Proceedings of the
13th International Conference on Applied Informatics
Eger, Hungary, February 19–21, 2026
DOI: 10.17048/icai.2026.21
pp. 21–33

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

Farouk Aziz†, Norbert Oláh

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

Abstract. Nowadays, email is the world's primary communication channel, making it a dominant vector for phishing attacks that exploit urgency, authority, and trust. Advances in Large Language Models (LLMs) have intensified this threat by enabling attackers to generate persuasive, context-aware, and multilingual phishing emails at low cost. Traditional machine learning (ML) approaches are no longer sufficient, as phishing techniques have evolved with the widespread public use of AI. These systems also face the scarcity of datasets for non-English content and the lack of human-friendly explanations of model decisions. Large Language Models can address these limitations, but are costly and impractical for large-scale deployment given that billions of emails are sent daily. This paper proposes a selective hybrid framework combining ML with LLMs in a cost-aware architecture. LLMs are invoked selectively to correct ML mistakes, translate non-English emails for a single English-trained classifier, and generate on-demand human-readable explanations of model decisions. GPT-4, Claude, and DeepSeek were independently evaluated across these three tasks to identify which performs best and fastest. The ML mistakes were forwarded to each LLM using task-specific prompts, translation was assessed across 18 languages, and explainability was evaluated by converting Local Interpretable Model-agnostic Explanations (LIME) feature attributions into natural language and measuring readability. A Term Frequency–Inverse Document Frequency (TF-IDF) based Linear Support Vector Machine (SVM) trained on approximately 14.000 English

*This research was conducted for the 13th International Conference on Applied Informatics.
†Corresponding authors: farouk.aziz@mailbox.unideb.hu, olah.norbert@inf.unideb.hu

emails serves as the baseline classifier. The results confirm improvements in classification robustness, multilingual coverage, and explainability, while keeping computational overhead well below universal LLM pipelines. The three LLMs showed distinct strengths across tasks, highlighting that model selection should be tailored to each role within the architecture.

Keywords: phishing detection, large language models, machine learning, selective hybrid architecture, explainability, multilingual classification

2020 *Mathematics Subject Classification:* Primary 68T50; Secondary 68T05, 94A60

1. Introduction

Email is the dominant attack vector for cybercriminals. Phishing accounts for approximately 90% of data breaches, with an average loss of $3.86 million per incident [5]. Attacks take several distinct forms. Bulk phishing sends identical lure messages to millions of recipients simultaneously, relying on volume rather than personalisation – a single convincing template can yield thousands of victims. Spear phishing targets specific individuals by incorporating personalised details such as the recipient's name, role, organisation, or recent activity, making the message appear to come from a trusted colleague or institution. Whaling is a high-value variant of spear phishing directed exclusively at senior executives, board members, or regulators, typically impersonating legal counsel, auditors, or government authorities to authorise large transfers or disclose confidential data. Business Email Compromise (BEC) involves hijacking or spoofing a legitimate corporate email account to redirect payments, alter supplier banking details, or extract sensitive information, causing average losses exceeding $125.000 per incident [1, 16]. Beyond these established categories, AI-generated phishing has emerged as a new threat dimension: attackers leveraging LLMs can now produce grammatically flawless, contextually convincing, and personalised messages in any language at negligible cost, effectively eliminating the surface errors – misspellings, awkward phrasing, generic salutations – that rule-based and classical ML detectors have historically relied upon [15, 16].

Existing defensive systems either apply LLMs uniformly to all inputs [7] or use preprocessing filters that still route every non-trivial email through expensive inference [20]. While these systems are described as hybrid, they rely solely on LLMs for the actual classification decision, making the "hybrid" label a description of their pipeline structure rather than a reflection of combining fundamentally different classification approaches.

2. Scientific Literature

Phishing detection has progressed from static blacklists and rule-based filters through classical ML, deep learning, and transformer-based systems to current

adaptive multimodal approaches, as illustrated in Figure 1 [18]. These solutions are discussed in the following sections.

