The Phish, The Spam, and The Valid: Generating Feature-Rich Emails for Benchmarking LLMs

Rebeka Toth
Department of Informatics
University of Oslo
Oslo, Norway
rebekat@ifi.uio.no

Nils Gruschka
Department of Informatics
University of Oslo
Oslo, Norway
nilsgrus@ifi.uio.no

Tamas Bisztray*
Department of Informatics
University of Oslo
Oslo, Norway
tamasbi@ifi.uio.no

***Abstract*—In this paper, we introduce a metadata-enriched generation framework (PhishFuzzer) that seeds real emails into Large Language Models (LLMs) to produce 23,100 diverse, structurally consistent email variants across controlled entity and length dimensions. Unlike prior corpora, our dataset features strict three-class labels (Phishing, Spam, Valid), provides full URLs and attachment metadata, and annotates each email with attacker intent. Using this dataset, we benchmark two state-of-the-art LLMs (Qwen-2.5-72B and Gemini-3.1-Pro) under both Basic (body, subject) and Full (+URL, sender, attachment) settings. Using formal confidence metrics (Task Success Rate and Confidence Index), we analyze model reliability, robustness to linguistic fuzzing, and the impact of structural metadata on detection accuracy. Our fully open-source framework and dataset provide a rigorous foundation for evaluating next-generation email security systems. To support open science, we make the PhishFuzzer Dataset, the generation scripts, and prompts available on GitHub: https://github.com/DataPhish/PhishFuzzer**

***Index Terms*—phishing, spam, email security, large language models, email detection, emotion analysis, dataset**

I. Introduction

Phishing and spam emails continue to pose a significant threat to cybersecurity, targeting individuals and organizations with deceptive messages designed to steal sensitive information, compromise systems, and facilitate fraudulent activity [1]. According to the ENISA Threat Landscape 2025 report, phishing remains a dominant initial access vector across major incident categories, with continued growth in both volume and sophistication [1].

Cybercriminals increasingly exploit Large Language Models (LLMs) to produce highly convincing and linguistically polished messages at scale. Recent work shows that LLM-generated phishing can achieve click-through rates comparable to human-crafted attacks [2], while traditional detection systems that utilize rule-based heuristics and static features, such as keywords, sender reputation, or metadata, degrade significantly when confronted with LLM-rephrased content [3], [4]. This highlights the need for more robust and adaptive email security solutions.

Meanwhile, existing phishing email datasets are often outdated, lack separate classes for phishing, spam, and valid emails, and lack metadata and linguistic variability required for modern email classification. To bridge this critical gap in dataset quality and model evaluation, we make the following contributions:

- **PhishFuzzer**: An LLM-based generation pipeline that uses real-world emails as templates to produce structurally consistent synthetic variants.
- **PhishFuzzer Dataset**: The first open-source, three-class (Phishing, Spam, Valid) email dataset comprising 3,300 real seeds and 19,800 synthetic variants, complete with attacker motivation annotations, structural metadata, and strict provenance tracking.
- **Rigorous LLM Benchmarking:** A comprehensive evaluation of Qwen-2.5-72B and Gemini-3.1-Pro demonstrating that while LLMs achieve high zero-shot phishing detection, their performance is sensitive to the inclusion of metadata and they struggle with the subjective boundary between spam and valid email.
- **Systematic Failure Analysis:** Through the introduction of the Total Flip Score (TFS@K) metric, we isolate model blind spots and inaccuracies in human labels in legacy public datasets.

This paper is structured as follows: Section II discusses related work, Section III presents our methodology, Section IV presents our results, Section V discusses limitations and future work, while Section VI summarizes and concludes our work.

II. Related Literature

A. Existing Email Datasets

Table I summarizes the properties of existing open-source email corpora, highlighting several critical shortcomings. Most widely used datasets were collected before 2010 and often suffer from data quality issues, including encoding inconsistencies, malformed characters, and residual HTML artifacts. Because of their age, they do not reflect the improved linguistic quality and structural sophistication seen in modern AI-assisted campaigns [2], [4]. Furthermore, some legacy datasets strip important real-world signals – such as full URL structures, attachment names, and sender-domain features – and provide no annotations of the attacker's underlying intent.

*Corresponding author: tamasbi@ifi.uio.no

TABLE I
COMPARISON OF PHISHING DATASETS BASED ON GRANULARITY, ORIGIN, AND AVAILABILITY.

