Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Explainable Machine Learning Framework with Large Language Model Features for Predicting Monetary Redress in Bank Fraud and Scam Complaints

Tahsien Al-Quraishi
Melbourne Institute of Technology https://orcid.org/0000-0003-3814-0431
O. S. Albahri
o.albahrey@federation.edu.au
Federation University
Raza Nowrozy
Victoria University
Naseer Ali Hussien
Al-Ayen Iraqi University
Urvashi Rahul Saxena
Melbourne Institute of Technology
Islam Mohammad Tahidul
Melbourne Institute of Technology
Jemal Abawajy
Deakin University

Research Article

Keywords: Scam redress, Authorised push payment fraud, Large language models, Explainable artificial intelligence, Cost-sensitive decision-making

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11721933/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

An Explainable Machine Learning Framework with Large Language Model Features for Predicting Monetary Redress in Bank Fraud and Scam Complaints

Tahsien Al-Quraishi[1], O. S. Albahri[2*], Raza Nowrozy[3], Naseer Ali Hussien[4], Urvashi Rahul Saxena[5], Islam Mohammad Tahidul[5], Jemal Abawajy[6]

[1]School of Business, Melbourne Institute of Technology, Melbourne, Victoria, Australia.
[2*]Institute of Innovation, Science and Sustainability, Federation University, Melbourne, Victoria, Australia.
[3]College of Art, Business, Law, Education & IT (CoABLEIT), Victoria University, Melbourne, Victoria, Australia.
[4]Department of Computer Engineering Technology, College of Technical Engineering, Al-Ayen Iraqi University (AUIQ), Nasiriyah, Iraq.
[5]School of IT and Engineering (SITE), Melbourne Institute of Technology, Melbourne, Victoria, Australia.
[6]Faculty of Science, Engineering and Built Environment, Deakin University, Melbourne, Victoria, Australia.

*Corresponding author(s). E-mail(s): o.albahrey@federation.edu.au;
Contributing authors: tquraishi@mit.edu.au; rnowrozy@mit.edu.au; naseerali@alayen.edu.iq; usaxena@mit.edu.au; mtislam@mit.edu.au; jemal.abawajy@deakin.edu.au;

Abstract

Liability for scam losses is shifting towards banks, yet few interpretable tools can anticipate which fraud and scam complaints will end in redress. This study develops SCAM-XD, an explainable framework that predicts whether a complaint to the US Consumer Financial Protection Bureau (CFPB) flagged as possible fraud or a scam closes with monetary relief reported by the company. From the CFPB

narratives archive, the study builds a benchmark of these complaints, about half of which human coders confirm as fraud or scams. A large language model (LLM) codes each narrative for scam type, payment rail, authorisation, manipulation cues and signals of the bank's response, agreeing closely with human coding on scam type. These features are combined with structured fields, company identity and text representations in a stacking ensemble, explained with SHapley Additive exPlanations (SHAP) and linked to a cost-sensitive triage layer. Tested on a later period, SCAM-XD improves significantly on strong baselines across seeds, including a reimplemented topic model approach. LLM features yield a compact, interpretable model that nearly matches a high-dimensional text baseline but do not significantly improve the full model. Company identity is the strongest driver of predicted redress, yet firm identity alone predicts far less well, and payment rail and bank response cues act in directions consistent with liability rules. Under assumed costs, thresholds selected on validation data substantially reduce modelled expected cost relative to standard thresholds, whereas model choice matters much less. Redress depends on both the firm and the facts that determine liability.

Keywords: Scam redress, Authorised push payment fraud, Large language models, Explainable artificial intelligence, Cost-sensitive decision-making

1 Introduction

Scams that persuade victims to move their own money have become a central problem in retail banking. In the United States (US), consumers reported losing US$15.9 billion to fraud in 2025, up from more than US$12 billion in 2024, with imposter scams the most reported category (Federal Trade Commission, 2026). In the United Kingdom (UK), losses to authorised push payment (APP) fraud rose by 19% in 2025 to £576.4 million across 248,070 cases (UK Finance, 2026), and Australians lost A$2.18 billion to scams in the same year (National Anti-Scam Centre, 2026). Unlike unauthorised fraud, in which a third party makes a payment without the account holder's consent, APP fraud occurs when a person is tricked into authorising a payment to a fraudster (UK Finance, 2026). Research on fraudulent communication shows that such attacks exploit psychological tactics including authority, urgency and impersonation, and that social engineering continues to cause substantial losses despite technical defences (Bera et al, 2023). Generative artificial intelligence (AI) is likely to increase the scale of the problem: in a survey experiment with 4,100 US adults, Heiding et al (2027) found an overall compliance rate of 16.5% with voice phishing requests generated by AI, rising to 36.1% for relative-in-distress scenarios, and concluded that AI-automated voice phishing appears economically viable.

Responsibility for these losses is shifting towards banks. In the UK, a mandatory reimbursement requirement for victims of APP scams came into force on 7 October 2024 (Payment Systems Regulator, 2023). In Australia, the Scams Prevention Framework Act 2025 requires regulated entities in designated sectors, which may include banking, to take reasonable steps to prevent, detect and disrupt scams, to report actionable scam intelligence and to provide mechanisms for responding to consumer

reports (Commonwealth of Australia, 2025). In the US, by contrast, the liability protections of Regulation E apply to unauthorised electronic fund transfers, a definition that excludes transfers initiated by the consumer (Consumer Financial Protection Bureau, 2026c). Whether a scam victim recovers money therefore depends on how the case is characterised and how the bank responds. For banks and regulators alike, the ability to anticipate which fraud and scam complaints will end in redress, and to explain why, is becoming an operational and supervisory need.

Most machine learning research on scams addresses *detection*: identifying a fraudulent entity before or while harm occurs. Examples include detecting fraudulent websites from their requests structures to third parties (Gopal et al, 2022), scam tokens in decentralised finance (DeFi) (Gunathilaka et al, 2024) and phishing accounts on Ethereum (Tang et al, 2024; Ye et al, 2026; Liu et al, 2026). These studies model the infrastructure or transactions of the scammer. They say little about what happens after a victim seeks redress, the stage at which bank decisions, consumer protection, and regulatory oversight come into play.

Research on complaint outcomes addresses this later stage, but only partly. Siering (2022) proposed an explainable and fair regulatory technology approach (RegTech) that predicts whether a complaint to the US Consumer Financial Protection Bureau (CFPB) results in monetary compensation, combining linguistic features grounded in information diagnosticity theory with a bag-of-words model and ensemble learning, and examining fairness for older consumers. That work established that complaint text carries information about outcomes, but it was not specific to scams, was trained on a balanced sample, did not use an out-of-time evaluation and predates large language models (LLMs). More recently, Wang et al (2026) combined narrative text, topic representations, engineered text features and structured attributes in a gradient-boosted classifier evaluated on more recent complaints, and reported that company identity carries predictive information about monetary relief. To our knowledge, however, no study has modelled redress specifically for fraud and scam complaints, represented complaint narratives through scam semantics extracted by an LLM, tested the stability of its explanations, or converted predicted risk into a cost-sensitive triage decision.

The data available for such research have also changed. In 2026, the CFPB removed consumer complaint narratives from its public database and made them available through an official narratives archive (Consumer Financial Protection Bureau, 2026b,a). Research based on CFPB complaint narratives now depends on that archive, which makes a documented scam subset drawn from it valuable in its own right.