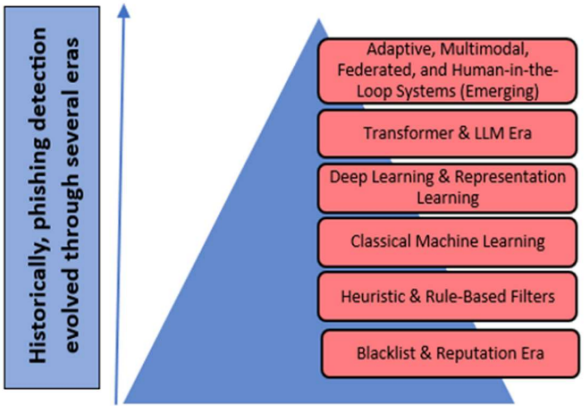

Figure 1. Evolution of phishing detection techniques, from blacklist and rule-based systems to adaptive multimodal approaches [18].

2.1. Classical Machine Learning Approaches

Supervised ML with Term Frequency–Inverse Document Frequency (TF-IDF) feature extraction and classifiers such as SVMs, Random Forests, and Logistic Regression established the first generation of scalable phishing detectors [12, 18]. Machine learning algorithms facilitate adaptive classification rules that outperform blacklisting or rule-based filtering, which rely on hand-coded rules susceptible to the changing nature of phishing attacks [12]. Tawil et al. [26] found TF-IDF with Random Forest and Logistic Regression achieved F1 of 0.98 at high throughput. Rastenis et al. [23] demonstrated Linear SVM with TF-IDF scales to multilingual datasets, noting its linear complexity as a deployment advantage. Despite their speed, classical models struggle with noisy or dynamic datasets [24] and fail entirely on non-English content [20].

2.2. Large Language Model-Based and Hybrid Approaches

LLM-based systems address the semantic limitations of classical ML but introduce prohibitive inference costs. LLMs require significant GPU resources, with training times ranging from hours to days, and API costs for processing 8 million tokens reaching $36–$128 depending on the model [13, 25]. Desolda, Greco, Viganò [7] proposed APOLLO, which integrates GPT-4o with VirusTotal and BigDataCloud APIs, achieving 97–99% detection accuracy and generating user-readable explanations through prompt chaining. Lim et al. [21] introduced EXPLICATE, combining a classical classifier with LIME/SHAP attributions and DeepSeek v3 to

translate feature weights into natural-language warnings, reaching 94.2% explanation accuracy. Lee [20] evaluated four open-source LLMs combining whitelisting and HTML parsing before LLM inference, improving throughput but still routing every non-whitelisted email through costly API calls. Tarapiah et al. [25] fine-tuned three transformer variants – DistilBERT (a lightweight, distilled version of BERT that retains 97% of its language understanding at 60% of the size), ALBERT (a parameter-efficient BERT variant using cross-layer parameter sharing to reduce memory footprint), and RoBERTa (a robustly optimised BERT with longer training and dynamic masking) – confirming high accuracy but at GPU-level inference cost. Kumar, Kushwaha, Rajpoot [18] confirm that hybrid ensemble approaches dominate production deployments and transformer-era models are the first to achieve meaningful multilingual generalisation. The critical gap is that LLMs are invoked uniformly rather than selectively across all prior systems.

Two further limitations remain unresolved: English-trained models drop below 60% accuracy on non-English emails [18, 23] with no general multilingual phishing dataset available [2]; and explainability tools such as LIME and SHAP require technical expertise to interpret, while LLM-generated explanations have only been evaluated universally rather than on demand [7, 21]. No prior system simultaneously addresses selective deployment, multilingual translation, and on-demand explainability within a single cost-aware architecture. Desolda, Greco, Viganò [7] achieve 97–99% accuracy with APOLLO but process all emails through GPT-4o, making real-time high-volume deployment economically challenging. Lee's system [20] reduces overhead through preprocessing but still applies LLM inference to every non-whitelisted email. Lim et al. [21] demonstrate comparable explanation quality with EXPLICATE but generate explanations universally rather than on demand, incurring multi-second overhead per email. The selective hybrid architecture proposed in this work goes further by using calibrated uncertainty as a routing signal, restricting LLM invocation to the cases where classical ML demonstrably fails.

