

Cross-Dataset Fragility of Phishing Detectors

Pushkar Kamma^{1*}

Received August 5, 2026

Accepted September 21, 2026

Electronic access October 31, 2026

Machine-learning phishing-email classifiers are typically tested on a held-out partition of their training dataset, which reveals little about their behavior on a different distribution. This test-deployment mismatch is especially relevant to phishing written by large language models (LLMs), widely assumed to remove many of the surface-level clues that legacy detectors use. It is tempting to blame LLM authorship when detectors fail, but that explanation is confounded: LLM email differs from a legacy corpus on two axes, authorship and source corpus. This study separates the two by evaluating frozen classical detectors across independent corpora and, separately, under a narrowed comparison that holds the phishing subtype and the legitimate email fixed. Frozen Random Forest, Multinomial Naive Bayes, and Logistic Regression classifiers trained on a legacy human dataset were evaluated on that dataset, an independent cross corpus human dataset, and an LLM-generated dataset. The Random Forest collapsed to near chance on cross corpus human email while Naive Bayes and Logistic Regression fell below it, and all three held on LLM email. At a frozen threshold the failure appeared as false alarms on legitimate email, not missed phishing. The Random Forest, for example, scored ROC-AUC 0.999 in-distribution but 0.529 and 0.888 respectively. Retraining on a broader mixed corpus recovered performance only on datasets included in that retraining. In the narrowed comparison, machine authorship gave no large measured advantage (ROC-AUC 0.9999 human, 0.9990 LLM, difference 0.0009). In conclusion, single-corpus benchmarks overstated robustness for the three detectors and six corpora tested here.

Keywords: phishing detection, distribution shift, machine-generated email, large language models, corpus shift, Random Forest, ROC-AUC, evaluation methodology

Introduction

Background and context

Phishing is one of the most prevalent cybersecurity attacks, acting as an entry point for account compromise, credential theft, and financial fraud^{1,2}. As email is mostly a text medium, many phishing detection studies rely on classical machine-learning pipelines: usually a term-frequency-inverse-document-frequency (TF-IDF) representation of the text, a set of metadata signals, and a classifier such as Multinomial Naive Bayes or Random Forest³. These studies frequently report high accuracy and F1 on a held-out test set drawn from the same dataset as the training data⁴. These scores make the model appear strong and dependable.

The problem with this assumption, however, is that in-distribution evaluation allows the training and test sets to share similar vocabulary, formatting, header quirks, and collection-time properties^{5,6}. These similarities cause the model to perform extremely well on in-distribution evaluations, but during deployment performance can drop sharply as these similarities disappear⁷. In this new era of LLM-generated phish-

ing the concern is more pressing, as language models produce clean, fluent messages. Commercial models will generate phishing content when prompted to⁸, and it is widely assumed that this fluency removes the surface-level clues that legacy detectors depend on^{9,10}. This has led to a concern that when tested against these cleaner, unfamiliar LLM-generated emails, legacy detectors will lose their accuracy rapidly. From this, it is tempting to blame LLM authorship for breaking legacy detectors.

The authorship and corpus-shift confound

That explanation is confounded. When a detector trained on legacy datasets performs poorly against LLM-generated email, the problem can be attributed to authorship (machine-written email versus human-written email) or corpus shift (the messages come from different datasets, with different topics, formatting, and metadata conventions). Any single comparison between a legacy human corpus and a modern LLM corpus mixes the effects of corpus shift and authorship as the corpora differ in both dimensions simultaneously. Attributing the failure to either factor, authorship or corpus shift, without separating the impact of the two risks blaming the wrong cause.

This study addresses the confound in two separate ways.

¹ Lake Washington High School; Kirkland, USA.

* Email: pushkarkamma@gmail.com

First, an intermediate group of unfamiliar human-written email, the cross corpus human dataset, is placed between the legacy and LLM data, so a comparison between the legacy and this intermediate set changes only the corpus while keeping the authorship constant. This tests whether the failure occurs during this transfer itself. Second, the comparison is narrowed to a single phishing subtype with the legitimate email held fixed, so that authorship varies while those two factors do not.

Objectives

A high score on one test dataset does not necessarily mean that the detector has understood the general notion of phishing emails. It may instead just reflect the lexical and metadata patterns of the training dataset. Whether a detector is generalizing or just memorizing formatting only becomes visible when tested against email drawn from a different distribution. If performance does drop under a certain distribution shift, then this knowledge can be used to determine if that problem can be solved and how. As a single benchmark test cannot reveal this on its own, the objective of this research is to measure whether a detector's in-distribution performance transfers to other corpora, what form any failure takes, whether retraining repairs it and how much machine authorship affects detection once the phishing subtype and legitimate email are held fixed. First, do near-perfect in-distribution scores survive on out-of-distribution email (cross corpus human and LLM) without retraining (RQ1)? Second, what is the failure mode, missed phishing or false alarms, when a detector fails on an out-of-distribution test set, and can a threshold change improve the performance of the detector (RQ2)? Third, does retraining on a broader dataset generalize to unseen corpora, or only recover corpora already in the training pool (RQ3)? Fourth, how large an evasion advantage, if any, does machine authorship give when the phishing subtype and the legitimate email are held fixed (RQ4)?

These research questions are bounded within a deliberately crafted scope. This study was designed to examine the behavior of classical models on public email datasets under a controlled distribution shift. It was not designed to build an optimal or production-ready detection system. All human datasets come from a single decade, which limits temporal drift without eliminating it. Any residual drift is folded into the corpus-shift comparison rather than measured separately.

Significance

Usually, a detector's capabilities and performance are predicted based on its benchmark accuracy measured on held-out email from the same corpus as it was trained on. In production, however, the detector analyzes email from many unfa-

iliar corpora, and studies of phishing at operational scale show how different that traffic is from a curated benchmark¹¹. An in-distribution benchmark does not test the detector's performance on these unfamiliar corpora and instead checks for fit to one narrow dataset, which is why detectors should be evaluated across several independently assembled datasets rather than one random split¹². The performance gap that the detector experiences when testing in-distribution versus out-of-distribution cannot simply be closed by collecting more data, because retraining recovers performance only on already seen corpora while unfamiliar legitimate email still gets flagged. The contribution of this work is that it offers a reusable and reproducible way to evaluate phishing detectors under corpus shift, supported by strict data-cleaning rules and by reporting across-seed variability over the ten random seeds, along with a narrowed test of whether machine authorship degrades detection once the phishing subtype and legitimate email are held fixed. The goal of this study was not to build a new, perfect detector but to show what a single-corpus benchmark fails to reveal.