Dataset	Size	Year	Classes	Metad.	Intent	Real/Syn.	Open
SpamAssas. [5]	6k	2003	S&V*	N	N	Real	Y
Enron [6]	517k	2004	–	Y	N	Real	Y
Nazario [6]	11k	≤21	P	N	N	Real	Y
CEAS-08 [6]	39k	2008	S&V	N	N	Real	Y
Nig.Fraud [6]	4k	≤07	P	N	N	Real	Y
Codebook [7]	503	≤21	P	N	Y**	Real	N
E-PhishLLM [4]	16k	2025	P&V	N	N	Syn	Y
PhishFuzzer	23k	2024-26	P&S&V	Y	Y	Both	Y

V = valid emails, S = spam, P = phishing
*Multiple difficulty levels are defined, but they are arbitrary.
**Unclear if dataset is annotated, not released.

To address the lack of modern data, Pajola proposed E-PhishGEN, a framework for generating fully synthetic phishing and valid emails from LLM-created user profiles [4]. Their cross-dataset evaluation revealed that classical Machine Learning (ML) approaches trained on legacy corpora and tested on E-PhishLLM experienced accuracy drops of up to 40 percentage points, signaling that legacy datasets are insufficient proxies for LLM-generated phishing content. However, their cross-training experiments were limited to classical ML pipelines, omitting more advanced transformer-based models. Furthermore, E-PhishLLM lacks the structural granularity, such as explicit URL strings and attachment names, that can be critical for both human and algorithmic decision-making.

B. LLM-Based Email Classification

Recent research has begun to examine phishing intent. Eilertsen [8] proposed an intent-based phishing taxonomy derived from MITRE ATT&CK T1566, categorizing emails as *Phishing via Link*, *Attachment*, *Service*, or *Other*. While highly valuable for threat modeling, this categorization relies entirely on manual annotation.

Saka [7] manually categorized 503 emails from the Nazario dataset based on the specific actions requested from the victim—'click', 'download', 'reply/email', 'call', 'other', and 'none'—however, they also did not provide an automated or LLM-based method to extract these labels at scale.

A growing amount of work has benchmarked LLMs for phishing and spam detection, with GPT-4o, Gemini 1.5, Llama-3.1, and Mistral-Large matching or exceeding fine-tuned transformer baselines [2], [9]–[12]. In parallel, studies on LLM-generated phishing show that GPT-4-crafted emails rival human-written attacks in persuasiveness [2]. Afane [3] found that detection accuracy drops substantially under LLM-based rephrasing when tested against traditional phishing detectors (e.g., Gmail Spam Filter, Apache SpamAssassin, Proofpoint) as well as classical machine learning models like SVM and Logistic Regression. These findings underscore a critical shift: as traditional filters fail against LLM-rephrased attacks, robust detection will increasingly rely on advanced architectures, ranging from fine-tuned local transformers (e.g., BERT) to zero-shot reasoning models (e.g., GPT-4o). To effectively train and benchmark these systems, the community requires realistic, large-scale datasets that preserve structural metadata and attacker intent. Our generation methodology directly addresses this need by resolving the fundamental shortcomings of existing approaches highlighted in Table I.

III. METHODOLOGY

This section provides an overview of both the dataset creation pipeline, and the LLM benchmarking approach. Figure 1 provides an overview of the four steps of the dataset creation.

A. Ethical Guidelines

This research was conducted in strict accordance with the ethical guidelines of the University of Oslo and Norwegian national regulations. All participants provided informed consent, and the submitted data underwent strict manual de-identification to ensure full anonymity.

B. Step 1: Seed Dataset Creation

The seed dataset combines two complementary sources:

Manually Curated Private (N = 300): Constructed from anonymized emails voluntarily submitted from personal and corporate inboxes, balanced across phishing (including enterprise phishing awareness campaigns [13]), spam, and legitimate (100 each). Each was annotated with subject, body, sender, all extracted URLs, and attachment filenames with extensions.

Aggregated Public (N = 3,000): Additional emails were sampled from multiple publicly available datasets [5], [6]. To mitigate the limitations of older corpora, we prioritized samples from the most recent portions of these datasets where available, while ensuring class balance and structural diversity across phishing, spam, and legitimate categories. This multi-source sampling strategy was chosen to maximize variability in content and writing styles, while retaining sufficient data quality for downstream processing. These datasets follow a coarser schema: URLs appear only when present as plaintext in the body, and attachments are recorded as binary flags rather than by filename. Steps 2–3 address this reduced granularity.

Combined Dataset (N = 3,300): Each email is represented as a structured record containing the following fields: *Subject*, *Body*, and *Sender* (textual content); *URL* and *File* (lists capturing embedded links and attachment filenames); *Type* (Phishing, Spam, Valid) and *Motivation* (primary requested action); *Source* and *Created by* (data provenance); *Year* (temporal context); and *Original_ID* (linking all variants to their seed template). A *Language* tag is assigned to each seed email based on its subject and body and propagated unchanged to all generated variants.