This study develops and evaluates SCAM-XD, an explainable, LLM-augmented framework for predicting whether a fraud- or scam-related complaint ends in monetary relief. We use 46,815 complaints flagged as potentially fraud- or scam-related, received between 2023 and 2026 and drawn from the CFPB narratives archive, of which 11.62% closed with monetary relief; human coding of a random sample indicates that about half concern fraud or a scam. An LLM codes each narrative against a fixed schema of scam type, payment rail, authorisation, manipulation cues and bank-response signals. These features enter a model comparison that ends in a stacking ensemble, which is explained with SHapley Additive exPlanations (SHAP) and linked to a cost-sensitive triage layer. All models are tuned and selected on a validation period, and test results

play no part in the selection criteria (Section 4.3). We address five research questions (RQs):

- **RQ1:** How accurately can models augmented with LLM features predict monetary relief for fraud and scam complaints?
- **RQ2:** Do scam features extracted by the LLM improve prediction relative to term frequency–inverse document frequency (TF-IDF) and topic-model representations?
- **RQ3:** Which narrative and institutional factors drive predicted redress, and are the explanations stable across model runs and faithful to the model?
- **RQ4:** Under assumed misclassification costs, how does validation-selected thresholding compare with threshold-only and fixed-capacity triage policies?
- **RQ5:** How do performance and explanations change over time and across subgroups?

The selected stacking model predicts monetary relief well above the base rate. It improves on the strongest baseline, a high-dimensional TF-IDF logistic regression (LR), by a margin that is statistically significant for every random seed, and it also outperforms a reimplementation of a recent topic-model approach on the same split. The LLM features are informative but not decisive: on their own they are weaker than TF-IDF, and adding them to the full LightGBM model did not yield a significant improvement. Combined with structured and company fields, however, they produce a compact, interpretable model that performs almost as well as the high-dimensional baseline. The explanations show that company identity accounts for the largest share of attributed importance, but a baseline built on firm identity alone predicts far less well than the full model, and the explanations are stable across seeds and test quarters.

This study makes four contributions:

- Whereas prior work detects scam entities (Gopal et al, 2022; Gunathilaka et al, 2024; Tang et al, 2024; Ye et al, 2026; Liu et al, 2026) or pools all complaint types (Siering, 2022; Wang et al, 2026), we provide, to our knowledge, the first out-of-time evaluation of a model that predicts whether consumers recover money after fraud and scam complaints, together with a documented benchmark from the CFPB narratives archive whose scam filter and LLM coding are validated against human coding.
- Features extracted by the LLM match a high-dimensional text baseline while remaining interpretable, but add no significant accuracy to the full model; their value is interpretability rather than accuracy.
- Company identity is the strongest driver of predicted redress, but firm identity alone falls well short of the full model. Payment-rail and bank-response characteristics act in directions consistent with the liability rules, and the ranking of drivers is stable across random seeds and test quarters, although it shifts when the model is retrained on later data.
- Cost-sensitive thresholds deliver most of the reduction in expected cost, while the choice between competitive models changes it little.

The remainder of the paper is organised as follows. Section 2 reviews the literature and develops the hypotheses. Section 3 describes the data and sample construction, Section 4 presents the research methodology and Section 5 reports the results.

Section 6 discusses the theoretical, managerial and policy implications, together with limitations, and Section 7 concludes.

2 Literature review and hypothesis development

Research relevant to scam redress spans five streams that have developed largely in isolation: machine learning for scam and fraud detection, the prediction of complaint outcomes, LLMs as information extractors, explainable AI (XAI) and cost-sensitive decision-making. This section reviews each stream, identifies where it falls short for the problem of predicting and explaining redress for scam victims, and closes with a structured comparison (Table 1) and the study's hypotheses.

Most machine learning research on scams focuses on detecting fraudulent entities. Gopal et al (2022) classify fraudulent websites from the structure of their requests to third parties, Gunathilaka et al (2024) detect scam tokens in DeFi by combining transaction event logs with social media sentiment, and Ethereum phishing accounts have been detected with attention-based recurrent networks (Tang et al, 2024), hypergraph-based multiview contrastive learning (Liu et al, 2026) and fine-tuned LLMs (Ye et al, 2026). A complementary stream studies scam content: Bera et al (2023) develop a framework of the psychological and design tactics used in fraudulent email, and Heiding et al (2027) show that some AI voice models achieve persuasiveness ratings comparable to those of human callers and that AI-generated voice-phishing requests elicit compliance from a substantial share of recipients. These studies focus on the scammer. The stage that follows victimisation, when a consumer seeks redress from a bank and liability rules and bank decisions determine the loss, has received far less attention.

The closest prior work predicts the outcomes of consumer financial complaints. Siering (2022) built a RegTech system that classifies complaints to the CFPB according to whether they result in monetary compensation, using linguistic features for explanation, bag-of-words representations and ensemble classifiers for accuracy, and consumer characteristics to examine fairness; a lift analysis showed the value of ranking complaints for review. Earlier analyses of the CFPB database documented marked differences between firms in how complaints are handled (Ayres et al, 2014), and topic models have been used to characterise complaint content at scale (Bastani et al, 2019). Wang et al (2026) combined narrative text, topic representations, engineered text features and structured attributes in a gradient-boosted classifier evaluated on more recent complaints, and found that company identity predicts monetary relief. This literature, however, pools all complaint types, so the dynamics of scams, in which the consumer authorised the payment, are not isolated. Its text representations are lexical or topic-based rather than variables that correspond to liability-relevant concepts, such as the payment rail or authorisation. And it stops at classification, without translating predictions into decisions under explicit costs.

LLMs offer a way to obtain such semantic variables. Rather than using an LLM as a predictor (Ye et al, 2026), this study uses it as a coder that converts narratives into a fixed set of inspectable variables. Evidence on this use is encouraging but mixed: Gilardi et al (2023) found that ChatGPT outperformed crowd workers on several

annotation tasks at a fraction of the cost, whereas Ziems et al (2024) found that LLMs reach only fair agreement with human annotators and recommend using them to augment rather than replace human coding. Other studies show that LLM annotations can vary with the prompt and between runs, and that their accuracy must be validated against human coding for each task (Reiss, 2023; Pangakis et al, 2023). Variables derived from LLM coding should therefore be judged by what they add to established text representations, not assumed to improve prediction.

Explainability is a design requirement for decision support in regulated settings (Siering, 2022). SHAP provides additive feature attributions grounded in cooperative game theory (Lundberg and Lee, 2017), and TreeSHAP computes them exactly for tree ensembles at both local and global levels (Lundberg et al, 2020). Explanations cannot, however, be taken at face value: interpretability methods are often not robust to small input changes (Alvarez-Melis and Jaakkola, 2018), and many feature importance methods perform no better than random assignment (Hooker et al, 2019). Explanations therefore need to be tested for stability and faithfulness.

Finally, triage operates under asymmetric costs, because missing a complaint that merits redress is usually costlier than reviewing one that does not. Cost-sensitive learning shows that optimal decisions follow from applying the cost matrix to predicted probabilities rather than from a default threshold (Elkan, 2001), and decision curve analysis summarises the net benefit of acting on a model across thresholds (Vickers and Elkin, 2006). To our knowledge, no study has evaluated the triage of fraud and scam complaints under explicit misclassification costs.

Table 1 summarises the gap. Detection studies address scams but not redress; complaint-outcome studies address redress but not scams, do not use features derived from LLM coding and stop short of decisions based on explicit costs. No prior study combines all of these elements, and SCAM-XD is designed to fill this gap.