Methodology overview

Three classical machine-learning models, a Random Forest, a Multinomial Naive Bayes, and a Logistic Regression, were used for the experiment. First, they were trained on the legacy human dataset, frozen, and then tested in-distribution and on the cross corpus human and LLM datasets. Next, they were retrained on a modern mixed corpus and retested to determine whether the degradation could be alleviated or completely repaired. ROC-AUC was the primary metric used to measure performance across all evaluations. Additionally, the effect of machine authorship was tested using a narrowed comparison that held the phishing subtype and the legitimate email fixed.

Methods

Research design

The research was conducted using a frozen-model, out-of-distribution design that measures corpus shift. Each classifier (Multinomial Naive Bayes, Random Forest, and Logistic Regression) was trained on one pool of data and then evaluated on three sets. No final-test examples influenced training, feature extraction, vocabulary construction, the decision threshold, or model settings. The decision threshold was set only on the separate calibration split. The move from the legacy human dataset to the cross corpus human dataset changed the source corpus while maintaining the same human authorship. The move from the cross corpus human dataset to the LLM dataset changed both the corpus and the authorship at the same time, so it does not isolate the effect of authorship and is

treated as a confounded comparison. In both human datasets, the legitimate and phishing email come from different source corpora, so a detector can partly separate the two by recognizing which corpus an email came from rather than whether it is truly phishing. The LLM dataset is different, as both its legitimate and phishing email come from the LLMGen corpus, so source identity alone does not determine the label there. This source confound is measured directly with a source-only probe in the robustness checks and discussed in the Limitations. A model that scores well by keying on an artifact of how a dataset was assembled is a documented failure mode¹³, and one that often becomes visible only under external validation¹⁴. The closest published analogue is a hip-fracture model whose accuracy fell to chance once the confounding variables it had exploited were balanced away¹⁵.

Data

All human corpora except for Nazario were obtained from Kaggle^{16,17}. Legacy Nazario was obtained from monkey.org, a website maintained by the creator of the corpus, rather than from a mirror. To check that the mirrored corpora reproduce their sources faithfully, the Enron and SpamAssassin files were compared against the original archives published by their maintainers, and of the 18,978 messages that could be matched by content the label agreed in every case. SHA-256 hashes and row counts for every input file are published with the code in a public repository, with the URL withheld here to preserve blind review, and every row carries a usable label. The same comparison was not possible for CEAS-08, whose original host no longer distributes the corpus, or for Nigerian-Fraud, whose original repository is no longer reachable. The three datasets had their text normalized and their features extracted by the same pipeline. One difference remains that the pipeline does not remove: the LLM source stores each message with its header lines inside the body, so a From line survives in 94.5% of those messages against 0.2% of the legacy corpus. Repeating the evaluation with those lines removed from the LLM text moved its ROC-AUC from 0.888 to 0.885, so this formatting difference does not account for the result reported below. Spam was excluded from the positive class at the source level rather than by message-by-message review. Three of the corpora contain both legitimate and spam email, and only their legitimate rows were taken: 15,787 of 31,326 rows from Enron, 17,311 of 39,153 from CEAS-08, and 4,089 of 7,863 from SpamAssassin. The positive class was drawn only from the two corpora that contain phishing exclusively, the classic Nazario archive and Nigerian-Fraud, so no spam-labeled message entered it. This means the class labels are inherited from each source dataset and were not independently re-verified. No message-level annotation, blinded review, or inter-rater agreement check was carried out, and the CLAIR

collection's designation of every Nigerian-Fraud message as fraud was accepted as given. This reliance on inherited labels is a limitation of the study. The datasets' composition after cleaning is given in Table 1, and the row counts at each stage of processing are given in Table 2.

Except for the LLM dataset, the rest are all early to mid-2000s human email (Enron 2001¹⁸, Nazario 2004-2007¹⁹, SpamAssassin 2002-2003²⁰, CEAS-08 2008²¹, Nigerian-Fraud 2003-2007²²). Only the LLM dataset²³ was created in 2025, and just the English portion, generated by OpenAI models, was used. The human corpora span 2001 to 2008, so temporal drift is reduced but not removed. The two principal legitimate sources are the furthest apart: Enron (2001) and CEAS-08 (2008), seven years, on the same axis that produces the false alarms reported below. The transfer gap therefore reflects corpus shift and residual temporal drift together. Some sources lost a substantial number of emails at the data cleaning stage: the legacy Nazario phishing corpus, for example, lost about 27% or about 1,200 emails mostly because of the near-duplicate entry removal described in the anti-leakage protocol. The equivalent loss for Enron was under 3%, and no other source lost more. The same rule was applied to every corpus, so the difference reflects how templated each archive is rather than uneven treatment, as Nazario is a template-heavy phishing collection while Enron is ordinary business correspondence. To check that this cleaning does not itself produce the transfer result, the whole evaluation was repeated with near-duplicate removal disabled, and the cross corpus ROC-AUC was 0.528 without deduplication against 0.529 with it, so the collapse reported below is not an artifact of the cleaning step.

Training conditions

Two training conditions were used. In the first condition, legacy, models were trained only on the legacy human dataset, with legitimate email capped at 12,000 and all phishing retained (10,172 emails: 8,499 legitimate / 1,673 phishing). The cap does not bind in this condition, where the legitimate pool is already smaller than 12,000. Its purpose is the mixed condition below, where the combined legitimate email would otherwise swamp the phishing set, and it is applied in both so that the two conditions differ only in which corpora they draw from. All phishing was retained in both. This first condition was designed to recreate the real-world 'train on legacy, deploy on modern' scenario and demonstrate the effect of distribution shift. In the second condition, mixed, models were trained on a combined pool of the legacy human and cross corpus human datasets under the same rule (15,524 emails: 12,000 legitimate / 3,524 phishing). This mixed condition was designed to determine whether retraining the model would reduce the effect of distribution shift. In both cases the Random

Table 1 Composition of the three evaluation datasets after per-source cleaning and before global deduplication.

Datasets	Role	Source(s)	Count (ham / phishing)
Legacy human	Training and in-distribution test	Enron (ham), Nazario (phishing)	18,574 (15,369 / 3,205)
Cross corpus human	Independent human email	CEAS-08 + SpamAssassin (ham), Nigerian-Fraud (phishing)	24,575 (21,291 / 3,284)
LLM	Machine-generated email	Dizzy0x00 LLMGen (cleaned)	3,754 (1,274 / 2,480)

Table 2 Row counts at each stage, from the cleaned datasets to the final evaluation partitions. Percentages are of the deduplicated dataset. The three partitions sum to it exactly.