The dataset is predominantly English (96.79%), with additional multilingual coverage including Norwegian (0.88%), Japanese (0.45%), French (0.36%), Portuguese (0.33%), and several other low-frequency languages, reflecting the heterogeneous nature of real-world email communication.

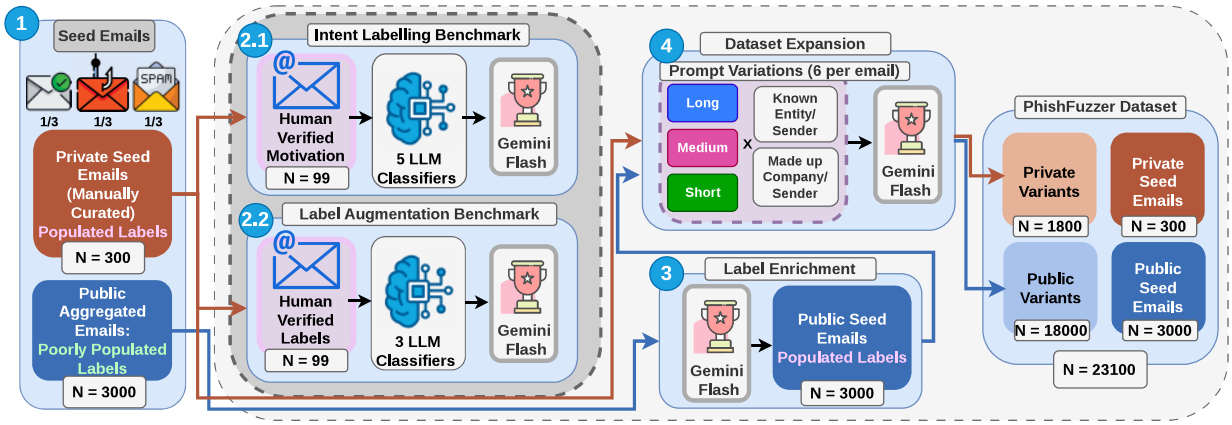

Fig. 1. Methodology for the PhishFuzzer Framework

C. Step 2.1: Intent Benchmarking

We benchmarked LLMs at identifying the *primary explicitly requested user action* in an email, not merely the presence of an artifact. Capturing attacker intent enables modeling of the underlying behavioral objective of an email (e.g., inducing a click or response), providing a more robust signal than surface-level features such as keywords or the mere presence of links and attachments.

A subset of 99 emails (33 per class) was independently labeled by two domain experts with one of four intent categories: *Follow the link*, *Open attachment*, *Reply*, or *Unknown*. Five LLMs (*Claude 3.5 Sonnet*, *GPT-5.2-Chat*, *Gemini-2.5-Flash*, *Qwen 2.5-7B-Instruct*, *DeepSeek-Chat*) each labeled all 99 emails five times ($k = 5$, temperature 0). Accuracy and internal consistency were measured using a confidence metric adapted from [14].

The prompt enforces a strict hierarchical decision process and structured JSON output to ensure consistent labeling across runs. The exact prompting template is available in the repository.[1]

D. Step 2.2: Label Augmentation Benchmarking

To populate the missing URL and attachment fields in the aggregated subset, we benchmarked three LLMs (*Claude 3.5 Sonnet*, *GPT-5.2-Chat*, *Gemini-2.5-Flash*) on the same 99 benchmark emails. Models were instructed to populate missing fields according to motivation-aligned structural rules: emails labeled "Follow the link" required a non-null URL, and "Open attachment" required a filename. Category-specific constraints governed domain plausibility.

For phishing variants, only deceptive look-alike or invented domains were permitted, assuming that Sender Policy Framework (SPF) checks would otherwise flag the use of official domains. Conversely, for spam and legitimate variants, real official domains were allowed, accurately reflecting real-world marketing and promotional practices. The full prompting strategy and constraint set are detailed in our repository. Outputs were evaluated on quantitative structural correctness and contextual plausibility via manual review. *Gemini-2.5-Flash* was selected based on the highest structural reliability and lowest hallucination rate.

E. Step 3: Label Enrichment

Using the validated prompting strategy, *Gemini-2.5-Flash* populated missing intent (motivation), URL, and attachment fields across all 3,000 aggregated emails. The enriched set was merged with the 300 manually curated emails, yielding a seed dataset of 3,300 structurally consistent emails.

Class-specific prompt variants were used for phishing, spam, and legitimate emails, with different constraints applied to domains and attachment generation to maintain realism. The exact prompting template is available in the repository.[2]

F. Step 4: Dataset Expansion with Seeding

We expand the dataset using structured LLM-based generation, where each of the $N = 3,300$ seed emails serves as an *email template*. For each template t_i ($i = 1, 2, \ldots, N$), we have $K = 7$ emails (the original seed plus six synthetic variants). We denote the ground-truth label for template t_i by y_i, which is shared across all K instances of t_i.