Building on the argument that semantic variables extracted from narratives capture concepts that lexical representations miss, we hypothesise:

H1. Features derived from LLM coding improve the area under the precision–recall curve (PR-AUC) when added to a model that already contains structured, company and conventional text features, on the same temporal split.

Liability rules give these characteristics a direct role in redress. In the US, Regulation E grants error resolution rights for unauthorised electronic transfers but not for payments the consumer authorised (Consumer Financial Protection Bureau, 2026c), and card payments carry dispute rights of their own (Consumer Financial Protection Bureau, 2026d). Complaints whose facts place them within these regimes, and in which the bank's own conduct is in question, should therefore be more likely to end in redress. We thus expect scam and bank-response characteristics to drive predictions in addition to the identity of the firm:

H2. Scam typology and bank-response cues rank among the top SHAP contributors, beyond company identity alone.

Finally, because review capacity is limited and missed redress is costly, we hypothesise:

Table 1 Comparison of the reviewed studies with the present study

Study	Data	Scam focus	Redress outcome	Compl. narr.	LLM feat.	Expl. tested	Out-of-time	Cost layer
Gopal et al (2022)	Websites	✓	×	×	×	×	–	×
Gunathilaka et al (2024)	DeFi tokens	✓	×	×	×	Partial	–	×
Tang et al (2024)	Ethereum accounts	✓	×	×	×	×	–	×
Liu et al (2026)	Ethereum accounts	✓	×	×	×	×	–	×
Ye et al (2026)	Ethereum accounts	✓	×	×	×	×	–	×
Bera et al (2023)	Phishing emails	✓	×	×	×	×	–	×
Heiding et al (2027)	Voice-phishing calls	✓	×	×	×	×	–	×
Siering (2022)	CFPB complaints	×	✓	✓	×	Partial	×	Partial
Wang et al (2026)	CFPB complaints	×	✓	✓	×	×	✓	×
This study	**CFPB fraud and scam complaints**	✓	✓	✓	✓	✓	✓	✓

H3. Under specified costs of false positives and false negatives, thresholds selected on the validation set lower the modelled expected cost of triage relative to threshold-only classification.

3 Data and sample construction

This section describes the data source, the construction of the analytical sample, the definition of the outcome variable, the measures taken to prevent information leakage, and the temporal partitioning used for model development and evaluation.

3.1 Data source: the CFPB narratives archive

The CFPB Consumer Complaint Database publishes complaints about consumer financial products and services after the company concerned has responded. In 2026, the CFPB removed consumer complaint narratives from the public database and made them available through an official archive of periodic exports (Consumer Financial Protection Bureau, 2026b,a). This study uses archive exports 4 to 21, which cover complaints received between 1 November 2022 and 31 August 2026 and together contain 14,483,005 rows. After the removal of eight malformed spillover rows, the data comprise 14,482,997 unique complaint identifiers, with no duplicate identifiers and no unparseable dates. All eighteen files were downloaded on 14 October 2026, and their sizes and SHA-256 hashes were recorded to ensure reproducibility. These records are available from the corresponding author.

3.2 Identification and validation of fraud- and scam-related complaints

Table 2 summarises the sample construction. We retain complaints received from 1 January 2023 that concern the three products through which scam losses typically arise: checking or savings accounts; credit cards; and money transfers, virtual currency and money services. Complaints filed under an earlier credit card product label were recoded to the current label (32,541 complaints). We then retain complaints with a nonempty narrative and apply a scam filter that flags a complaint if either of two rules matches. The issue rule matches the terms "fraud" or "scam" in the issue or subissue field. The narrative rule matches at least one of ten case-insensitive regular expressions in the narrative, covering the terms *scam*, *impersonat*, *Zelle*, *wire* (and its inflections *wired*, *wires* and *wiring*), *told me to*, *investment*, *romance*, *crypto*, *pretend* and *bank's fraud department*. Patterns are anchored at a word boundary, so prefix patterns capture inflected forms (for example, "scammer" and "scammed"), while "Zelle" and "romance" match whole words only and the wire pattern excludes unrelated words such as "wireless". No stemming or lemmatisation is applied, and the full patterns are available from the corresponding author.

Of the 64,271 complaints flagged, 19,213 satisfy the issue rule, 54,906 satisfy the narrative rule, 9,848 satisfy both, and 45,058 are admitted by the narrative rule alone. Because the narrative rule determines most of the sample, we assessed the filter's precision by hand. One author (R.N.), shown only the narrative, coded 400 flagged complaints drawn at random from the 64,153 with a recorded outcome, stratified by admission route (119 issue rule, 281 narrative keywords only), as a deception scam, an unauthorised transaction without deception, neither, or unclear; a second author (U.R.S.) independently coded a random 100 of these (agreement 93.0%, Cohen's $\kappa = 0.902$, 95% CI 0.818–0.971). Of the 300 sampled complaints in the modelling set, 47.3% (95% CI 41.5–53.2; unclear cases excluded) describe fraud or a scam and 29.2% (24.2–34.9) a deception scam (Table 4). Precision is much higher for the issue rule (67.6%) than for narrative keywords alone (34.3%), and most of the remaining complaints concern fees, account closures, billing errors or holds. Grouped into deception, unauthorised and not a scam, the LLM coding agrees with the human coder on 84.8% of these complaints ($\kappa = 0.764$, 0.694–0.827), although it labels fewer complaints as not a scam: 38.83% of the modelling set as deception scams, 26.07% as unauthorised transactions without a scam and 35.10% as not describing a scam (Table 3). The modelling set is therefore best described as complaints flagged as potentially fraud- or scam-related, of which roughly half concern fraud or a scam.

3.3 Sample scope and exclusions

Two groups of complaints whose outcome is effectively fixed were excluded from the modelling set. The first comprises all complaints against Early Warning Services, LLC (16,839 complaints), identified by an exact match on the company field. Early Warning Services operates the Zelle payment network rather than holding consumers' accounts, and none of its complaints in the sample closed with monetary relief, so their outcome

Table 2 Construction of the analytical sample from the CFPB narratives archive

Step	Complaints
Rows in archive exports 4–21	14,483,005
Unique complaint IDs (after removing 8 malformed spillover rows)	14,482,997
Received from 1 January 2023	14,320,813
Three selected products	673,524
Nonempty consumer narrative	353,599
Scam filter (issue rule or narrative keywords)	64,271
of which matched by issue rule	19,213
of which matched by narrative keywords only	45,058
Excluding *Untimely response* (117) and *In progress* (1)	64,153
Modelling set after scope exclusions	46,815
of which closed with monetary relief	5,440

Table 3 LLM-coded scam type in the modelling set (n = 46,815)

Scam type	Complaints	Share (%)
Not a scam	16,434	35.10
Unauthorised transaction without a scam	12,204	26.07
Purchase or marketplace	5,507	11.76
Other scam	3,854	8.23
Bank impersonation	3,392	7.25
Investment or crypto	1,701	3.63
Technical-support impersonation	1,374	2.93
Employment	903	1.93
Invoice or payee redirection	702	1.50
Government impersonation	469	1.00
Romance	275	0.59

Coded by the LLM (Section 4.2). Grouped into deception, unauthorised and not a scam, the codes match the consensus of two independent human coders on 91.6% of the 191 sampled complaints on which the coders agree ($\kappa = 0.872$; Table 4). Deception scams (all rows except the first two) account for 38.83%.

is determined by company identity alone. The second comprises a form-letter campaign that began on 15 January 2025, shortly after the CFPB filed a lawsuit against Early Warning Services and three large banks over fraud on the Zelle network (Consumer Financial Protection Bureau, 2024). Campaign complaints were identified by a case-insensitive regular expression matching the recurring phrase "formally express my concern(s) regarding [. . .] handling of disputes". The expression matched 12,098 complaints dated between 15 January 2025 and 30 June 2026, of which 95.9% named Early Warning Services and 499 named another company; only four closed with monetary relief. Without these exclusions, campaign letters with a nearly fixed outcome would account for 64% of complaints in the validation period.

