13th International Conference on Applied Informatics
Eszterházy Károly Catholic University
Eger, Hungary, February 19–21, 2026

LLM-Augmented Machine Learning for Phishing Email Detection: Enhancing Classification, Explainability, and Multilingual Support

Farouk Aziz[a], **Norbert Oláh**[a]

[a]University of Debrecen, Faculty of Informatics, Debrecen, Hungary
farouk.aziz@mailbox.unideb.hu,olah.norbert@inf.unideb.hu

Abstract

Phishing email detection systems face a trade-off between efficiency and capability. Traditional machine learning models perform well on English email bodies but struggle with multilingual content, uncertainty handling, and explainability. Large language models address these limitations but are costly and unsuitable for large-scale deployment. Although email systems operate at massive scale, the proposed architecture limits LLM invocation to a small fraction of emails, ensuring that LLM usage does not scale linearly with email volume. This paper proposes a selective hybrid framework that invokes LLMs only at critical failure points. Experimental results demonstrate improved robustness and interpretability with reduced computational overhead.

1. Introduction

Email remains the primary communication channel for organizations and individuals, making it a dominant vector for phishing attacks that exploit urgency, authority, and trust [2]. Advances in large language models have further intensified this threat by enabling attackers to generate persuasive, context-aware, and multilingual phishing emails at low cost [4].

2. Scientific Background

Early phishing detection relied on rule-based methods and blacklists, which proved ineffective against evolving attacks. Traditional machine learning using TF-IDF and classical classifiers later achieved strong performance on English email bodies but struggled with contextual variation, explainability, and linguistic diversity [2].

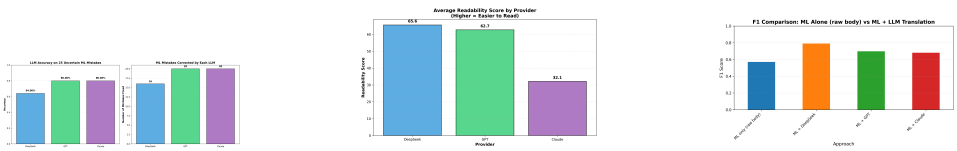
Recent work has explored large language models for phishing detection, often reporting high accuracy but incurring substantial computational cost due to universal LLM processing [3, 4]. Multilingual phishing detection remains challenging because most datasets and models are English-centric; translation-based approaches frequently suffer from semantic drift, motivating selective or context-aware LLM integration [5, 6].

3. Methodology and Experimental Design

This work evaluates a selective ML–LLM architecture as a proof of concept for cost-aware phishing email detection. A TF-IDF–based Linear SVM trained on approximately 14,000 English email bodies serves as the baseline classifier.

GPT-4, Claude 3.5 Sonnet, and DeepSeek v3.1 were independently evaluated for selective error correction, multilingual processing, and explainability. Only misclassified or uncertain emails were forwarded to a single LLM using task-specific prompts. Multilingual translation was assessed across 18 languages, while explainability was evaluated by converting LIME attributions into natural language and measuring readability.

4. Results and Analysis



(a) Error correction **(b)** Explainability **(c)** Translation

Figure 1. LLM performance comparison across selective hybrid tasks.

The evaluation confirms the technical feasibility and cost-effectiveness of selective LLM deployment at key ML failure points.

Error correction: LLM inference was triggered only for misclassified email bodies (1,598 of 16,478), reducing paid invocations by an order of magnitude com-

pared to universal LLM pipelines. On uncertain cases, GPT-4 and Claude achieved correction rates of approximately 80%, while DeepSeek achieved 64% (Figure 1a).

Multilingual translation: LLM-based translation enables a single English-trained ML model to process non-English email bodies without language-specific retraining. By preserving phishing-relevant semantic cues and deceptive patterns, this approach achieved 75% accuracy compared to a 50% ML-only baseline, with DeepSeek demonstrating the strongest performance (F1 0.791) (Figure 1c).

On-demand explainability: Explanations were generated only upon user request, avoiding the latency and cost overhead of universal XAI deployment. LLMs successfully transformed LIME feature attributions into human-readable explanations, with DeepSeek achieving the highest readability and GPT-4 the lowest latency (Figure 1b).

Overall, the results highlight an architectural advantage: selective LLM invocation achieves order-of-magnitude computational savings while preserving robustness, multilingual capability, and interpretability where traditional ML alone is insufficient.

5. Conclusion and Future Work

This work demonstrates that selective ML–LLM integration enables cost-aware phishing email detection. Future work will validate the approach at larger scale and compare LLM-based translation with ML models trained directly on non-English datasets. The implementation is available in the project repository [1].

References

[1] F. Aziz: *Hybrid Email Classifier Research: LLM-Augmented Machine Learning for Phishing Detection*, GitHub repository, 2025, URL: https://github.com/FaroukAziz2020/Hybrid-Email-Classifier-Research.

[2] P. Bountakas, K. Koutroumpouchos, C. Xenakis: *Enhancing Phishing Detection with a Combined LLM and Traditional Machine Learning Model*, in: Proceedings of the 16th International Conference on Availability, Reliability and Security (ARES 2021), 2021, pp. 1–12, DOI: https://doi.org/10.1145/3465481.3469205.

[3] G. Desolda, F. Greco, L. Viganò: *APOLLO: A GPT-based Tool to Detect Phishing Emails and Generate Explanations that Warn Users*, Proceedings of the ACM on Human-Computer Interaction 9.4 (June 2025), EICS003, DOI: https://doi.org/10.1145/3733049.

[4] F. Heiding, S. Lermen, A. Kao, B. Schneier, A. Vishwanath: *Evaluating Large Language Models' Capability to Launch Fully Automated Spear Phishing Campaigns: Validated on Human Subjects* (2024), arXiv: 2412.00586 [cs.CR], URL: https://arxiv.org/abs/2412.00586.

[5] D. Komosny: *Phishing Detection on Webpages in European Non-English Languages Based on Machine Learning*, Scientific Reports 15 (2025), p. 37472, DOI: https://doi.org/10.1038/s41598-025-21384-w.

[6] J. Rastenis, S. Ramanauskaitė, I. Suzdalev, K. Tunaitytė, J. Janulevičius, A. Čenys: *Multi-Language Spam and Phishing Classification by Email Body Text*, Electronics 10.6 (2021), p. 668, DOI: https://doi.org/10.3390/electronics10060668.