2.3. Our Contribution

This paper proposes a selective hybrid architecture that exploits the heterogeneous difficulty of email classification, restricting LLM invocation to three well-defined failure points. Table 1 summarises how the proposed system compares to prior approaches across key evaluation dimensions.

Table 1. Comparison of system capabilities across key dimensions.

System	Efficiency	Accuracy	Multilingual	Explainability
ML only (SVM)	High	Moderate	No	Partial
LLM only (APOLLO [7])	Low	High	Yes	Yes
EXPLICATE [21]	Moderate	High	No	Yes
Proposed (Selective)	High	High	Yes	Yes

The main contributions are: (1) a selective architecture invoking LLMs only for the most uncertain misclassifications (calibrated confidence near 0.5); (2) on-demand LLM translation of non-English emails, improving accuracy from 50% to 75% across 18 languages; (3) an on-demand explainability module converting Local Interpretable Model-agnostic Explanations (LIME) attributions into plain-language warnings; and (4) a head-to-head evaluation of GPT-4, Claude 3.5 Sonnet, and DeepSeek across all three tasks.

3. Methodology

3.1. Machine Learning Classification Pipeline

The pipeline requires a classifier with high accuracy on English email text, calibrated probability estimates for uncertainty routing, and feature-level attributions for explainability. Raw emails are preprocessed via tokenization, stop-word removal, lowercasing, stemming, and URL removal [12, 19]. A Linear SVM with TF-IDF vectorization was selected after evaluating neural, tree-based, and linear alternatives [18, 26].

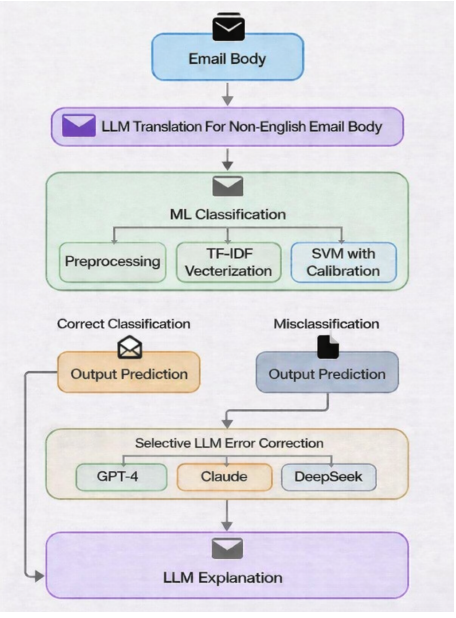

Figure 2. Selective hybrid architecture: ML processes all emails; LLMs are invoked only for uncertain misclassifications, non-English content, and on-demand explanations.

Figure 2 illustrates the overall architecture. TF-IDF (up to 10.000 features, n-gram range 1–3, min/max document frequency 2/0.95) was chosen for its sub-

millisecond CPU inference on sparse matrix representations, directly interpretable feature weights compatible with LIME, and natural emphasis on rare discriminative phishing terms [14, 26]. N-gram features capture contextual phrasal cues such as "verify your account" that distinguish phishing from legitimate messages [17]. SVM was selected for its strong generalisation in high-dimensional sparse spaces [18, 22, 26] and, critically, its calibration via Platt scaling (CalibratedClassifierCV, 5-fold) which transforms the decision function into posterior probabilities [10] – the mechanism enabling selective routing:

$$P(y = 1 \mid x) = \sigma\big(A \cdot f(x) + B\big)$$

where σ is the sigmoid function, $f(x)$ the SVM decision function, and A, B are learned calibration parameters. A well-calibrated classifier correctly quantifies the level of uncertainty associated with its predictions, which is essential for cost-sensitive classification and optimal decision making [10].

3.2. Selective Large Language Model Augmentation

The three LLM components share a common principle. LLM invocation occurs only when and where it provides measurable value over the ML baseline.