Table 4 Human validation of the scam filter and the LLM coding

A. Filter precision, modelling set (R.N., $n = 300$), % [95% CI]	All	Issue rule	Keywords only
Fraud or scam[a]	47.3 [41.5, 53.2]	67.6 [58.3, 75.7]	34.3 [27.6, 41.8]
Deception scam only[a]	29.2 [24.2, 34.9]	48.1 [39.0, 57.5]	17.2 [12.2, 23.6]
Neither (share of all)	48.7 [43.1, 54.3]	29.9 [22.4, 38.7]	60.7 [53.4, 67.4]
Unclear (share of all)	7.7 [5.2, 11.2]	7.7 [4.1, 14.0]	7.7 [4.6, 12.4]

B. Agreement on the filter categories	n	Agreement (%)	κ [95% CI]
R.N. vs U.R.S. (4 categories)	100	93.0	0.902 [0.818, 0.971]
LLM vs R.N. (3 groups)[b]	277	84.8	0.764 [0.694, 0.827]

C. Agreement on the LLM-coded fields ($n = 200$), κ (agreement %) Field	LLM vs U.R.S.	LLM vs R.N.	U.R.S. vs R.N.	LLM vs consensus[c]
Scam type (11)	0.798 (84.0)	0.804 (84.5)	0.889 (91.5)	0.852 (88.5; 183)
Scam group (3)[b]	0.847 (90.0)	0.832 (89.0)	0.931 (95.5)	0.872 (91.6; 191)
Payment rail	0.612 (68.5)	0.636 (71.5)	0.667 (73.5)	0.746 (79.6; 147)
Authorised	0.672 (78.5)	0.547 (69.5)	0.656 (77.5)	0.736 (82.6; 155)
Escalated[d]	0.423 (71.0)	0.490 (74.5)	0.409 (69.5)	0.650 (82.7; 139)

Percentages with Wilson 95% CIs; κ CIs from 1,000 bootstrap resamples. [a]Unclear cases excluded (277, 108 and 169 cases). [b]Deception, unauthorised, not a scam. [c]Cases on which the two coders agree; the number of cases follows the agreement percentage. [d]The LLM was asked whether the consumer escalated the complaint within the institution; the coders recorded escalation by the bank.

Because the two groups overlap, the exclusions remove 17,338 complaints (16,839 plus 499), yielding a modelling set of 46,815 complaints from the 64,153 eligible complaints. As the exclusions reflect researcher judgement, Section 5.5 reestimates the main comparison on the full set of 64,153 complaints.

3.4 Outcome definition and leakage control

The outcome is binary. It takes the value 1 if the company's response to the consumer is *Closed with monetary relief*, and 0 if it is *Closed with explanation* or *Closed with nonmonetary relief*. Nonmonetary relief is grouped with closure with explanation because the target is the category of monetary relief reported by the company; nonmonetary remedies, such as correcting an account record, do not restore a loss. The category does not establish whether a loss was reimbursed in full, how much was recovered or whether the complaint was independently adjudicated. Complaints with an *Untimely response* (117) or still *In progress* (1) were removed because their outcome is unknown. To assess whether archived responses were final, we matched the complaints to a public export dated 13 October 2026: of the 16,845 matched complaints, 16,844 retained the same response. This check covers money-transfer complaints only, because the public export was restricted to that product. An escalation outcome was

also considered but could not be constructed: the *Consumer disputed?* field was discontinued in 2017 and is absent from the archive, the database publishes no consumer identifier through which repeat complaints could be linked, and the remaining candidate fields describe either the consumer (tags) or company behaviour after submission. The study therefore models monetary relief only.

SCAM-XD is designed to support triage at the point when a complaint is received, so it uses only information available at submission: structured fields recorded at submission and the consumer's narrative, which is written at submission. Narrative cues about the bank's earlier response, such as a partial refund already given, describe events before the complaint and are therefore available at that point, but they may partly anticipate the recorded outcome; Section 6.4 returns to this. Table 5 lists each raw field and its role in the analysis. Fields that describe company behaviour after submission (*Timely response?* and *Company public response*) and the CFPB routing date are excluded to prevent information leakage. ZIP codes are partially masked and have high cardinality, so state is used instead. The *Submitted via* field is retained in the table for completeness but takes a single value in the modelling set and therefore carries no predictive information.

Table 5 CFPB data fields and their role in the analysis

Field	Role	Rationale
Date received	Feature	Known at submission; split key and calendar features.
Product	Feature	Selected by the consumer at submission.
Subproduct	Feature	Selected by the consumer at submission.
Issue	Feature	Selected by the consumer at submission.
Subissue	Feature	Selected by the consumer at submission.
Consumer complaint narrative	Feature	Written by the consumer at submission (TF-IDF, text features, LLM extraction).
Company	Feature	Named by the consumer at submission (company identity; ablated).
State	Feature	Provided by the consumer at submission.
ZIP code	Excluded	Partially masked by the CFPB and of high cardinality; state used instead.
Tags	Feature	Older American or servicemember status, provided by the consumer at submission.
Submitted via	Feature	Channel of submission (single value in the modelling set).
Company response to consumer	Target source	Used only to construct the outcome; recorded after resolution.
Timely response?	Excluded	Leakage: company behaviour after submission.
Company public response	Excluded	Leakage: written by the company after submission.
Date sent to company	Excluded	Leakage: CFPB routing date, recorded after submission.
Complaint ID	Identifier	Identifier and cache key only.

3.5 Temporal partitioning and descriptive statistics

Although export 21 contains complaints received up to 31 August 2026, no narratives are available for the study products after 29 July 2026: all 18,641 complaints concerning the study products received in August 2026, and those received on 30 and 31 July 2026, lack a narrative and are removed by the narrative requirement. This is consistent with the CFPB's cessation of narrative publication in 2026 (Consumer Financial Protection Bureau, 2026b). The modelling set therefore covers complaints received between 1 January 2023 and 29 July 2026.

Complaints are partitioned by date received into a training period (2023–2024), a validation period (January to June 2025) and a test period (July 2025 to July 2026), as reported in Table 6. All hyperparameter tuning, feature choices, and model selection use the validation period only; test results were not used as a selection criterion, although they were visible between selection rounds (Section 4.3).

Table 6 Temporal partitioning, class balance and product composition. %, monetary-relief rate; product cells report complaints (relief rate, %)

Split	Period	n	Relief	%	Check./sav.	Credit card	Money tr.
Training	2023–2024	22,220	2,776	12.49	7,831 (12.87)	5,295 (18.72)	9,094 (8.54)
Validation	Jan–Jun 2025	8,387	888	10.59	2,502 (12.51)	1,220 (21.39)	4,665 (6.73)
Test	Jul 2025–Jul 2026	16,208	1,776	10.96	5,429 (11.60)	2,616 (18.92)	8,163 (7.98)

The modelling set contains 21,922 money-transfer complaints, 15,762 checking or savings complaints and 9,131 credit card complaints. The monetary relief rate was 11.65% in 2023, 13.36% in 2024, 10.44% in 2025 and 11.64% in 2026, and it varies by product: in the test period it is 18.92% for credit cards, 11.60% for checking or savings accounts and 7.98% for money transfers. Narratives have a median length of 212 words, with a mean of 295.25 and a standard deviation (SD) of 326.14, and the CFPB's redaction token "XXXX" appears a median of ten times per narrative. Complaints are concentrated among a small number of firms whose relief rates differ markedly (Table 7): 26.70% of complaints against Bank of America closed with monetary relief, compared with 0.10% for Block, Inc. and none of the 845 complaints against Robinhood.