The six variants ($2 \cdot 3$) are generated along two orthogonal dimensions, as shown in Phase 4 of Figure 1:

- **2 Entity Types:** Globally recognized entities (e.g., Amazon, Disney, Microsoft) versus fabricated but realistic ones.
- **3 Length Types:** Short (4–8 sentences), medium (10–16 sentences), and long (25–40 sentences).

Six distinct prompt templates cover each entity-length combination, strictly preserving the original email's intent and structural constraints. Consistent with Step 2.2, phishing variants require deceptive domains for well-known entities, while spam and legitimate emails allow realistic corporate domains.

[1]https://github.com/DataPhish/PhishFuzzer/blob/main/DataSet_Creation/prompts.md#1-intent-labeling-prompt

[2]https://github.com/DataPhish/PhishFuzzer/blob/main/DataSet_Creation/prompts.md#2-populating-the-dataset-prompt-structural-augmentation

This generation process inherently sanitizes encoding artifacts and residual HTML from the aggregated seeds, yielding clean plaintext variants. Additionally, non-English variants were retained in their original languages. This process produced 19,800 synthetic variants, yielding a final evaluation dataset of 23,100 emails.

The full prompting strategy is available in the repository.[3]

G. Evaluation and Metrics

Qwen-2.5-72B and Gemini-3.1-Pro are evaluated using a single inference pass per email. A fixed classification prompt enforces a single-label output (*phishing*, *spam*, or *valid*) without explanation. The exact prompt is available in the repository.[4] To measure classification reliability across linguistic variations, we group the predictions by their originating template. To measure classification reliability across linguistic variations, we group the predictions by their originating template. Let $q_{i,j}$ denote the predicted label for the j-th instance ($j = 1, \ldots, K$) of template t_i. We adapt the metrics introduced by Tihanyi [14]:

The **Task Success Rate** (TSR) counts the number of correctly classified variants for a given template t_i such that:

$$\mathrm{TSR}(t_i, K) = \sum_{j=1}^{K} I\big(q_{i,j} = y_i\big), \tag{1}$$

where $I(\cdot)$ is the indicator function returning 1 if its argument is true and 0 otherwise, such that $0 \leq \mathrm{TSR}(t_i, K) \leq K$.

The **Confidence Index** ($\mathrm{Conf@}K$) measures the percentage of templates that are perfectly classified across all K variants:

$$\mathrm{Conf@}K = \frac{100}{N} \sum_{i=1}^{N} I\big(\mathrm{TSR}(t_i, K) = K\big). \tag{2}$$

We additionally introduce the **Total Flip Score** ($\mathrm{TFS@}K$), which counts the absolute number of templates where the model consistently fails across all K variants:

$$\mathrm{TFS@}K = \sum_{i=1}^{N} I\big(\mathrm{TSR}(t_i, K) = 0\big). \tag{3}$$

By isolating $\mathrm{TSR}(t_i, K) = 0$ cases, $\mathrm{TFS@}K$ pinpoints threat vectors that are reliably evasive, distinguishing fundamental model blind spots from mere stochastic errors.

The two LLMs are evaluated across four dimensions:

1) **Label Configuration:** Three-class classification (Phishing vs. Spam vs. Valid).
2) **Prompting Strategy:** *Basic* (subject, body) vs. *Full* (+sender, URLs, filenames).
3) **Dataset Condition:** *Original* seed emails vs. LLM-*Generated* variants.
4) **Source Type:** *Private* inbox collections vs. *Public* datasets.

[3]https://github.com/DataPhish/PhishFuzzer/blob/main/DataSet_Creation/prompts.md#3-rephrasing--dataset-expansion

[4]https://github.com/DataPhish/PhishFuzzer/blob/main/DataSet_Creation/prompts.md#4-classification-prompt

IV. Results

A. Intent and Label Augmentation Benchmark

Before expanding the dataset, we benchmarked LLMs on two tasks using 99 manually validated emails: (1) inferring attacker intent and (2) populating missing structural metadata (URLs and attachments).

Intent Labeling: Each model processed the benchmark $r = 5$ times per email. To make a single label, we prioritized motivation hierarchically: requests to "open attachment" superseded "follow link". We report *Strict Accuracy* (all 5 runs match ground truth), and *Consistency* (giving the same answer regardless of correctness).

TABLE II
Intent Label Results (99 emails, $r = 5$ runs).