Error correction. The SVM assigns a calibrated posterior probability $P(y = 1 \mid x)$ to every email. An email is considered uncertain when this probability falls within the band $[0.5 - \delta,\ 0.5 + \delta]$ for a configurable threshold δ, meaning the model has low confidence in either class. The k most uncertain ML misclassifications – those with calibrated confidence closest to 0.5 – are routed to LLM re-classification, where k is a budget parameter controlling the accuracy–cost trade-off:

$$\text{Confidence}(x) = \max_{c}\ P(y = c \mid x)$$

Ambiguous cases benefit from the LLM's ability to reason about intent, tone, and deceptive context that lexical features miss [20, 25].

Multilingual translation. Non-English emails are routed to an LLM using a phishing-aware prompt that explicitly instructs the model to preserve urgency markers, grammatical errors, and suspicious phrasing rather than produce fluent translations. This is deliberate: these surface anomalies are the diagnostic signals the English-trained SVM relies on for classification [23]. Translation is performed on demand rather than universally, avoiding the per-email latency and cost of preprocessing all incoming email [27].

On-demand explainability. LIME feature attributions are generated for the ML prediction and passed to an LLM only when a user requests an explanation. Local Interpretable Model-agnostic Explanations (LIME) is an explainability technique that approximates any classifier locally with a simple interpretable model, identifying which input features most influenced a specific prediction [21]. SHapley Additive exPlanations (SHAP) is an alternative that assigns each feature a contribution score derived from cooperative game theory, measuring its average impact across all possible feature combinations [21]. LIME was chosen over SHAP for

its local operation on individual predictions and direct compatibility with the TF-IDF vocabulary, enabling the LLM to map numerical weights to concrete phishing tactics in accessible natural language [7, 21].

3.3. Large Language Model Provider Selection

GPT-4 (openai/gpt-4.1), Claude 3.5 Sonnet (anthropic/claude-3.5-sonnet), and DeepSeek (deepseek/deepseek-chat-v3.1) were selected as leading commercial models with strong multilingual pretraining and documented reasoning performance [13, 16]. Independent benchmarking by Artificial Analysis [3] confirms that these three providers represent distinct positions on the intelligence–speed–cost frontier, making them well-suited for comparative evaluation across latency-sensitive and accuracy-critical tasks. All API calls were made via OpenRouter at temperature 0.3 [13, 25].

4. Experimental Design

ML Training Dataset: 17.496 English emails sourced from the Phishing Email Collection on Kaggle [9] (after removing 1.154 duplicates from 18.650 originals), split 80/20 into training (13.996) and held-out test (3.500) sets using stratified sampling.

Hybrid Test Dataset: 16.478 emails from the Phishing Emails dataset on HuggingFace [8], completely separate from training. Used to assess real-world generalization and evaluate selective LLM error correction.

Multilingual Dataset: 36 manually curated emails across 18 languages (Hungarian, Spanish, French, German, Italian, Russian, Japanese, Portuguese, Polish, Dutch, Swedish, Turkish, Arabic, Korean, Czech, Greek, Norwegian, Finnish), one phishing and one legitimate per language, each reflecting culturally specific phishing tactics.

Model performance was assessed using 10-fold stratified cross-validation [11]. Cross-validation is a resampling technique that partitions the dataset into k equally sized folds; the model is trained on $k-1$ folds and evaluated on the remaining fold, repeating this process k times so that every sample is used for evaluation exactly once. Using $k = 10$ is a standard choice that balances bias and variance while preserving class distribution across folds. Stability was confirmed across ten random data splits.

Explainability quality was quantified using the Flesch Reading Ease score [6], a standardised readability metric scored on a 0–100 scale. Scores in the range 70–80 correspond to text readily understood by a 7th-grade student, scores of 60–70 indicate standard readability accessible to an 8th–9th grade reader, scores of 50–60 reflect content suited to a high school student, scores of 30–50 indicate college-level difficulty, and scores below 30 describe highly technical or professional text typically requiring advanced education to comprehend. This metric was used to measure the accessibility of LLM-generated explanations for non-expert users.