Stage	Legacy human	Cross corpus human	LLM
Before global deduplication	18,574	24,575	3,754
Removed as duplicates within the dataset	1,620	1,179	0
Removed as duplicates shared with another dataset	0	2	0
After global deduplication	16,954	23,394	3,754
Training partition (60%)	10,172	14,036	2,252
legitimate / phishing	8,499 / 1,673	12,185 / 1,851	764 / 1,488
Calibration partition (15%)	2,543	3,509	563
Test partition (25%)	4,239	5,849	939
legitimate / phishing	3,542 / 697	5,078 / 771	319 / 620

Forest and Logistic Regression used balanced class weights to account for the resulting class imbalance, and Naive Bayes used its class priors. Each dataset was split into 60% training, 15% calibration, and 25% testing. The calibration split set the decision threshold, and the test split was used only for the final evaluation. The testing dataset kept its natural class ratio, with no per-class cap. The main evaluation, including the feature-set ablation, was repeated across 10 random seeds, and its metrics are aggregated over those seeds. The leave-one-source-out test and the corpus-distance measurement each used five seeds, and the individual-feature comparison a single seed. Each is reported with its own count.

Features and anti-leakage protocol

Every email was represented using a TF-IDF encoding of the text over unigrams and bigrams (maximum vocabulary of 5,000 terms, stop words and accents removed, everything lowercased), a count of uniform-resource locators (URLs) and an urgency-lexicon score, which is the proportion of a fixed list of 17 phishing-related terms, such as urgent, verify, and winner, that appear anywhere in the message. The features were kept simple to mimic existing detectors rather than optimized ones, so the findings reflect how distribution shift affected the three classical detectors tested here. As metadata features can sometimes unintentionally capture which dataset an email came from rather than genuine phishing signal, a text-only versus metadata-only versus combined ablation was used to test this. Collection and labeling artifacts of this kind have been shown to make a security dataset, and any classifier trained on it, look more reliable than it is²⁴. Before splitting,

near-duplicate emails that differed only in the URLs, email addresses, or numbers were discarded from each dataset to prevent templated phishing from leaking across the training and test split.

Models and procedure

Three classical models were trained using scikit-learn²⁵: a Multinomial Naive Bayes classifier²⁶ (default settings, with learned empirical class priors), a Random Forest (200 trees, balanced class weights), and a Logistic Regression (balanced class weights). All three models and their feature pipelines were frozen after training and applied unchanged to all three test sets. The decision threshold was set on the calibration partition, which is held out from both training and testing, to a 5% false-positive rate, and then held fixed across all test sets. On that partition the achieved rate was 0.0488, and the resulting threshold was 0.139. The forest size was fixed at 200 trees in advance rather than tuned, because the study measures how a detector transfers rather than how well it can be optimized, and leaving the setting untuned keeps that comparison clean. The 5% false-positive rate was chosen because it reflects a realistic operating point for a deployed filter, where the cost of blocking legitimate mail is what limits how aggressive a detector can be.

Evaluation and statistics

For each dataset and model, three performance indicators were recorded: false-negative rate (FNR), false-positive rate (FPR), and ROC-AUC, because class imbalance affects

threshold-based summaries and threshold-independent ones differently²⁷. The main conclusions were based on the threshold-independent ROC-AUC as it does not depend on a particular class ratio²⁸. This matters because the class balance of the LLM dataset is inverted relative to the other two, holding more phishing than legitimate email (2,480 against 1,274), while the legacy and cross corpus human datasets are mostly legitimate. Training used balanced class weights for the Random Forest and Logistic Regression so that the larger class could not dominate the fit. Multinomial Naive Bayes has no equivalent setting and used its learned empirical class priors. To confirm the inversion does not affect the comparison, the LLM test set was resampled to match the legacy phishing prevalence, which moved its ROC-AUC by 0.0029, from 0.8705 to 0.8734, while F1 at the same threshold moved by 0.3626, from 0.845 to 0.483, which is why ROC-AUC is the headline metric. The LLM dataset is also the smallest of the three at 3,754 emails, and its test partition holds 319 legitimate emails against 3,542 and 5,078 for the other two, so its estimates rest on the fewest cases. Each headline metric is the mean across the 10 seeds, reported with a 95% across-seed band computed from the spread of those results. The seeds change the data split and, for the stochastic models, the model fit as well, so this band describes how stable the pipeline is rather than sampling uncertainty for a wider population. As real-world inboxes' phishing frequency is well below 50%, each dataset was also re-tested at a 10% phishing base rate. All trials were seeded, and every reported metric traces back to a committed output file. Corpus shift was also measured directly to show how far the evaluation corpora were from the training corpus. This measure was based on a classifier that identified corpora instead of phishing. A score of 0.5 means that the two corpora are indistinguishable, and a score of 1.0 means they can be perfectly separated.

Two additional evaluations were run. The first held out each legitimate-email source in turn, training on the remaining sources and calibrating the threshold on a held-out split of that familiar legitimate email rather than on email the model had already been trained on. The second used all three legacy-trained classifiers and compared the same held-out legacy human legitimate email first against credential phishing from the human dataset and then against credential phishing from the LLM dataset.

Ethical considerations

This study did not involve any human subjects or collect any new data, so no informed consent or institutional review was needed. All email datasets used for this research were publicly available and are standard benchmarks widely used in phishing and spam-detection research. The datasets were used for training and testing purposes only, and no attempt was made

to identify the people whose emails appear in them. No new phishing or attack content was generated for this study. All LLM-generated emails were acquired from an existing public dataset. Each dataset was used in accordance with the terms under which it is distributed. As phishing-detection research can be misused, the analysis was strictly defensive, aiming to show where the detectors fail rather than to help individuals with malicious intent.

Results

In-distribution performance is near-perfect (RQ1)

With a held-out test set, 25% of the legacy human dataset, the Random Forest reached ROC-AUC = 0.999 [0.999, 1.000] over 10 seeds. This high benchmark score is similar to the high in-distribution results that are commonly reported in the phishing-detection literature and provides the in-distribution ceiling against which out-of-distribution performance can be compared.