Model	Strict Accuracy	Consistency
Claude 3.5 Sonnet	0.9091	0.9838
GPT-5.2-chat	0.8788	0.9798
Gemini 2.5 Flash	**0.9798**	**1.0000**
Qwen 2.5-7B	0.6162	0.9434
DeepSeek Chat	0.8586	0.9939

As shown in Table II, Gemini-2.5-Flash achieved the strongest performance, reaching 97.98% strict accuracy with perfect internal consistency.

Structural Label Augmentation. To populate missing fields in legacy emails, we evaluated Gemini 2.5 Flash, GPT-5.2-chat, and Claude 3.5 Sonnet. Quantitative checks verified logical consistency: emails labeled "follow link" required a generated URL, and "open attachment" required a non-empty filename. Manual review by two experts assessed the plausibility of the generated artifacts. Claude and GPT frequently hallucinated placeholders, generating text such as `[insert link here]` instead of a genuine URL, based on the email's narrative. Gemini-2.5-Flash consistently generated realistic, context-aware domains and filenames.

B. Provenance: Private, Public, and their rephrased variants

We track classification accuracy across four distinct data subsets: Private Seeds ($N = 300$), Public Seeds ($N = 3,000$), Private Variants ($N = 1,800$), and Public Variants ($N = 18,000$). Figure 2 presents these comparisons for Qwen-2.5-72B and Gemini-3.1-Pro, respectively, evaluating both under Basic (body and subject only) and Full (including metadata) prompting configurations.

Private vs. Public: Examining just the Basic setting, Qwen-2.5-72B performs significantly worse on private seed. In contrast, Gemini-3.1-Pro maintains a uniform baseline accuracy across both private and public corpora prior to metadata addition.

Metadata on Provenance: At first glance, the introduction of metadata benefits private emails more than public ones. However, overall accuracy scores can mask underlying class-level shifts, in which the recall or F1 score of one class might improve while another drops. **When we investigate the confusion matrices, new insights will be gained**. Until then,

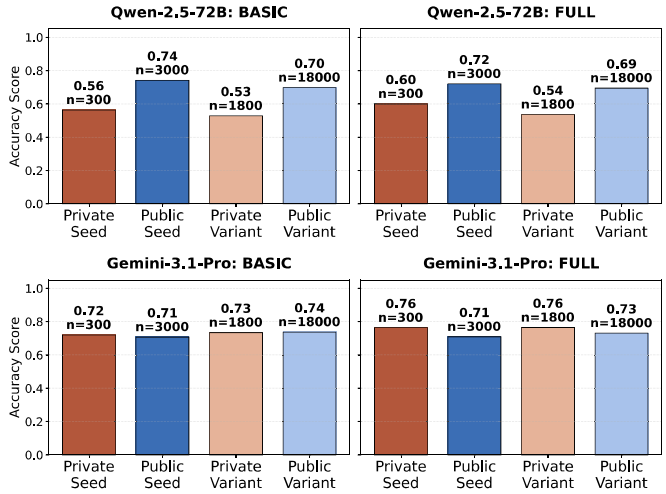

Fig. 2. Provenance Analysis: Comparison of classification accuracy between BASIC and FULL prompts.

we cannot draw definitive conclusions from these aggregated figures about whether the semi-synthetic metadata for Public seed, or the fully synthetic metadata for Private variants and Public variants, improves or degrades classification decisions.

Variants: On the other hand, rephrasing the emails (the variants) had a noticeably different impact on Qwen-2.5-72B and Gemini-3.1-Pro. For Qwen, performance drops on variants, suggesting that the "fuzzing" is successful and using different wording confuses the classifier, regardless of whether metadata is present. Gemini, on the other hand, shows the opposite: performance on variants improves by 0–2 pp., suggesting that once it understands the email's logic, it stays more consistent with its classification.

C. 3-Class Performance

While aggregate accuracy suggests stable performance (Figure 2), the confusion matrices (Figures 3) and corresponding F1 scores (Table III) reveal severe class-specific imbalances and a distinct trade-off when metadata are introduced.

Phishing Detection: Qwen-2.5-72B acts as a "relaxed" classifier, initially misclassifying 1,190 phishing emails as Valid (BASIC); adding metadata reduces this to 882, raising its Phishing F1 from 0.893 to 0.917. Gemini-3.1-Pro demonstrates a stricter posture, misclassifying only 33 phishing emails as Valid, dropping to 13 with metadata, achieving an exceptional Phishing F1 of 0.958.

Spam Detection: Both models struggle to isolate Spam, as reflected by the low Spam F1 scores (≤ 0.428). Under the BASIC prompt, Qwen-2.5-72B misclassifies 5,419 Spam emails as Valid. Gemini-3.1-Pro similarly marks 4,930 Spam emails as Valid while confusing 434 for Phishing. Counterintuitively, introducing metadata increases these errors. It severely impacts models' spam sensitivity, driving Qwen-2.5-72B's already low Spam Recall from 25.11% to 19.06% and Gemini-3.1-Pro's from 28.65% to 24.74%, with the vast majority of these false negatives absorbed by the Valid class.