5. Results and Analysis

5.1. ML Baseline

The SVM model achieved 98.37% accuracy on the held-out test set (F1: 0.9781, ROC-AUC: 0.9986) with training time of 3.13 seconds and inference at 0.34 ms per email. On the independent hybrid test dataset, the model processed all 16.478 emails in 5.55 seconds, achieving 90.30% accuracy (1.598 misclassifications), reflecting realistic domain shift between training and deployment data.

5.2. Selective Error Correction

The 1.598 misclassified emails were ranked by uncertainty and the 250 most uncertain cases were selected for LLM review. In this evaluation, ground truth labels were available to identify confirmed misclassifications; in production deployment, the system would instead route all predictions below a configurable confidence threshold to LLM review, with the optimal threshold determined empirically per deployment context. Table 2 presents the correction results across all three providers, and Figure 3 visualises the accuracy and number of mistakes corrected.

Table 2. LLM Error Correction on 250 Most Uncertain ML Misclassifications.

Model	Accuracy	Fixed	Avg Time/Email
Claude 3.5 Sonnet	**93.60%**	**234/250**	2.06 s
GPT-4	68.00%	170/250	**0.67 s**
DeepSeek	57.94%	135/250	14.46 s

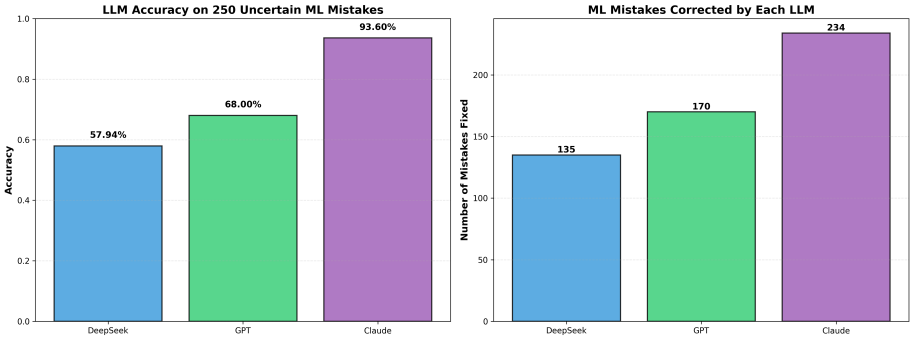

Figure 3. LLM accuracy and number of mistakes corrected on 250 most uncertain ML misclassifications.

Claude 3.5 Sonnet achieved the highest correction accuracy (93.60%, 234/250), while GPT-4's lower accuracy (68%) is offset by its 3× faster response time, making

it preferable for latency-sensitive deployments. The wide performance gap (57.94–93.60%) confirms that semantic reasoning about deceptive intent varies substantially across LLM architectures [16].

5.3. Multilingual Translation

The multilingual dataset was designed for representative coverage across diverse language families rather than volume, balancing evaluation breadth against the per-call API costs inherent to LLM-based processing. Table 3 reports classification performance across all translation configurations, and Figure 4 compares F1 scores across providers.

Table 3. Multilingual Classification Performance (36 Emails, 18 Languages).

Approach	Accuracy	Recall	F1	Time/Email
ML Only (Raw)	50.0%	66.7%	0.571	6.83 ms
ML + GPT-4	63.9%	83.3%	0.698	**847 ms**
ML + Claude 3.5	58.3%	88.9%	0.681	2.033 ms
ML + DeepSeek	**75.0%**	**94.4%**	**0.791**	2.611 ms