Out-of-distribution performance collapses on one corpus but holds on another (RQ1)

When the frozen legacy-trained Random Forest model processed the two other datasets, ROC-AUC dropped from 0.999 in-distribution to 0.529 [0.513, 0.545] on the cross corpus human dataset while remaining at a high 0.888 [0.878, 0.899] on the LLM dataset. The collapse to near chance on the cross corpus human dataset cannot be blamed on a change in authorship as it was held constant. The degradation cannot cleanly be attributed to a single cause, however, as the legacy and cross corpus human datasets also differ in their phishing examples and their legitimate email sources. So, the collapse reflects the combined effect of all these factors. The LLM dataset, which differs in both corpus and authorship, degraded only modestly, to 0.888. Out-of-distribution degradation is therefore dataset-dependent, so for the three detectors and six corpora tested here a near-perfect in-distribution score did not predict how well a detector worked on an unfamiliar corpus (Figure 1). Whether machine authorship carries an effect once the phishing subtype and the legitimate email are held fixed is examined in the narrowed comparison reported below.

The same test was then done on the Multinomial Naive Bayes and Logistic Regression models. The Naive Bayes and Logistic Regression performed worse than the Random Forest, achieving ROC-AUCs of 0.365 [0.358, 0.371] and 0.477 [0.471, 0.484] on the cross corpus human dataset respectively. Because the ROC-AUC of the models fell below chance (0.500), their ranking was inverted on this dataset. The scores alone do not show which cues caused the inversion.

Testing on multiple models also confirmed that the effect is classifier-independent (Table 3).

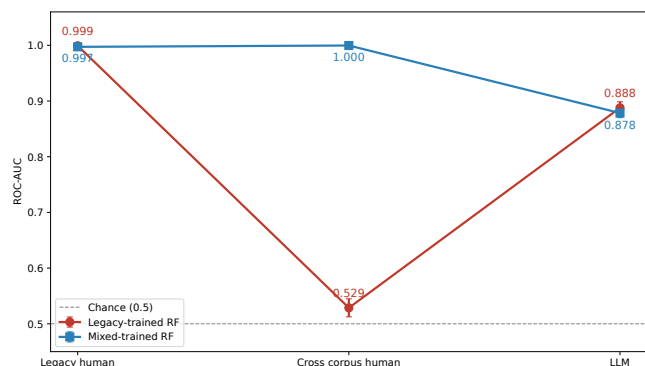


Figure. 1 Frozen-model ROC-AUC across the three evaluation datasets for the legacy-trained and mixed-trained Random Forest. The legacy-trained model degrades steeply under corpus shift, falling from 0.999 in-distribution to 0.529 on the cross corpus human dataset, then recovering to 0.888 on the LLM dataset. Mixed training restores cross corpus performance to 1.000, though that dataset is in-source for the mixed model. Both models sit near 0.88 on the LLM dataset. Error bars show across-seed variability over 10 seeds, and the dashed line marks chance at 0.5.

At deployment thresholds the failure is false alarms, not missed phishing (RQ2)

At the frozen 5% operating point, the degradation was not caused by phishing passing through. Instead, the legacy Random Forest's false-positive rate climbed sharply from 0.049 in-distribution to 0.811 on the cross corpus human dataset and 0.789 on the LLM dataset, while its false-negative rate stayed low at 0.003, 0.037, and 0.001 respectively (Figure 2). This means that a detector trained on one corpus of legitimate email raised false alarms on unfamiliar but legitimate email from other corpora. The same pattern, in which ranking quality holds up while a fixed threshold stops working, has been reported for classifiers that tell human-written phishing from LLM-generated phishing, a separate task, when they are evaluated across different LLM generators²⁹. This finding contradicts the common intuition that degradation occurs due to false negatives. On the cross corpus human dataset this issue cannot be fixed simply by moving the threshold as the ranking is near chance (ROC-AUC 0.529) and lowering the false alarms will also raise missed phishing³⁰, increasing the false-negative rate to 0.605 at the default 0.5 threshold. The behavior of the LLM dataset is different. As the detector ranks the emails reasonably well (ROC-AUC 0.888), the excess false alarms are mainly a result of a miscalibrated threshold. A model can retain its ability to rank while the threshold placed on its

scores stops being trustworthy, so calibration and discrimination are distinct properties³¹. To check that this account of the false alarms was more than an inference, a random sample of 25 false positives was drawn from each dataset and read, with the sampling fixed in code before any message was seen. In-distribution, 11 of the 25 contained phishing vocabulary such as password, account, or security, the clearest case being a genuine Enron notice about deleting user id passwords. Here the vocabulary explanation holds, as legitimate mail was flagged for using the same words the legacy phishing set is built from. On the cross corpus human dataset only 2 of the 25 contained any such vocabulary and 18 were mailing-list or developer threads, and one of those two was a developer thread about reading passwords from standard input. The mechanism therefore differs by dataset: in-distribution the model reacts to shared vocabulary, while on the cross corpus human dataset it flags mail that is simply unlike its training corpus.

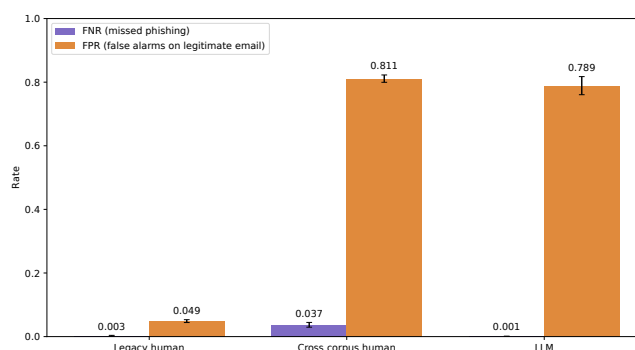


Figure. 2 False-positive rate (false alarms on legitimate email) and false-negative rate (missed phishing) for the legacy-trained Random Forest at the frozen 5% operating point, across the three evaluation datasets. At this operating point the failure is dominated by a rising false-positive rate while the false-negative rate stays low.

Retraining recovers the trained corpus in-source but leaves the LLM gap (RQ3)

When the Random Forest detector was retrained on a mixed pool of the legacy human and cross corpus human datasets, its performance on the cross corpus human dataset recovered almost completely. The cross corpus human ROC-AUC increased from 0.529 to 1.000, and the cross corpus human false-positive rate decreased from 0.811 to 0.039. The recovery was consistent across all ten seeds. However, the cross corpus human data was part of the mixed training set, so it measures in-distribution performance on a corpus the model has now seen, not an ability to generalize to new corpora. The leave-one-source-out test addresses this directly.

The LLM dataset, however, behaved differently. Retraining on a mixed dataset did not change its results meaningfully.

Table 3 Frozen legacy-trained ROC-AUC by model and evaluation dataset (combined features, mean over 10 seeds with across-seed variability). Every model is near-perfect in-distribution but collapses to near or below chance on the cross corpus human dataset, while remaining high on the LLM dataset.

