Robustness:**The framework integrates adversarial augmentation directly into the training loop, enabling the model to withstand realistic rephrasing and structural evasion attempts rather than reacting to attacks post-deployment.
- **Integrated Concept Drift Management**:Drift detection mechanisms combined with confidence monitoring provide a principled, data-driven approach to identifying shifts in phishing tactics and triggering timely model adaptation.
- **Cost-Aware LLM Integration:**By selectively invoking the LLM only for uncertain or drift-affected samples, CADET-Phish balances semantic reasoning capability with strict latency and computational budget constraints.
- **Modularity and Open Implementability:**The architecture is modular and can be implemented using widely available open-source tools, facilitating reproducibility and adaptation across datasets and deployment environments.

VII. LIMITATIONS

While CADET-Phish is designed to support adaptive phishing detection in dynamic environments, several limitations should be acknowledged.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, March 2026
Impact Factor: 8.2

- **Dependence on LLM Capabilities:**The effectiveness of the semantic analysis and adversarial augmentation components depends on the quality and alignment of the underlying Large Language Model. Inadequate fine-tuning or model hallucinations may introduce bias or reduce the linguistic realism of generated samples.
- **Approximation of Drift and Threat Models:**Drift detection relies on statistical indicators rather than direct observation of ground truth, which may lead to delayed detection of gradual shifts or occasional false positives. Similarly, adversarial augmentation focuses on known evasion strategies and may not fully capture highly novel, targeted attacks.
- **Sensitivity to Label Noise:**Incorporating user feedback for continual learning introduces the risk of noisy or incorrect labels. Without appropriate filtering and validation, such noise may adversely affect retraining outcomes.
- **Operational Complexity:**The multi-layer architecture requires coordinated deployment, monitoring, and maintenance. Effective operation therefore depends on the availability of mature MLOps infrastructure, which may pose challenges for resource-constrained environments.
- **Evaluation Constraints:**Empirical validation is limited by the availability of representative, up-to-date phishing datasets. Accurately simulating adaptive adversaries remains challenging, which may affect the generalizability of experimental results to live production settings.

VIII. FUTURE SCOPE

Future research may extend the CADET-Phish framework along several directions to further enhance adaptability and robustness.

- **Richer Threat Models and Human-in-the-Loop Evaluation:**Future work may model campaign-level attacker behavior and incorporate structured feedback from security analysts to assess usability, trust, and interpretability in operational settings.
- **Multi-Modal Phishing Signals:**Integrating additional modalities such as URL features, visual rendering cues, and attachment metadata could provide a more comprehensive defense beyond text-only analysis.
- **Active Learning and Label Efficiency:**Active learning strategies may reduce labeling overhead by prioritizing analyst input for samples flagged by high uncertainty or detected drift.
- **Cross-Organization Transfer and Federated Learning:**Privacy-preserving collaboration mechanisms, including federated learning, could enable shared adaptation to large-scale phishing campaigns without exposing sensitive data.
- **Formal Robustness Guarantees:**Future studies may explore theoretical robustness bounds to complement empirical evaluation and provide stronger guarantees under constrained adversarial perturbations.Collectively, these directions aim to advance phishing detection systems toward greater resilience, adaptability, and real-world deployability.

IX. CONCLUSION

Contemporary phishing detection systems, despite strong performance on static benchmarks, remain vulnerable to adversarial manipulation and concept drift—limitations that are increasingly exploited by AI-enabled phishing campaigns. To address these challenges, this paper introduced **CADET-Phish**, a cost-aware continual adversarial framework for phishing email detection.

By combining an efficient deep learning base classifier with real-time drift monitoring, selective LLM-assisted semantic analysis, and adversarial data augmentation, CADET-Phish provides a modular architecture designed for robust and economically viable deployment in high-volume email streams. The framework explicitly balances detection accuracy, adaptability, and operational cost, moving beyond static evaluation toward resilience in dynamic environments.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, March 2026
Impact Factor: 8.2

Although large-scale empirical validation remains a key direction for future work, the proposed design establishes a concrete foundation for systematic investigation of robustness, drift adaptation, and cost–performance trade-offs. In doing so, CADET-Phish bridges the gap between academic phishing detection research and the practical demands of real-world email security systems.

X. ACKNOWLEDGMENT

The authors gratefully acknowledge the guidance and constructive feedback provided by their academic mentors, as well as the broader research community whose work in phishing detection, adversarial machine learning, and concept drift informed this study.

REFERENCES

[1] MITRE & SEI, "Applications of Adversarial Machine Learning to Phishing Detection," technical report, 2018.
[2] Y. Li et al., "Improving Phishing Email Detection Performance through Deep Learning with Adaptive Optimization," *Scientific Reports*, 2025.
[3] A. Senol et al., "Joint Detection of Fraud and Concept Drift in Online Conversations," arXiv preprint, 2025.
[4] A. Panum et al., "Evolution of Phishing Detection with AI: A Comparative Study of ML, DL, and Small LLMs," arXiv preprint, 2025.
[5] M. Alsharnouby et al., "Advancing Phishing Email Detection: A Comparative Study of Deep Learning Models," 2024.
[6] Viso.ai, "Concept Drift vs Data Drift: Why It Matters in AI," technical article, 2025.
[7] SEI, "Applications of Adversarial Machine Learning to Phishing," presentation, 2018.
[8] H. Alqahtani et al., "Advancing Phishing Email Detection: A Comparative Study of Deep Learning Approaches," 2024.
[9] Evidently AI, "What is Concept Drift in ML, and How to Detect and Address It," technical guide, 2025.
[10] R. De Gaspari et al., "Towards Adversarial Phishing Detection," in *Proceedings of CSET*, USENIX, 2020.
[11] S. Sharma, "A Survey on Phishing Email Detection Techniques using LSTM and Deep Learning," *IJRASET*, 2025.
[12] A. Mallick, "Best Practices for Dealing with Concept Drift," Neptune.ai, 2023.
[13] P. Papadopoulos et al., "SpacePhish: The Evasion-space of Adversarial Attacks against Phishing Website Detectors using Machine Learning," ACM, 2022.
[14] N. K. Sahu et al., "Spam Email Detection Using Deep Learning Techniques," *Procedia Computer Science*, 2021.
[15] Activeloop, "Concept Drift Adaptation in Machine Learning," technical glossary article, 2023.