4 Methodology

This section formalises the prediction problem and presents the SCAM-XD framework. It then describes the feature representation and LLM-based narrative coding, the predictive models and ensemble, the evaluation and explanation procedures, and the cost-sensitive triage policy and robustness analyses. Full implementation details are available from the corresponding author upon reasonable request.

Table 7 Companies with the most complaints in the analytical sample (top 15)

Company (as named by the CFPB)	Complaints	Relief rate (%)
JPMORGAN CHASE & CO.	5,140	6.58
Block, Inc.	4,830	0.10
WELLS FARGO & COMPANY	4,456	9.13
BANK OF AMERICA, N.A.	4,228	26.70
Paypal Holdings, Inc	3,151	19.01
CITIBANK, N.A.	2,402	23.86
Chime Financial Inc	2,142	6.49
CAPITAL ONE FINANCIAL CORPORATION	2,010	10.50
Coinbase, Inc.	1,204	18.19
NAVY FEDERAL CREDIT UNION	899	14.46
ROBINHOOD MARKETS INC.	845	0.00
SYNCHRONY FINANCIAL	801	21.97
U.S. BANCORP	770	8.83
TRUIST FINANCIAL CORPORATION	698	5.73
AMERICAN EXPRESS COMPANY	673	17.98

4.1 Problem formulation and framework overview

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i, d_i)\}_{i=1}^{N}$ denote the set of complaints, where $\mathbf{x}_i$ contains the information available when complaint i is submitted (structured fields, the named company and the narrative), $y_i \in \{0, 1\}$ indicates whether the complaint closed with monetary relief, and d_i is the date received. The task is to learn a scoring function $f : \mathbf{x} \mapsto \hat{p} \in [0, 1]$ that estimates $\Pr(y = 1 \mid \mathbf{x})$ for complaints not yet resolved. The data are partitioned by date into training, validation and test sets,

$$\mathcal{D}_{\text{tr}} = \{i : d_i < \tau_1\}, \quad \mathcal{D}_{\text{va}} = \{i : \tau_1 \le d_i < \tau_2\}, \quad \mathcal{D}_{\text{te}} = \{i : d_i \ge \tau_2\}, \tag{1}$$

with $\tau_1 = 1$ January 2025 and $\tau_2 = 1$ July 2025 (Section 3). A temporal rather than random partition is used because it mirrors deployment, in which a model trained on past complaints scores new ones, and because it prevents information from later periods from leaking into training. Every feature transformer is fitted only on the rows on which the corresponding model is trained. Figure 1 summarises the framework. An LLM codes each narrative into structured variables; seven feature families are constructed; benchmark models and candidate classifiers are trained and combined in a stacking ensemble; and the selected model is evaluated, explained, embedded in a cost-sensitive decision layer and stress-tested over time and across subgroups. Each stage corresponds to one of the RQs.

4.2 Feature representation, LLM-based narrative coding and validation

Each complaint is represented by up to seven feature families, so that

$$\mathbf{z}_i = \big[\, \mathbf{s}_i \,;\, \mathbf{c}_i \,;\, \mathbf{w}_i \,;\, \mathbf{t}_i \,;\, \mathbf{l}_i \,;\, x_i \,;\, \mathbf{g}_i \,\big], \tag{2}$$

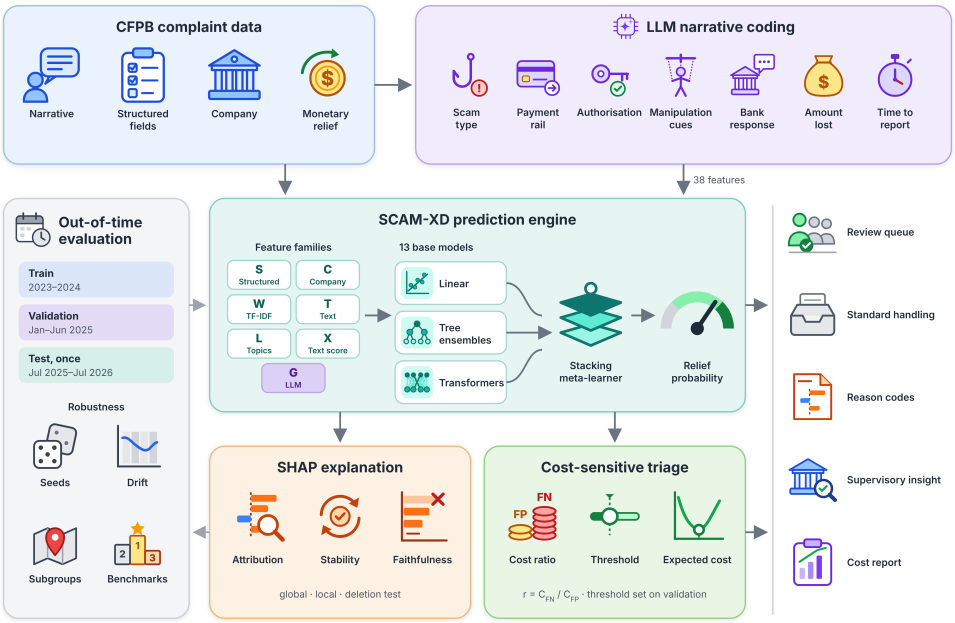


Fig. 1 Overview of the SCAM-XD framework

where different models use different subsets of the families. The families represent the main ways in which complaint text has been modelled in prior work, so that features derived from LLM coding can be tested against them on equal terms: lexical representations, the standard baseline for complaint-outcome prediction (Siering, 2022), and topic models with engineered text features, the most recent approach (Wang et al, 2026). After CFPB redaction markers are replaced by placeholder tokens, the families are constructed as follows. The structured fields $\mathbf{s}_i$ encode product, subproduct, issue, subissue, state, consumer tags, submission channel and calendar features. Company identity $\mathbf{c}_i$ is a one-hot encoding of the named company, with companies that have fewer than 30 training complaints grouped into a single category; it is kept separate from $\mathbf{s}_i$ so that its contribution can be isolated in ablations and explanations. The lexical representation $\mathbf{w}_i$ is a word unigram and bigram TF-IDF vector. The vector $\mathbf{t}_i$ contains the 14 deterministic narrative features of Wang et al (2026), covering length, lexical diversity, readability (Flesch, 1948), monetary amounts, dates, capitalisation and punctuation. The topic features $\mathbf{l}_i$ come from latent Dirichlet allocation (LDA) (Blei et al, 2003) models fitted separately by issue and by company, following Wang et al (2026). The stacked text score x_i is an out-of-fold prediction from an LR model fitted to TF-IDF features, so that no complaint's own label influences its score. Finally, $\mathbf{g}_i$ contains 38 features derived from LLM coding of the narrative.