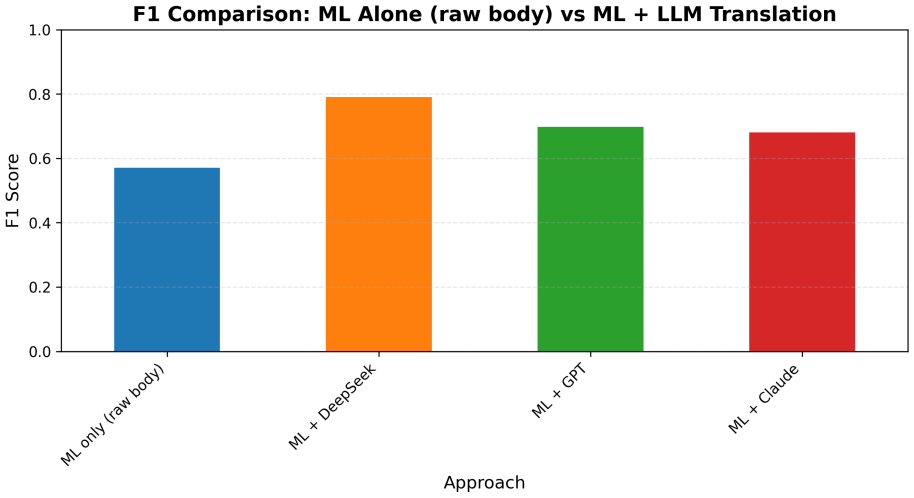

Figure 4. F1 score comparison: ML alone vs ML with LLM translation across three providers.

Without translation, the ML baseline scored 50% – equivalent to random guessing. DeepSeek achieved the best results (75% accuracy, 94.4% recall, F1 0.791), a 38% relative F1 improvement over baseline, while GPT-4 offered a favorable speed–accuracy trade-off at 847 ms per email. High recall is critical: missed phishing emails reach users [27].

5.4. Explainability Enhancement

This evaluation prioritised qualitative depth and comparison of explanation quality across providers over large-scale quantitative evaluation, reflecting practical API cost constraints for a three-provider comparison. Table 4 presents the readability scores and latency of LLM-generated explanations across all three providers, and Figure 5 visualises the mean Flesch scores.

Table 4. Readability of LLM-Generated Explanations (5 Emails).

Model	Mean Flesch	Std Dev	Time/Explanation
DeepSeek	**65.59**	6.65	6.72 s
GPT-4	62.70	14.67	**4.15 s**
Claude 3.5 Sonnet	32.13	16.05	7.04 s

DeepSeek achieved the highest mean Flesch Reading Ease score (65.59, 8th–9th grade) with the lowest variance (6.65), indicating consistent accessibility across email types. GPT-4 produced comparable readability (62.70) at the lowest latency (4.15 s). Claude's low score (32.13, college level) suggests it prioritizes thoroughness over simplicity. All three providers successfully transformed LIME feature-weight pairs into plain-language warnings [21].

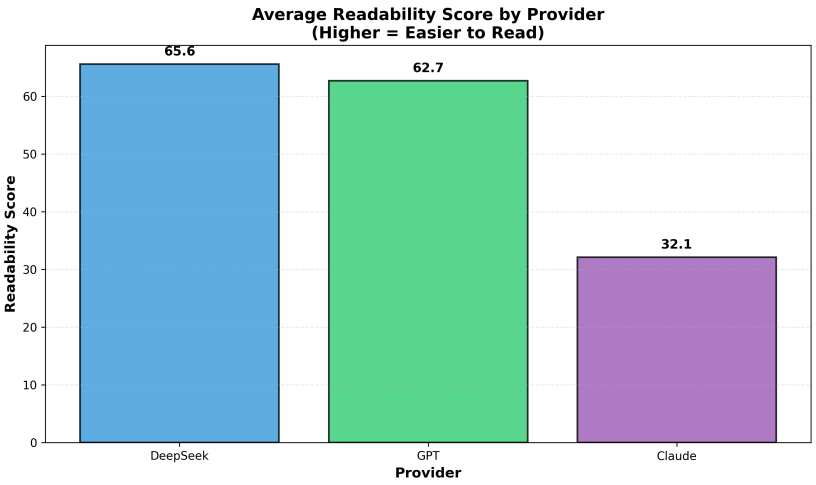

Figure 5. Average Flesch Reading Ease scores by LLM provider. Higher scores indicate more accessible explanations for non-expert users.

6. Discussion and Conclusion

Architectural Effectiveness

Routing only 1.5% of emails (250 of 16.478) to LLM review preserves ML-level throughput while recovering the most difficult misclassifications. The results confirm the core hypothesis: email classification exhibits heterogeneous difficulty, and restricting LLM invocation to calibrated failure points delivers LLM-level accuracy at a fraction of the cost.