Model	Legacy human (in-distribution)	Cross corpus human	LLM
Random Forest	0.999 [0.999, 1.000]	0.529 [0.513, 0.545]	0.888 [0.878, 0.899]
Multinomial Naive Bayes	0.999 [0.999, 0.999]	0.365 [0.358, 0.371]	0.879 [0.871, 0.888]
Logistic Regression	1.000 [0.999, 1.000]	0.477 [0.471, 0.484]	0.881 [0.874, 0.889]

The ROC-AUC of the LLM dataset changed from 0.888 to 0.878. In a dependent per-seed sensitivity analysis, paired DeLong tests³² were significant at 0.05 in 1 of the 10 seeds, with per-seed p-values from 0.012 to 0.937, so the difference was not consistently detectable. Its false-positive rate stayed high, from 0.789 to 0.810. Importantly, the LLM source was never added to the mixed training pool. This means the comparison does not measure the authorship effect but instead serves as a specificity check. The same intervention that recovers the cross corpus human data did not produce a consistently detectable change on the LLM dataset, indicating that the two failures are separate phenomena. Training on a wider variety of human email datasets can therefore increase the cross corpus human ROC-AUC without affecting the LLM problem (Figure 1).

The false-alarm effect recurs across three independent legitimate-email sources (RQ3)

To test whether exposure to multiple legitimate sources protects a genuinely unseen one, each legitimate source was held out in turn. Each time, a detector was trained on the remaining sources, and its threshold was calibrated to a 5% false-positive rate on a held-out split of that familiar legitimate email. Holding this threshold fixed, the false-positive rate was then measured on the held-out source. Every held-out source was flagged well above the calibrated rate. The false-positive rate on the held-out source's legitimate email was 0.500 for Enron, 0.374 for CEAS-08, and 0.124 for SpamAssassin. Across the three sources held out here, a threshold calibrated on familiar legitimate email produced between twelve and fifty per cent false positives on an unseen legitimate source, two to ten times the five per cent target. The pattern was consistent across the three sources tested, though three sources are not enough to establish it as a general property of cross corpus deployment. It also confirms the earlier caveat that retraining only recovers results for familiar corpora.

A small measured difference in an unmatched authorship comparison (RQ4)

To separate authorship from corpus shift as far as the data allow, the evaluation narrowed the test datasets to a sin-

gle subtype of phishing that was present in both human and LLM datasets, credential phishing, while holding the legitimate email fixed. The legacy-trained model performed almost equally well against both human and LLM-generated phishing, achieving a high ROC-AUC of 0.9999 [0.9998, 1.0000] and 0.9990 [0.9984, 0.9997] respectively, a difference of 0.0009. The comparison used 586 human and 555 LLM credential-phishing messages on average across the ten seeds (ranges 575–596 and 546–566), scored against a fixed set of 3,542 legitimate messages. All three classifiers agree: the difference was 0.0009 for the Random Forest, 0.0004 for Naive Bayes, and 0.0004 for Logistic Regression. With the subtype and legitimate email held fixed, the measured difference was 0.0009 [0.0004, 0.0013] for the Random Forest, 0.0004 [0.0000, 0.0008] for Naive Bayes, and 0.0004 [0.0003, 0.0005] for Logistic Regression. The human and LLM messages were not paired on content, era, or collection process, so this comparison cannot attribute the difference to authorship. Both scores also sit near the ceiling, so it cannot separate a small effect from no effect. The LLMGen dataset card does not document model versions, prompts, or decoding settings.

Robustness checks

Multiple additional evaluations were conducted to determine the robustness of the conclusions. The first evaluation was a text-only versus metadata-only ablation on the legacy-trained Random Forest (Figure 3). When the model was run on the cross corpus human dataset, the text-only ROC-AUC (0.749) was the highest, well above the combined model (0.529), while metadata alone fell below chance (0.483). So, adding the metadata features lowered the cross corpus performance rather than increasing it. On the LLM dataset the pattern flipped. Adding the metadata features increased the ROC-AUC instead of decreasing it. The combined model was the highest at 0.888, closely followed by metadata-only at 0.864 and text-only at 0.852. The two hand-crafted signals behave very differently. Taken alone, the URL count scores 0.828 in-distribution but 0.338 on the cross corpus human data, inverting below chance, while the urgency-lexicon score transfers intact at 0.851 and 0.899. The cross corpus liability is therefore specific to the URL count rather than a general property of metadata features, and only these two signals were tested.

The second evaluation was a source-only probe, run to measure how much of the label a detector could recover from collection source alone. In the legacy human dataset the two sources are split by class, Enron contributing 14,166 legitimate messages and Nazario 2,788 phishing messages, so a classifier that sees only the source name reaches ROC-AUC 1.000. The cross corpus human dataset behaves the same way across its three sources, also reaching 1.000. Within-source performance cannot be computed for either human dataset, because no single human source contains both classes. The LLM dataset is different. Its one source supplies both classes, 1,274 legitimate and 2,480 phishing, so the source-only probe reaches 0.500, which is chance. The frozen detector still scored 0.888 on that dataset. Source identity therefore explains the label completely in the two human datasets and not at all in the LLM dataset, and the transfer result reported above is not reducible to source identity alone. The two human results are the reason the generalization claim in this paper is limited to transfer across these specific nested corpus shifts.

The third test was re-testing at a realistic 10% base phishing rate, because a classifier's decisions have to account for a shift in class priors³³. Adjusting the threshold on a disjoint split behaved very differently across datasets. On the LLM dataset, where the ranking is sound, F1 rose from 0.374 to 0.479. On the cross corpus human dataset, however, F1 barely moved, from 0.153 to 0.215. This was because the ROC-AUC on the 10%-prevalence subsample was 0.521, close to chance, so the ranking carried little signal for any threshold to exploit. The low-prevalence collapse on the cross corpus human dataset is therefore genuinely a ranking failure and not a calibration problem, reinforcing the earlier threshold results. The same evaluation was then repeated at 1, 5, and 20 percent prevalence alongside the 10 percent point. Ranking quality was almost unaffected by the base rate. Across that whole range the cross corpus human ROC-AUC stayed between 0.520 and 0.522, the LLM dataset between 0.877 and 0.890, and the in-distribution score between 0.999 and 1.000. The collapse is therefore a property of the ranking rather than of the class balance at deployment.

Finally, the fourth test measured the distance between the corpora directly, using a classifier trained to tell them apart, which is an established way of detecting dataset shift³⁴. That classifier reached a ROC-AUC of 0.984 for legacy versus cross corpus human and 1.000 for legacy versus LLM. Both evaluation corpora are therefore almost perfectly distinguishable from the training corpus, yet the detector collapses on the cross corpus human dataset (0.529) while holding up on the LLM dataset (0.888). How distinguishable a corpus is from the training data does not by itself predict how well the detector will transfer to it.