The LLM is used as a coder rather than as an end-to-end classifier for two reasons. Coding yields a small set of variables that correspond to concepts relevant to liability and redress, such as the payment rail and whether the consumer authorised the payment, which regulators can inspect and to which tree-based explanations can assign

meaningful contributions. Separating extraction from prediction also allows the value of these variables to be tested against conventional text representations in controlled ablations. Formally, the LLM acts as a fixed coding function $\mathcal{M}$ that maps a narrative n_i to a structured record $\mathbf{r}_i = \mathcal{M}(n_i; \pi, \mathcal{S})$ under a prompt π and a JavaScript Object Notation (JSON) schema $\mathcal{S}$, and a deterministic mapping ϕ converts the record into features, $\mathbf{g}_i = \phi(\mathbf{r}_i) \in \mathbb{R}^{38}$. Each of the 46,815 narratives was coded once by Claude Haiku 5.5 against the schema in Table 8, which defines every field and category, requires all fields and permits no others; the manipulation cues draw on typologies of persuasion tactics in fraudulent communication (Bera et al, 2023). No sampling temperature was set, so repeated calls may not return identical outputs; each response was therefore cached under its complaint identifier together with a hash of π and $\mathcal{S}$, so the coding is reproducible from the cache. Coding returned no errors and cost US$5.48 in total, which makes it practical to repeat. The mapping ϕ produces indicators for scam type, payment rail and authorisation; ordinal values with "unknown" flags for amount lost and time to report; the manipulation cues and bank-response signals; and a count of manipulation cues. Because LLM annotations can diverge from human coding (Ziems et al, 2024), two authors (U.R.S. and R.N.), shown only the narrative and blind to the LLM codes, independently coded scam type, payment rail, authorisation and escalation for 200 further complaints drawn at random from the modelling set. The two coders agree closely on scam type (91.5%, $\kappa = 0.889$; 95.5% and $\kappa = 0.931$ when grouped into deception, unauthorised and not a scam) and less closely on payment rail ($\kappa = 0.667$), authorisation ($\kappa = 0.656$) and escalation ($\kappa = 0.409$). On the cases where they agree, the LLM matches their consensus for 88.5% of scam types ($\kappa = 0.852$, 95% CI 0.786–0.913; 91.6% and $\kappa = 0.872$ for the three groups), 79.6% of payment rails ($\kappa = 0.746$), 82.6% of authorisation codes ($\kappa = 0.736$) and 82.7% of escalation codes ($\kappa = 0.650$; Table 4). Agreement between the LLM and each coder separately is similar for scam type ($\kappa = 0.798$ and 0.804) and lower for the other fields. Escalation agrees least well largely because of a difference in definition: the prompt asked whether the consumer escalated the complaint within the institution, whereas the coders recorded escalation by the bank, and almost all disagreements between the LLM and U.R.S. (57 of 58) are cases the LLM codes as escalated and the coder does not. The most frequent scam-type disagreement is a narrative coded as a deception scam by the LLM but as not a scam by a coder (13 cases for U.R.S.). The S+C+G representation contains 286 features, compared with 50,248 for the lexical benchmark B1.

4.3 Predictive models and ensemble design

Five benchmark models span the main alternatives to LLM-derived features. **B1** is an LR on structured fields, company identity and TF-IDF (50,248 features). **B2** is XGBoost on TF-IDF. **B3** reimplements the approach of Wang et al (2026) on our temporal split. It is included because it is the most recent model of monetary relief on CFPB complaints and combines topic and engineered text features; because the original study is a preprint that uses different data and a different base rate, we compare with the method as reimplemented rather than with the figures reported in that study. **B4a** is an LR on sentence embeddings from `all-MiniLM-L6-v2` (Wang et al, 2020;

Table 8 Coding schema for LLM-based narrative extraction

Field	Values
Scam type	Bank impersonation; government impersonation; technical-support impersonation; romance; investment or crypto; purchase or marketplace; employment; invoice or payee redirection; other scam; unauthorised transaction without a scam; not a scam
Payment rail	Zelle or peer-to-peer; wire; card; crypto; cash or gift card; ACH or cheque; other; unknown
Authorised by customer	Yes; no; unclear
Manipulation cues	Urgency; authority impersonation; secrecy; emotional pressure; coached to lie to the bank (each true or false)
Bank-response signals	Reported promptly; bank refused claim; bank cited authorisation; account closed; partial refund given; consumer escalated within the institution (each true or false)
Amount lost	Less than $500; $500–2,000; $2,000–10,000; $10,000–50,000; more than $50,000; unknown
Time to report	Same day; within a week; within a month; over a month; unknown

Reimers and Gurevych, 2019), and **B4b** is a fine-tuned `distilroberta-base` classifier (Sanh et al, 2019; Liu et al, 2019). Together they allow H1 to be tested against lexical, topic-based and neural text representations.

Seven learners with different inductive biases were trained on combinations of feature families, so that conclusions do not depend on a single model family: LR, a calibrated linear support vector machine (SVM), complement naive Bayes, random forest, extra trees, XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al, 2017). Gradient-boosted trees were included because tree-based models remain highly competitive on medium-sized tabular data (Grinsztajn et al, 2022). For each model m with hyperparameter grid Θ_m, the setting is chosen by

$$\theta_m^* = \arg\max_{\theta \in \Theta_m} \text{PR-AUC}\big(f_{m,\theta};\, \mathcal{D}_{\text{va}}\big), \tag{3}$$

where $f_{m,\theta}$ is fitted on $\mathcal{D}_{\text{tr}}$, with early stopping on validation average precision for gradient-boosted models. The chosen setting is then refitted on $\mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{va}}$ and evaluated on the test set.

The final SCAM-XD model is a stacking ensemble (Wolpert, 1992) of $M = 13$ base models: the seven learners above, trained on structured, company, TF-IDF and LLM features; four further base models; and the two neural benchmarks. The four further models are (i) LightGBM on S+C+G with a wider hyperparameter search, which also serves as the explanation model (Section 4.4); (ii) LightGBM tuned over a 144-setting grid on S+C+G and the stacked text score (287 features); (iii) an LR ($C = 0.3$, no class weighting) on structured fields, company identity, the 38 LLM features scaled to $[0, 1]$, word 1–2-gram TF-IDF (50,000 terms) and character 3–5-gram TF-IDF (61,057 terms in the training period); and (iv) LightGBM trained on the features with nonzero coefficients in an ℓ_1-penalised LR fitted to the feature matrix of model

(iii) (428 features in the training period). Two simpler combiners were also evaluated: soft voting, which averages the base-model probabilities, and a weighted ensemble, which averages them with weights chosen on the validation set. Let $\hat{p}_{m,i}$ denote the predicted probability of base model m for complaint i. The ensemble combines them through a logistic metalearner,

$$\hat{p}_i = \sigma\left(\beta_0 + \sum_{m=1}^{M} \beta_m \hat{p}_{m,i}\right), \qquad \sigma(u) = \frac{1}{1+e^{-u}}, \tag{4}$$

whose coefficients are estimated using ℓ_2-regularised LR on the validation-period predictions from base models fitted in the training period. At test time, the metalearner is applied to predictions from base models refitted on the training and validation periods. Stacking is used because it learns how to weight heterogeneous base models with complementary strengths. Because refitted base models may produce score distributions that differ from those on which the metalearner was trained, the calibration of the ensemble is checked on the test set (Section 5.1). Model selection proceeded in six prespecified rounds, each selecting the model with the highest validation PR-AUC so far. Selection never used the test set, but because test scores were computed as models were run, earlier results were visible when later rounds were designed; we acknowledge this as a source of researcher degrees of freedom. A fresh holdout period could not be added, because the CFPB stopped publishing complaint narratives after July 2026 (Section 3.5), so no later complaints with narratives are available.