Task-Specific LLM Recommendations

No single provider dominates all tasks: Claude 3.5 Sonnet leads on error correction (93.60%), DeepSeek on translation (75%, F1 0.791) and readability (65.59 Flesch), and GPT-4 on speed across all tasks. The modular architecture allows independent provider selection per task and naturally extends to a multi-agent collaboration paradigm, where specialized agents (one optimized for classification correction, one for translation, one for explanation generation) operate in parallel under a lightweight routing controller [13].

Future Work

Future work should validate thresholding in production, compare LLM translation against language-specific ML models, run user studies on explanation comprehension, and assess prompt injection resilience. Although the current system does not explicitly analyse URLs or attachments, LLMs may still recognise suspicious URL patterns through their broad pretraining knowledge, even without task-specific training on phishing URLs [16]; incorporating raw URL strings and attachment metadata into the LLM prompt, and potentially fine-tuning on phishing-specific datasets, would further strengthen this capability. As selective hybrid architectures of this kind are novel, the present work prioritised breadth of exploration across three distinct tasks and three LLM providers rather than depth of scale within any single task; larger-scale multilingual and explainability evaluations represent a natural next step as API costs continue to decrease and access to richer phishing datasets improves.

Conclusion

The proposed architecture demonstrates that targeting LLMs only at ML failure points delivers LLM-level accuracy at a fraction of the cost. This provides a practical path to robust, explainable, and multilingual phishing detection. The implementation is publicly available [4].

Acknowledgements

This work has been supported by the TalentUD Talent Management Programme of the University of Debrecen, Hungary.

References

[1] A. A. Alsuwaylimi: *Enhancing Arabic phishing email detection: A hybrid machine learning based on genetic algorithm feature selection*, International Journal of Advanced Computer Science & Applications 15.8 (2024), doi: 10.14569/IJACSA.2024.0150826.

[2] P. An, R. Shafi, T. Mughogho, O. A. Onyango: *Multilingual email phishing attacks detection using OSINT and machine learning*, arXiv preprint (2025), arXiv: 2501.08723.

[3] Artificial Analysis: *Artificial Analysis: Independent Analysis of AI Models and API Providers*, Online platform, Accessed: 2025, 2024, url: https://artificialanalysis.ai/.

[4] F. Aziz: *Hybrid email classifier research: LLM-augmented machine learning for phishing detection*, GitHub repository, Accessed: 2025, 2025, url: https://github.com/FaroukAziz2020/Hybrid-Email-Classifier-Research.

[5] P. Bountakas, K. Koutroumpouchos, C. Xenakis: *A comparison of natural language processing and machine learning methods for phishing email detection*, in: Proceedings of the 16th International Conference on Availability, Reliability and Security (ARES 2021), New York, NY, USA: ACM, 2021, pp. 1–12, doi: 10.1145/3465481.3469205.

[6] I. Chanis, A. Arampatzis: *Enhancing phishing email detection with stylometric features and classifier stacking*, International Journal of Information Security 24.1 (2025), p. 15, doi: 10.1007/s10207-024-00928-7.

[7] 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 (2025), pp. 1–33, doi: 10.1145/3733049.

[8] Duong Tin: *Phishing Emails*, Hugging Face dataset, Accessed: 2025, 2024, url: https://huggingface.co/datasets/duongtino/phishing-email.

[9] Ethan Nguyen: *Phishing Email Collection*, Kaggle dataset, Accessed: 2025, 2024, url: https://www.kaggle.com/datasets/ethanhnguyen/phishing-email-collection.

[10] T. S. Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, P. Flach: *Classifier calibration: A survey on how to assess and improve predicted class probabilities*, Machine Learning 112.9 (2023), pp. 3211–3260, doi: 10.1007/s10994-023-06336-7.

[11] T. Fontanari, T. C. Fróes, M. Recamonde-Mendoza: *Cross-validation strategies for balanced and imbalanced datasets*, in: Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS 2022), Cham: Springer, 2022, pp. 626–640, doi: 10.1007/978-3-031-21686-2_43.