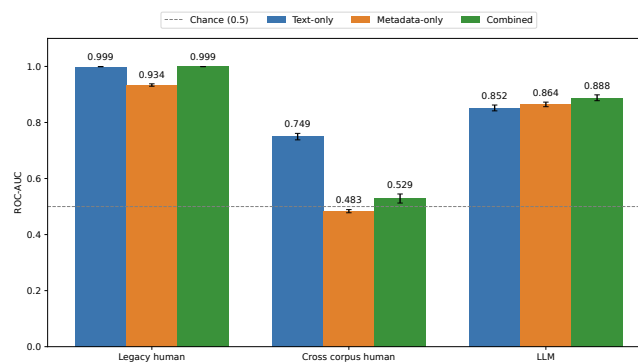


Figure. 3 Legacy-trained Random Forest ROC-AUC by feature set across the three evaluation datasets (mean over 10 seeds with across-seed variability, dashed line marks chance). Metadata features help on the LLM dataset but drag the combined model below the text-only model on the cross corpus human dataset, suggesting that the URL count in particular encodes corpus identity rather than a transferable phishing signal. Only these two hand-crafted signals were tested.

Discussion

Principal findings

There are three main findings from this experiment. First, for the three detectors and six corpora tested here, in-distribution benchmark scores overstated robustness: a detector at ROC-AUC 0.999 in-distribution falls to 0.529, near chance, on the cross corpus human dataset while holding at 0.888 on the LLM dataset. Second, the cross corpus collapse is a ranking failure rather than a mis-calibrated threshold: as the detector's scores no longer separate phishing from legitimate email (ROC-AUC 0.529), changing the cutoff cannot rescue it. Retraining on a mixed pool increased the ROC-AUC from 0.529 to 1.000, but this was only because the corpus was part of the training data. A leave one source out test, which trains on a few legitimate email corpora and measures the false-positive rate on a corpus not seen by the model before, still produced false-positive rates as high as 0.50, so retraining recovers performance on familiar corpora but not on new ones.

Third, no large difference was measured once the subtype and legitimate email were held fixed: on credential phishing with the legitimate email fixed, the detector scored 0.9999 on human and 0.9990 on machine-generated email, a difference of 0.0009. This suggests that any authorship effect is at most small. Overall, a near-perfect in-distribution benchmark score is a poor indicator of how a detector will behave on a different corpus, which is the central benchmark validity concern of this study.

Interpretation

How and why phishing detectors fail is more mundane than what the popular narrative suggests. The claim that LLM authorship is the main factor breaking phishing detectors is not supported by the narrowed comparison reported above. Instead, the detector's actual weakness is that its accuracy does not transfer reliably across corpora: in-distribution test's near perfect scores drastically drop on one human corpus (0.529) yet hold high on the LLM corpus (0.888). A detector trained and benchmarked on a single narrow corpus can therefore be unreliable when it is asked to judge email from a different corpus. Broader, more representative training data helps for the corpora that can be collected, and methods that use unlabeled target data or that optimize for the worst-performing group have been proposed for exactly this kind of sampling bias^{35,36}. However, it cannot be the whole solution because a detector will always encounter unfamiliar data during deployment. For the detectors and corpora tested here, a single-corpus benchmark was an unreliable proxy for performance on an unfamiliar corpus, which suggests cross corpus evaluation is worth reporting alongside the standard benchmark, and methods for estimating how a model will perform under distribution shift have been developed for exactly this purpose³⁷. Corpus distance alone does not forecast this failure either, as both evaluation corpora were almost perfectly distinguishable from the training corpus yet only one of them broke the detector.

In the narrowed comparison, where the phishing subtype and legitimate email were held fixed, machine-generated email carried a difference of 0.0009 in ROC-AUC. Both scores sit near the ceiling, and the messages were not paired on content, so this comparison cannot attribute the difference to authorship. Because the failure mode is false alarms rather than missed phishing, the practical danger of a detector deployed on a different corpus is operational rather than a security problem. A rising false-alarm rate should therefore be treated as an early warning sign of corpus shift, prompting developers to retrain on more diverse data or recalibrate the threshold. Recalibration, however, helps only where the ranking still separates the classes. Where it has collapsed, as on the cross corpus human dataset, performance can only be restored by retraining the model on that specific corpus, and a corpus it has never seen is still flagged heavily. For anyone running a detector in production, the practical sequence follows from this. A false-alarm rate that climbs on live traffic is the signal to act on, not a falling catch rate. Check first whether the ranking still holds by scoring a small labeled sample from the new source: if it does, recalibrating the threshold on that sample will recover most of the loss, and if it does not, no threshold will help and the model has to be retrained on data from that source.

Comparison with prior work

Previous studies have largely focused on whether LLMs can write convincing phishing emails^{9,10} or whether detectors decay due to temporal shift³⁸. Other work has examined whether machine-written phishing can be told apart from legitimate email³⁹, and has compared human-written and machine-written phishing drawn from separate corpora⁴⁰, which are different questions from the one asked here. This study tackles the problem from a different angle, by measuring how a frozen detector transfers across independent corpora and by testing whether machine-written phishing has any evasion advantage over human phishing once the subtype and legitimate email are held fixed. A recent independent study reached a compatible conclusion, finding that cross corpus transfer and prevalence shift degrade detector performance, and recommending pooled multi-corpus training and prevalence-aware thresholding as partial mitigations⁴¹. That study therefore supports the transfer-failure characterization reported here, while the narrowed test of machine authorship and the finding that transfer is dataset-dependent rather than uniformly degrading are contributions specific to this study.

Limitations

There are four key limitations that affect this study. Firstly, the authorship effect reported here is derived from the narrowed comparison reported in the Results, where the phishing subtype and the legitimate email were held fixed. This is the closest the available corpora allow, but because both scores sit at the ceiling (ROC-AUC 0.9999 and 0.9990), the test cannot separate a small difference from no difference. The human and LLM messages were also never paired on content, era, or collection process, so authorship is not the only factor that varies between them. The raw gap between the cross corpus human and LLM datasets is not used to measure the effect of authorship as the comparison is confounded: the LLM dataset's legitimate emails are also machine-generated and from a different corpus.