Valid Classification: Qwen-2.5-72B's "relaxed" nature benefits legitimate emails, incorrectly flagging only 29 Valid as Phishing. Conversely, Gemini-3.1-Pro's "paranoid" posture penalizes legitimate traffic, misclassifying 398 Valid emails as Phishing in the BASIC setting. However, metadata provides a crucial corrective signal, reducing Gemini's false positives to 215.

The Metadata Trade-off: The integration of structural metadata forces a zero-sum shift in model behavior. While it is highly beneficial for isolating true phishing threats and correcting Valid false positives, it systematically pushes borderline Spam into the Valid category. This specific failure mode causes the overall Macro F1 score to decline for both models (e.g., Qwen-2.5-72B drops from 0.655 to 0.636) despite metadata improving strict phishing detection.

D. Reliability and Systematic Failure Modes

1) Absolute Blind Spots – Analysis of Systematic Failures: To investigate the nature of model weaknesses, we analyze the error distributions of "Blind Spots": templates where a model failed to correctly classify the original seed and all six variants ($TFS@7 = 7$). The resulting Systematic TFS matrices (Figure 3) reveal distinct security postures across these consistently misclassified families.

Actual Label – Phishing: Qwen-2.5-72B misclassified 43 Phishing templates as Valid, though this failure more than halves when metadata is added. Gemini-3.1-Pro, under the Basic setting, exhibits consistent blind spots only when Phishing is pushed into Spam. However, adding metadata pushed one Phishing template from Spam into Valid. Upon manual OSINT investigation, this email (`No. 388`, sourced from the Nazario Dataset) is a legitimate inquiry from an MIT Lincoln Laboratory researcher. We retain the original label for this experiment, but we have corrected it for the published dataset for all variants.

TABLE III
CLASSIFICATION PERFORMANCE AND TEMPLATE RELIABILITY METRICS ACROSS MODELS AND PROMPTS.

		Classification Metrics ($N = 23,100$)				**Template Reliability** ($N = 3,300$ **Templates**)			
Model	**Prompt**	**Accuracy**	**Macro F1**	**F1 (Phish)**	**F1 (Spam)**	**#Conf@7**	**Conf@7 (%)**	**#TFS@7**	**TFS@7 (%)**
Qwen-2.5-72B	BASIC	0.688	0.655	0.893	0.387	1,745	52.88%	510	15.45%
Qwen-2.5-72B	FULL	0.684	0.636	0.917	0.310	1,826	55.33%	575	17.42%
Gemini-3.1-Pro	BASIC	0.733	**0.694**	0.939	**0.428**	1,998	60.55%	**467**	**14.15%**
Gemini-3.1-Pro	FULL	**0.731**	0.685	**0.958**	0.383	**2,075**	**62.88%**	491	14.88%