[12] T. Gangavarapu, C. D. Jaidhar, B. Chanduka: *Applicability of machine learning in spam and phishing email filtering: Review and approaches*, Artificial Intelligence Review 53.7 (2020), pp. 5019–5081, doi: 10.1007/s10462-020-09814-9.

[13] G. Goldenits, P. Koenig, S. Raubitzek, A. Ekelhart: *Small language models for phishing website detection: Cost, performance, and privacy trade-offs*, arXiv preprint (2025), arXiv: 2511.15434.

[14] E. S. Gualberto, R. T. D. Sousa, T. P. D. B. Vieira, J. P. C. L. D. Costa, C. G. Duque: *The answer is in the text: Multi-stage methods for phishing detection based on feature engineering*, IEEE Access 8 (2020), pp. 223529–223547, doi: 10.1109/ACCESS.2020.3044229.

[15] J. Hazell: *Large language models can be used to effectively scale spear phishing campaigns*, arXiv preprint (2023), arXiv: 2305.06972.

[16] 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*, arXiv preprint (2024), arXiv: 2412.00586.

[17] I. Kanaris, K. Kanaris, I. Houvardas, E. Stamatatos: *Words versus character n-grams for anti-spam filtering*, International Journal on Artificial Intelligence Tools 16.6 (2007), pp. 1047–1067, doi: 10.1142/S0218213007003692.

[18] P. Kumar, C. Kushwaha, A. K. Rajpoot: *Advances in Phishing Detection: A Comprehensive Review of Machine Learning, Deep Learning, and Transformer-based Approaches*, SSRN preprint (2025), Available at SSRN, url: https://ssrn.com/abstract=6079506.

[19] B. A. Kumara, M. M. Kodabagi, T. Choudhury, J.-S. Um: *Improved email classification through enhanced data preprocessing approach*, Spatial Information Research 29.2 (2021), pp. 247–255, doi: 10.1007/s41324-020-00349-5.

[20] C. Lee: *Enhancing phishing email identification with large language models*, arXiv preprint (2025), arXiv: 2502.04759.

[21] B. Lim, R. Huerta, A. Sotelo, A. Quintela, P. Kumar: *EXPLICATE: Enhancing phishing detection through explainable AI and LLM-powered interpretability*, arXiv preprint (2025), arXiv: 2503.20796.

[22] M. J. Nabet, L. E. George: *Phishing attacks detection by using support vector machine*, Journal of Al-Qadisiyah for Computer Science and Mathematics 15.2 (2023), pp. 180–191, doi: 10.29304/jqcsm.2023.15.21378.

[23] J. Rastenis, S. Ramanauskaitė, I. Suzdalev, K. Tunaitytė, J. Janulevičius, A. Čenys: *Multi-language spam/phishing classification by email body text: Toward automated security incident investigation*, Electronics 10.6 (2021), p. 668, doi: 10.3390/electronics10060668.

[24] K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed, S. Chen, D. Liu, J. Li: *Performance comparison and current challenges of using machine learning techniques in cybersecurity*, Energies 13.10 (2020), p. 2509, doi: 10.3390/en13102509.

[25] S. Tarapiah, L. Abbas, O. Mardawi, S. Atalla, Y. Himeur, W. Mansoor: *Evaluating the effectiveness of large language models (LLMs) versus machine learning (ML) in identifying and detecting phishing email attempts*, Algorithms 18.10 (2025), p. 599, doi: 10.3390/a18100599.

[26] A. A. Tawil, L. Almazaydeh, D. Qawasmeh, B. Qawasmeh, M. Alshinwan, K. Elleithy: *Comparative analysis of machine learning algorithms for email phishing detection using TF-IDF, Word2Vec, and BERT*, Computers, Materials & Continua 81 (2024), pp. 3395–3416, doi: 10.32604/cmc.2024.058217.

[27] W. Zhu, H. Liu, Q. Dong, J. Xu, S. Huang, L. Kong, J. Chen, L. Li: *Multilingual machine translation with large language models: Empirical results and analysis*, in: Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 2765–2781, doi: 10.18653/v1/2024.findings-naacl.176.