Secondly, the cross corpus human dataset varies from the legacy dataset along two distinct axes at once. The phishing changes subtype, from the general phishing of the Nazario corpus to the advance-fee fraud of Nigerian-Fraud, while the legitimate email changes source, from Enron to CEAS-08 and SpamAssassin. Both shifts are near-total: a classifier separates the two phishing sets with a ROC-AUC of 0.9991, and the two legitimate sets with 0.9988, where 1.0 denotes perfect separability. The design therefore shows that transfer fails without identifying exactly which factor had how much effect. Thirdly, every machine-generated email used here came from a single family of models, as the LLM dataset was generated entirely by unspecified OpenAI models and its documentation records no model versions, prompts, or decoding settings.

Other families such as Claude, Gemini, or open-source models may write differently, so this result applies to the model family tested here rather than to machine-generated phishing in general. Finally, the study only evaluates simple, classical detectors using just the message's text and two metadata signals, so whether a modern neural or LLM-based detector shows similar cross-dataset fragility remains untested. This feature set was carefully picked to mimic legacy detectors, and the ablation reported in the robustness checks shows the findings hold on text alone. This research could be furthered in the future by testing whether the findings hold up on modern detectors with richer features.

Conclusion

A high in-distribution benchmark score does not provide a good indication of how the model behaves on a different corpus. A classical detector that scored a ROC-AUC of 0.999 in-distribution collapsed to near chance on one human corpus (0.529) while still holding high on the LLM corpus (0.888). So, for the three detectors and six corpora tested here, cross corpus transfer was unreliable and was not predicted by an in-distribution benchmark. At the deployed detection threshold, this collapse occurred in the form of false alarms on legitimate email and as the model's ranking itself had failed, no threshold could reduce the false-positive rate without increasing missed phishing. The model's ROC-AUC improved after retraining, but only if the test corpus was also present in the training pool. So, it is not a general fix. With the subtype and legitimate email held fixed, no large difference was measured, so the popular narrative that LLM-written phishing is what breaks detectors is not supported by these results. Because the two message sets were not paired on content, this comparison cannot isolate authorship. These findings suggest that cross corpus evaluation is a useful complement to the standard in-distribution benchmark, threshold-independent metrics reported with their across-seed variability give a more reliable picture of robustness than a single-threshold score, and corpus shift should be accounted for when attributing a detector's failure to LLM authorship.

Acknowledgments

The author thanks the maintainers of the public email corpora (Enron, Nazario, SpamAssassin, CEAS-08, Nigerian-Fraud, and the Dizzy0x00 LLMGen set) and the scikit-learn and Hugging Face open-source projects, without which this study would not have been possible. This research was conducted independently by the author. This research received no external funding, and the author declares no competing interests.

References

- 1 Anti-Phishing Working Group. Phishing activity trends report. https://docs.apwg.org/reports/apwg_trends_report_q4_2023.pdf, 2024.
- 2 Verizon. Data breach investigations report (DBIR). <https://www.verizon.com/business/resources/reports/2023-data-breach-investigations-report-dbir.pdf>, 2023.
- 3 P. Bountakas, K. Koutroumpouchos, C. Xenakis. A comparison of natural language processing and machine learning methods for phishing email detection. *16th International Conference on Availability, Reliability and Security*. pg. 1–12, 2021, <https://doi.org/10.1145/3465481.3469205>.
- 4 N. Altwaijry, I. Al-Turaiki, R. Alotaibi, F. Alakeel. Advancing phishing email detection: a comparative study of deep learning models. *Sensors*. Vol. **24**, pg. 2077, 2024, <https://doi.org/10.3390/s24072077>.
- 5 D. Arp, E. Quiring, F. Pendlebury, A. Warnecke, F. Pierazzi, C. Wressnegger, L. Cavallaro, K. Rieck. Dos and don'ts of machine learning in computer security. *31st USENIX Security Symposium*. pg. 3971–3988, 2022, <https://www.usenix.org/conference/usenixsecurity22/presentation/arp>.
- 6 F. Pendlebury, F. Pierazzi, R. Jordaney, J. Kinder, L. Cavallaro. TESSER-ACT: eliminating experimental bias in malware classification across space and time. *28th USENIX Security Symposium*. pg. 729–746, 2019, <https://www.usenix.org/conference/usenixsecurity19/presentation/pendlebury>.
- 7 J. Quiñero-Candela, M. Sugiyama, A. Schwaighofer, N. D. Lawrence (eds.). *Dataset shift in machine learning*. MIT Press, 2009.
- 8 S. S. Roy, P. Thota, K. V. Naragam, S. Nilizadeh. From chatbots to phishbots?: phishing scam generation in commercial large language models. *2024 IEEE Symposium on Security and Privacy*. pg. 36–54, 2024, <https://doi.org/10.1109/SP54263.2024.00182>.
- 9 F. Heiding, B. Schneier, A. Vishwanath, J. Bernstein, P. S. Park. Devising and detecting phishing emails using large language models. *IEEE Access*. Vol. **12**, pg. 42131–42146, 2024, <https://doi.org/10.1109/ACCESS.2024.3375882>.
- 10 F. Heiding, S. Lermen, A. Kao, C. Mayrink Verdun, B. Schneier, A. Vishwanath. Evaluating large language models' ability to automate spear phishing. *Expert Systems with Applications*. Vol. **314**, pg. 131546, 2026, <https://doi.org/10.1016/j.eswa.2026.131546>.
- 11 G. Ho, A. Cidon, L. Gavish, M. Schweighauser, V. Paxson, S. Savage, G. M. Voelker, D. Wagner. Detecting and characterizing lateral phishing at scale. *28th USENIX Security Symposium*. pg. 1273–1290, 2019, <https://www.usenix.org/conference/usenixsecurity19/presentation/ho>.
- 12 A. Alhuzali, A. Alloqmani, M. Aljabri, F. Alharbi. In-depth analysis of phishing email detection: evaluating the performance of machine learning and deep learning models across multiple datasets. *Applied Sciences*. Vol. **15**, pg. 3396, 2025, <https://doi.org/10.3390/app15063396>.
- 13 S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*. Vol. **10**, pg. 1096, 2019, <https://doi.org/10.1038/s41467-019-08987-4>.
- 14 A. J. DeGrave, J. D. Janizek, S.-I. Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*. Vol. **3**, pg. 610–619, 2021, <https://doi.org/10.1038/s42256-021-00338-7>.
- 15 M. A. Badgeley, J. R. Zech, L. Oakden-Rayner, B. S. Glicksberg, M. Liu, W. Gale, M. V. McConnell, B. Percha, T. M. Snyder, J. T. Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*. Vol. **2**, pg. 31, 2019, <https://doi.org/10.1038/s41746-019-0105-1>.
- 16 N. A. Alam. Phishing email dataset. <https://www.kaggle.com/d>