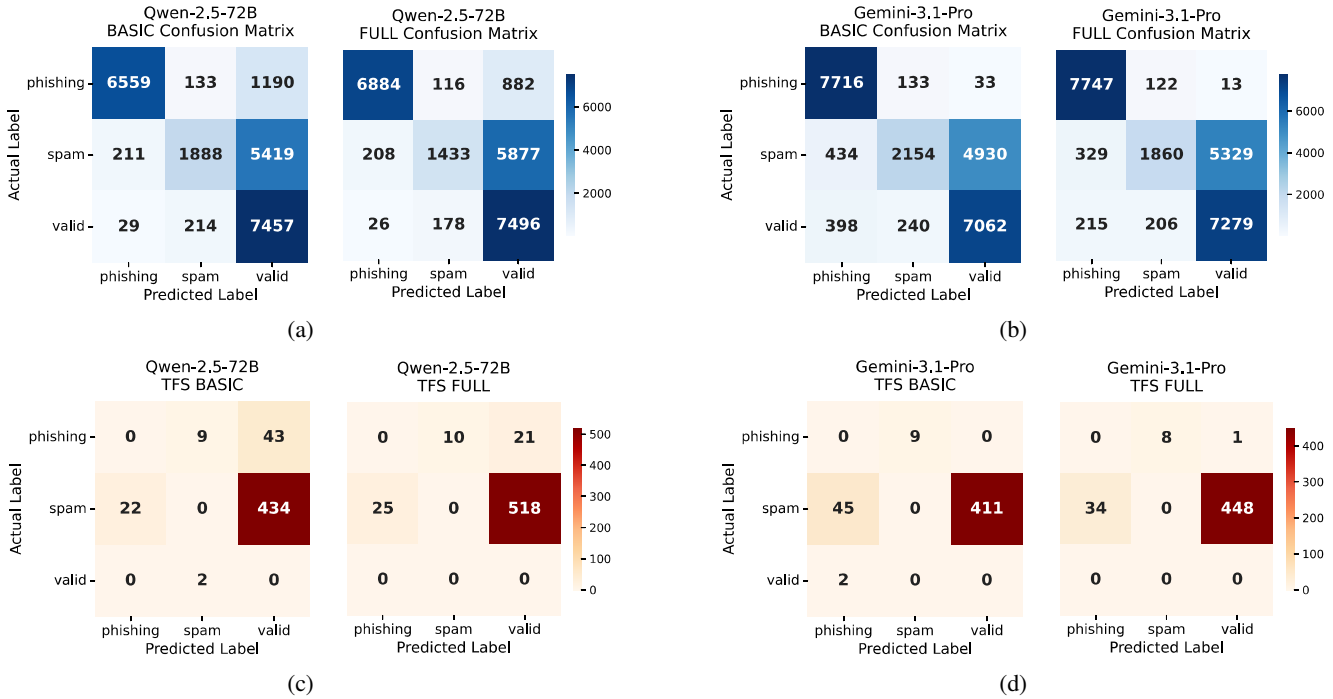

Fig. 3. Confusion matrices (a and b), Total Flip Score matrices (c and d) under the Basic and Full prompting strategies.

Actual Label – Spam: Both models systematically misclassify Spam templates as Valid, as there is an inherently thin boundary between the two classes. In practice, modern inboxes categorize legitimate email into subcategories (Promotions, Updates, Forums) and reserve Spam for aggressive marketing or gray-area communications that do not fit elsewhere. The models' struggles reflect this real-world ambiguity.

Actual Label – Valid: Systematic misclassification of Valid emails is exceptionally rare. In the Basic setting, Qwen-2.5-72B completely misclassified two templates as Spam, while Gemini-3.1-Pro misclassified two as Phishing. Crucially, introducing metadata eliminated these systematic blind spots for both models, leaving the Valid row empty.

2) Model Confidence: Table III demonstrates that Gemini-3.1-Pro has stronger semantic understanding, allowing it to better interpret a template's logic and remain consistent across variants. Overall, Full prompts containing metadata result in higher model consistency for both classifiers. However, while Qwen-2.5-72B is generally less confident in its correct predictions, it is paradoxically more *confidently wrong* than Gemini-3.1-Pro, systematically failing on 575 templates compared to Gemini's 491 under the Full setting.

V. Limitations and Threats to Validity

This study only benchmarked two LLMs in a zero-shot setting. Future work should compare these results with fine-tuned encoder models, such as BERT, and with traditional machine-learning classifiers to assess their efficacy for three-class attribution. While structural constraints were enforced, the 19800 synthetic variants were not human-validated at scale. Because these variants were generated by an LLM (Gemini-2.5-Flash), downstream evaluation by other LLMs risks introducing implicit evaluation bias. Specifically, classifiers may leverage shared underlying linguistic distributions for the different classes. However, our methodology mitigates this risk by grounding the generation process in real-world seed emails, which preserves inherent linguistic diversity and structural authenticity. Furthermore, the underlying framework is highly extensible, allowing researchers to apply it to new corpora to synthesize diverse variants for stress-testing future classification techniques.

The observed trade-off between phishing detection and spam classification reflects the inherent ambiguity between these classes, where the distinction between unsolicited and legitimate communication is often context-dependent.

VI. Conclusion

This study evaluates the reliability and systematic failure modes of LLMs in email security classification, focusing on model reliability and the influence of contextual metadata. We distill our findings by formalizing and answering two primary research questions regarding model resilience and feature representation.

RQ1: How do state-of-the-art LLMs generalize across unseen private data and rephrased email variants?
Our evaluation reveals a stark contrast between the examined models. Qwen-2.5-72B exhibits strong evidence of data contamination, showing notable accuracy degradation on unseen 2026 emails (both real and synthetic) compared to emails from public datasets. Conversely, Gemini-3.1-Pro maintains robust performance on unseen data. It also shows higher resilience to semantic perturbation (62.88% confidence index vs. Qwen-

2.5-72B's 55.33%), suggesting a reliance on underlying logic rather than brittle lexical patterns. Despite this architectural divergence, both models share a persistent, systematic weakness in differentiating spam from valid emails.