4.4 Evaluation, statistical inference and explanation

Because monetary relief is a minority outcome, the primary metric is PR-AUC,

$$\text{PR-AUC} = \int_0^1 \text{Precision}(R)\,\mathrm{d}R, \tag{5}$$

where Precision(R) is the precision at recall R. PR-AUC is more informative than the area under the receiver operating characteristic curve (ROC-AUC) under class imbalance (Saito and Rehmsmeier, 2015), because its baseline equals the positive-class rate rather than 0.5. We also report ROC-AUC, recall and precision among the 10% and 20% highest-scored complaints (fixed review capacities), F1 at the validation-optimal threshold, the Brier score and decile calibration curves. For two models A and B, the test-set difference $\Delta = \text{PR-AUC}_A - \text{PR-AUC}_B$ is assessed with a paired bootstrap (Efron and Tibshirani, 1993): complaints are resampled with replacement 2,000 times, the difference Δ_b^* is recomputed on each resample b, and the percentile interval of $\{\Delta_b^*\}$ gives the 95% confidence interval (CI). The bootstrap requires no distributional assumptions, and resampling complaints in pairs accounts for the correlation between models evaluated on the same complaints. Complaint-level resampling does not, however, account for clustering by company or over time, so the intervals are conditional on the observed test sample. ROC-AUC differences are tested with DeLong's test (DeLong et al, 1988), and the selected model is retrained with seeds 1 to 5 to assess sensitivity to randomness.

A stacking ensemble of heterogeneous models has no single exact SHAP decomposition, so explanations are computed for the LightGBM model on structured, company and LLM features (S+C+G), the strongest single interpretable model; the same model is used for the stability, faithfulness and drift analyses. SHAP decomposes each prediction, on the log-odds scale, into additive feature contributions (Lundberg and Lee, 2017),

$$f(\mathbf{z}_i) = \varphi_0 + \sum_{j=1}^{J} \varphi_{ij}, \tag{6}$$

where φ_0 is the expected model output and φ_{ij} is the contribution of feature j to complaint i, computed exactly with TreeSHAP (Lundberg et al, 2020). Global importance and the share of a feature family $\mathcal{F}$ are defined as

$$I_j = \frac{1}{|\mathcal{D}_{\text{te}}|} \sum_{i \in \mathcal{D}_{\text{te}}} |\varphi_{ij}|, \qquad S_{\mathcal{F}} = \frac{\sum_{j \in \mathcal{F}} I_j}{\sum_{j=1}^{J} I_j}, \tag{7}$$

and four local explanations illustrate individual cases.

Because attributions can vary across runs and periods (Alvarez-Melis and Jaakkola, 2018), stability is measured with Kendall's rank correlation (Kendall, 1938) between two importance rankings computed on the union of their top-20 features,

$$\tau = \frac{n_c - n_d}{\frac{1}{2}\, n(n-1)}, \tag{8}$$

where n is the number of features compared and n_c and n_d are the numbers of concordant and discordant pairs. Rankings are compared across seeds 1 to 5 and across test quarters; a rank-based measure is used because the quantity of interest is feature ordering, not the scale of SHAP values. Because importance methods may not reflect what a model actually relies on (Hooker et al, 2019), faithfulness is assessed with a deletion test. For $k \in \{1, 5, 10, 20, 50\}$, the k features with the highest importance on the training and validation rows are removed, the model is refitted, and the drop in test performance,

$$\Delta_k = \text{PR-AUC}_{\text{full}} - \text{PR-AUC}_{-k}, \tag{9}$$

is compared with the mean drop from removing k randomly chosen features across three draws. Ranking features on the training and validation rows ensures that the test set plays no part in choosing which features are removed.

4.5 Cost-sensitive triage policy and robustness analyses

The decision layer is applied to SCAM-XD and to the strongest benchmark, B1. A policy flags complaint i for review when $\hat{p}_i \geq t$. With an unnecessary review, or false positive (FP), costing one unit, and a missed relief case, or false negative (FN), costing

$r = C_{\text{FN}}/C_{\text{FP}}$ units ($1 \leq r \leq 20$), the expected cost per complaint is

$$\text{EC}(r,t) = \frac{1}{N}\sum_{i=1}^{N}\Big[\,\mathbf{1}\{\hat{p}_i \geq t,\ y_i = 0\} + r \cdot \mathbf{1}\{\hat{p}_i < t,\ y_i = 1\}\,\Big] = \frac{\text{FP}(t) + r \cdot \text{FN}(t)}{N}, \tag{10}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function and reviewing a relief case is treated as costless, because that review is part of providing redress that is owed rather than a cost of a triage error. The true cost ratio is bank-specific and unobserved, so results are reported across $1 \leq r \leq 20$, from a missed relief case costing the same as one review to costing as much as twenty reviews. The analysis models expected cost under these assumptions; it does not measure realised savings. Under this definition, reviewing no complaints costs r times the base rate, and reviewing all complaints costs one minus the base rate, whatever the value of r. Following cost-sensitive decision theory (Elkan, 2001), the threshold is derived from the cost ratio rather than fixed in advance:

$$t^*(r) = \arg\min_{t \in \mathcal{Q}} \text{EC}_{\text{va}}(r,t), \tag{11}$$

where $\mathcal{Q}$ is a grid of 1,001 quantiles of the validation scores and EC_{va} is the expected cost on the validation set. The selected threshold is applied once to the test set and compared with two threshold-only policies, a fixed threshold of 0.5 and the threshold that maximises F1 on the validation set, and with policies that review complaints with the top 10% or top 20% of scores, all complaints or no complaints. Decision curves (Vickers and Elkin, 2006) report the net benefit across threshold probabilities p_t,

$$\text{NB}(p_t) = \frac{\text{TP}(p_t)}{N} - \frac{\text{FP}(p_t)}{N} \cdot \frac{p_t}{1 - p_t}, \tag{12}$$

where $\text{TP}(p_t)$ is the number of true positives (TP) at threshold p_t and $p_t/(1-p_t)$ is the implied ratio of the FP cost to the FN cost.

Three robustness analyses address RQ5. First, temporal drift is examined over five six-month windows $W_1, \ldots, W_5$ from January 2024 to July 2026: for each window W_w, the explanation model is trained on all complaints received before W_w and evaluated on W_w, mirroring periodic retraining in deployment, and its SHAP ranking is compared with that of W_{w-1} using Equation (8). Second, test performance is reported for the ten largest states, for each product and for each consumer tag (older American and servicemember), with bootstrap CIs and the lift of PR-AUC over the subgroup base rate, $\text{PR-AUC}_g/\bar{y}_g$. Third, the main comparison is reestimated on the full set without scope exclusions.

5 Results

This section reports the results in the order of the RQs. Section 5.1 compares the predictive performance of SCAM-XD with the benchmark models, and Section 5.2 assesses the incremental value of the LLM-extracted features (H1). Section 5.3 examines the drivers of predicted redress and the stability and faithfulness of the explanations (H2). Section 5.4 evaluates the expected cost of the triage policies (H3), and Section 5.5

tests robustness over time, across subgroups and to changes in sample scope. Unless stated otherwise, results refer to the out-of-time test set (July 2025 to July 2026; n = 16,208; relief rate 0.1096) and to seed 42, and differences between models are tested with paired bootstrap resampling and the DeLong test.