- atatasets/naserabdullahalam/phishing-email-dataset, 2024.
- 17 A. Al-Subaiey, M. Al-Thani, N. A. Alam, K. F. Antora, A. Khandakar, S. M. A. Uz Zaman. Novel interpretable and robust web-based AI platform for phishing email detection. *Computers and Electrical Engineering*. Vol. **120**, pg. 109625, 2024, <https://doi.org/10.1016/j.compeleceng.2024.109625>.
 - 18 B. Klimt, Y. Yang. The Enron corpus: a new dataset for email classification research. *Proceedings of the 15th European Conference on Machine Learning (ECML 2004), Lecture Notes in Computer Science*. Vol. **3201**, pg. 217–226, 2004, https://doi.org/10.1007/978-3-540-30115-8_22.
 - 19 J. Nazario. Phishing corpus. <https://monkey.org/~jose/phishing/>, 2005–2007.
 - 20 Apache SpamAssassin Project. SpamAssassin public corpus. <https://spamassassin.apache.org/old/publiccorpus/>, 2002–2003.
 - 21 Conference on Email and Anti-Spam. CEAS 2008 live spam challenge laboratory corpus. <http://www.ceas.cc/> (no longer distributed by this host), 2008.
 - 22 D. R. Radev. CLAIR collection of fraud email (ACL Data and Code Repository, ADCR2008T001). [https://www.aclweb.org/aclwiki/CLAIR_collection_of_fraud_email_\(Repository\)](https://www.aclweb.org/aclwiki/CLAIR_collection_of_fraud_email_(Repository)), 2008.
 - 23 Dizzy0x00. LLMGen-Phishing-Email-Dataset (LLM-generated phishing email dataset). <https://huggingface.co/datasets/Dizzy0x00/LLMGen-Phishing-Email-Dataset>, 2025.
 - 24 G. Engelen, V. Rimmer, W. Joosen. Troubleshooting an intrusion detection dataset: the CICIDS2017 case study. *2021 IEEE Security and Privacy Workshops*. pg. 7–12, 2021, <https://doi.org/10.1109/SPW53761.2021.00009>.
 - 25 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*. Vol. **12**, pg. 2825–2830, 2011, <https://www.jmlr.org/papers/v12/pedregosa11a.html>.
 - 26 I. Androustopoulos, J. Koutsias, K. V. Chandrinou, G. Paliouras, C. D. Spyropoulos. An evaluation of naive Bayesian anti-spam filtering. *Proceedings of the Workshop on Machine Learning in the New Information Age (ECML 2000)*. pg. 9–17, 2000, <https://arxiv.org/abs/cs/0006013>.
 - 27 A. Luque, A. Carrasco, A. Martín, A. de las Heras. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*. Vol. **91**, pg. 216–231, 2019, <https://doi.org/10.1016/j.patcog.2019.02.023>.
 - 28 J. Li. Area under the ROC curve has the most consistent evaluation for binary classification. *PLOS ONE*. Vol. **19**, pg. e0316019, 2024, <https://doi.org/10.1371/journal.pone.0316019>.
 - 29 R. Gutierrez, W. Villegas-Ch, J. Govea. Cross-model evaluation of phishing detectors against LLM-generated emails. *Frontiers in Big Data*. Vol. **9**, pg. 1883452, 2026, <https://doi.org/10.3389/fdata.2026.1883452>.
 - 30 C. Elkan. The foundations of cost-sensitive learning. *Proceedings of the 17th International Joint Conference on Artificial Intelligence*. pg. 973–978, 2001.
 - 31 J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. H. S. Torr, P. K. Dokania. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*. Vol. **33**, pg. 15288–15299, 2020, <https://proceedings.neurips.cc/paper/2020/hash/ae7b30ef1d024a76f21ald40e30c302-Abstract.html>.
 - 32 E. R. DeLong, D. M. DeLong, D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. Vol. **44**, pg. 837–845, 1988, <https://doi.org/10.2307/2531595>.
 - 33 M. Saerens, P. Latinne, C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation*. Vol. **14**, pg. 21–41, 2002, <https://doi.org/10.1162/089976602753284446>.
 - 34 S. Rabanser, S. Günnemann, Z. C. Lipton. Failing loudly: an empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*. Vol. **32**, pg. 1396–1408, 2019, <https://proceedings.neurips.cc/paper/2019/hash/846c260d715e5b854ffad5f70a516c88-Abstract.html>.
 - 35 S. Thirumuruganathan, F. Deniz, I. Khalil, T. Yu, M. Nabeel, M. Ouzzani. Detecting and mitigating sampling bias in cybersecurity with unlabeled data. *33rd USENIX Security Symposium*. pg. 1741–1758, 2024, <https://www.usenix.org/conference/usenixsecurity24/presentation/thirumuruganathan>.
 - 36 S. Sagawa, P. W. Koh, T. B. Hashimoto, P. Liang. Distributionally robust neural networks for group shifts: on the importance of regularization for worst-case generalization. *Eighth International Conference on Learning Representations*. 2020, <https://openreview.net/forum?id=ryxGuJrFvS>.
 - 37 M. Chen, K. Goel, N. S. Sohoni, F. Poms, K. Fatahalian, C. Ré. Mando-line: model evaluation under distribution shift. *Proceedings of the 38th International Conference on Machine Learning, PMLR*. Vol. **139**, pg. 1617–1629, 2021, <https://proceedings.mlr.press/v139/chen21i.html>.
 - 38 S. J. Delany, P. Cunningham, A. Tsymbal, L. Coyle. A case-based technique for tracking concept drift in spam filtering. *Knowledge-Based Systems*. Vol. **18**, pg. 187–195, 2005, <https://doi.org/10.1016/j.knosys.2004.10.002>.
 - 39 C. Opara, P. Modesti, L. Golightly. Evaluating spam filters and stylometric detection of AI-generated phishing emails. *Expert Systems with Applications*. Vol. **276**, pg. 127044, 2025, <https://doi.org/10.1016/j.eswa.2025.127044>.
 - 40 C. S. Eze, L. Shamir. Analysis and prevention of AI-based phishing email attacks. *Electronics*. Vol. **13**, pg. 1839, 2024, <https://doi.org/10.3390/electronics13101839>.
 - 41 I. Bhuiyan, T. Bhuiyan. The fragility of phishing detection models: evidence from cross-corpus transfer, prevalence shift, artifact learning, and evasion risk. *Big Data and Cognitive Computing*. Vol. **10**, pg. 211, 2026, <https://doi.org/10.3390/bdcc10070211>.