RQ2: What is the impact of incorporating structural email metadata on LLM classification performance? While both models establish strong baselines when relying only on email body and subject, the integration of structural metadata (URLs, sender domains, attachments) decisively enhances phishing detection across architectures—elevating Gemini-3.1-Pro's Phishing F1 to 0.958 from 0.939 and Qwen-2.5-72B's from 0.893 to 0.917, representing an approximate $20 - 30\%$ reduction in classification errors. However, this extended context introduces severe class-specific trade-offs, degrading spam detection and causing both models to more frequently misclassify spam emails as valid. Consequently, while metadata is a necessary catalyst for hardening critical phishing defenses, its depressive effect on the overall Macro F1 score underscores the fundamental semantic challenge LLMs face when navigating the subjective boundary between spam and valid communication.

Future work should investigate systematically the failure cases identified by the TFS metric and address labeling inconsistencies in legacy public datasets. Additionally, research must move beyond the strict three-class taxonomy to encompass granular, real-world inbox categories, such as Promotions, Social, and Updates. As demonstrated by modern commercial filters, accurately disambiguating aggressive marketing from unsolicited spam remains a persistent challenge regardless of the underlying detection architecture. Finally, evaluating models under adversarial prompt injections when embedded in email bodies, and optimizing the computational overhead of LLMs for real-time email gateways represent critical next steps for the field.

References

[1] ENISA, "Enisa threat landscape 2025," European Union Agency for Cybersecurity, Tech. Rep., October 2025. [Online]. Available: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2025

[2] F. Heiding, B. Schneier, A. Vishwanath, J. Bernstein, and P. S. Park, "Devising and detecting phishing emails using large language models," *IEEE Access*, vol. 12, pp. 42 131–42 146, 2024.

[3] K. Afane, W. Wei, Y. Mao, J. Farooq, and J. Chen, " Next-Generation Phishing: How LLM Agents Empower Cyber Attackers ," in *2024 IEEE International Conference on Big Data (BigData)*. Los Alamitos, CA, USA: IEEE Computer Society, Dec. 2024, pp. 2558–2567. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/BigData62323.2024.10825018

[4] L. Pajola, E. Caripoti, S. Banzer, S. Pizzi, M. Conti, and G. Apruzzese, "E-phishgen: Unlocking novel research in phishing email detection," in *Proceedings of the 18th ACM Workshop on Artificial Intelligence and Security*, ser. AISec '25. New York, NY, USA: Association for Computing Machinery, 2026, p. 64–76. [Online]. Available: https://doi.org/10.1145/3733799.3762967

[5] Apache SpamAssassin Project, "The SpamAssassin Public Email Corpus," https://spamassassin.apache.org/old/publiccorpus/, 2006, public email corpus for spam filtering research.

[6] N. A. Alam, "Phishing email dataset," 2024. [Online]. Available: https://www.kaggle.com/ds/5074342

[7] T. Saka, R. Jain, K. Vaniea, and N. Kökciyan, "Phishing codebook: A structured framework for the characterization of phishing emails," *arXiv preprint arXiv:2408.08967*, 2024.

[8] E. Eilertsen, V. Mavroeidis, and G. Grov, "Llm-powered intent-based categorization of phishing emails," in *2025 IEEE International Conference on Cyber Security and Resilience (CSR)*. IEEE, 2025, pp. 753–758.

[9] J. Zhang, P. Wu, J. London, and D. Tenney, "Benchmarking and evaluating large language models in phishing detection for small and midsize enterprises: A comprehensive analysis," *IEEE Access*, vol. 13, pp. 28 335–28 352, 2025.

[10] K. Mardiansyah and W. Surya, "Comparative analysis of chatgpt-4 and google gemini for spam detection on the spamassassin public mail corpus," March 2024, preprint. [Online]. Available: https://doi.org/10.21203/rs.3.rs-4005702/v1

[11] S. Mahendru and T. Pandit, "Securenet: A comparative study of deberta and large language models for phishing detection," in *2024 IEEE 7th International Conference on Big Data and Artificial Intelligence (BDAI)*, 2024, pp. 160–169.

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

[13] R. Toth, R. A. Dubniczky, O. Limonova, and N. Tihanyi, " Sustaining Cyber Awareness: The Long-Term Impact of Continuous Phishing Training and Emotional Triggers ," in *2025 IEEE International Conference on Big Data (BigData)*. Los Alamitos, CA, USA: IEEE Computer Society, Dec. 2025, pp. 7854–7862. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/BigData66926.2025.11402180

[14] N. Tihanyi, T. Bisztray, R. A. Dubniczky, R. Toth, B. Borsos, B. Cherif, R. Jain, L. Muzsai, M. A. Ferrag, R. Marinelli, L. C. Cordeiro, M. Debbah, V. Mavroeidis, and A. Jøsang, "Dynamic intelligence assessment: Benchmarking llms on the road to agi with a focus on model confidence," in *2024 IEEE International Conference on Big Data (BigData)*, 2024, pp. 3313–3321.