5.1 Predictive performance (RQ1)

Table 9 reports validation and test performance. SCAM-XD achieves a test PR-AUC of 0.3552 and a ROC-AUC of 0.8237 against a base rate of 0.1096, and the 10% of complaints with the highest scores contain 35.08% of all relief cases. The strongest benchmark, B1, reaches a PR-AUC of 0.3344, and the reimplemented method of Wang et al (2026) (B3) reaches 0.3174. The two transformer benchmarks are the weakest: fine-tuned DistilRoBERTa (B4b) reaches 0.2708 and the model using MiniLM embeddings (B4a) reaches 0.2652. Every model scores lower in the test period than in the validation period, which is consistent with the temporal shift examined in Section 5.5. The drop is largest for B4b (0.3100 to 0.2708).

Table 9 Validation and test performance of SCAM-XD, alternative ensembles and benchmark models (seed 42)

	Val.	Test				
Model	PR-AUC	PR-AUC	ROC-AUC	Recall@10%	F1	Brier
SCAM-XD (stack, 13 models)	0.3833	0.3552	0.8237	0.3508	0.4070	0.0844
Stack, 11 models[a]	0.3828	0.3527	0.8237	0.3502	0.4028	0.0844
Weighted ensemble	0.3779	0.3463	0.8216	0.3457	0.3969	0.0833
B1: S+C+W, LR	0.3553	0.3344	0.8155	0.3378	0.3921	0.0844
B3: Wang et al. (re-impl.)	0.3272	0.3174	0.7940	0.3226	0.3768	0.0858
B2: TF-IDF, XGBoost	0.2941	0.2870	0.7710	0.3041	0.3522	0.0880
B4b: DistilRoBERTa, FT	0.3100	0.2708	0.7560	0.2838	0.3253	0.0924
B4a: MiniLM, LR	0.2680	0.2652	0.7573	0.2793	0.3337	0.0893

Test base rate 0.1096 (n = 16,208). S, structured fields; C, company identity; W, TF-IDF; LR, logistic regression; FT, fine-tuned. [a]SCAM-XD without B4a and B4b. Recall@10%, share of relief cases among the 10% highest-scored complaints; F1 at the threshold that maximises F1 on the validation set.

Figure 2 shows where the advantage lies. SCAM-XD leads at the top of the ranking: reviewing the highest-scored 10% of complaints identifies 35.1% of relief cases at a precision of 0.384, and reviewing 20% identifies 57.6% at 0.316. Beyond a recall of about 0.5, the SCAM-XD and B1 curves largely converge, so the two models perform similarly when about half of all complaints are reviewed (Section 5.4).

Table 10 tests the differences. Against B1, SCAM-XD increases PR-AUC by 0.0208 (95% CI 0.0082–0.0328, $p < 0.001$) and ROC-AUC by 0.0082 (DeLong $p = 0.0015$). Against B3, the improvements are 0.0378 in PR-AUC (95% CI 0.0254–0.0501) and 0.0298 in ROC-AUC, both with $p < 0.001$. Adding the two transformer models to the stack, however, makes no significant difference: compared with the 11-model stack, PR-AUC rises by 0.0025 ($p = 0.055$) and ROC-AUC is unchanged.

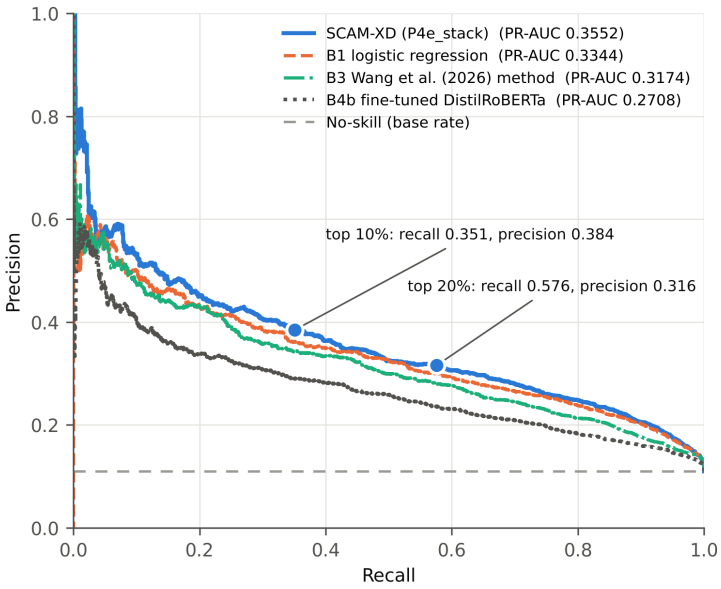


Fig. 2 Test-set precision–recall curves for SCAM-XD and three benchmarks

Table 10 Paired comparison of SCAM-XD with benchmark models and with the stack without transformer models (test set, seed 42)

Comparison	ΔPR-AUC	95% CI	p (boot.)	ΔROC-AUC	p (DeLong)
vs B1	0.0208	[0.0082, 0.0328]	<0.001	0.0082	0.0015
vs B3	0.0378	[0.0254, 0.0501]	<0.001	0.0298	<0.001
vs 11-model stack	0.0025	[−0.00003, 0.00489]	0.055	0.0000	0.999
vs firm-rate baseline	0.1173	[0.1011, 0.1350]	<0.001	–	–

p (boot.), paired bootstrap with 2,000 resamples. Firm-rate baseline: each complaint is scored with its company's relief rate in the training and validation periods (test PR-AUC 0.2378, ROC-AUC 0.7611).

The improvement over B1 holds across random seeds (Table 11). It ranges from 0.0172 to 0.0227, its CI excludes zero for every seed, and $p \leq 0.003$ in each case. Over the five seeds, test PR-AUC averages 0.3542 (SD 0.0023), against 0.3123 (SD 0.0026) for B3. In these runs, the 11 classical base models and the metalearner are retrained for each seed, while B4a and B4b are held at seed 42 because of the cost of fine-tuning. The improvement over B1 is therefore consistent across seeds.

Among the learners trained on the same features (Table 12), stacking had the highest validation PR-AUC, followed by extra trees. In the test period, however, LR (0.3389) outperformed extra trees (0.3269), which shows that validation rankings among close competitors do not always carry over to a later period. Naive Bayes was clearly the weakest. SCAM-XD and B1 have the same Brier score (0.0844); their calibration curves are shown in Figure 3.

Table 11 SCAM-XD across random seeds and its test PR-AUC gain over B1

Seed	Val. PR-AUC	Test PR-AUC	Test ROC-AUC	ΔPR-AUC	95% CI	p
1	0.3836	0.3515	0.8198	0.0172	[0.0057, 0.0292]	0.003
2	0.3848	0.3541	0.8234	0.0197	[0.0068, 0.0324]	0.001
3	0.3830	0.3571	0.8232	0.0227	[0.0105, 0.0340]	<0.001
4	0.3806	0.3526	0.8222	0.0183	[0.0060, 0.0298]	0.002
5	0.3832	0.3557	0.8267	0.0214	[0.0089, 0.0334]	<0.001
Mean	0.3830	0.3542	0.8230	0.0199		
SD	0.0015	0.0023	0.0025	0.0023		

B4a and B4b are fixed at seed 42; the other 11 base models and the metalearner are retrained for each seed.

Table 12 Comparison of learning algorithms on structured, company, TF-IDF and LLM features, ranked by validation PR-AUC

Learner	Val. PR-AUC	Test PR-AUC	Test ROC-AUC	Precision@10%	Brier
Stacking	0.3783	0.3463	0.8198	0.3769	0.0846
Extra trees	0.3752	0.3269	0.8119	0.3603	0.0845
Soft voting	0.3653	0.3291	0.8045	0.3566	0.0915
Logistic regression	0.3616	0.3389	0.8162	0.3683	0.0841
LightGBM	0.3613	0.3349	0.8107	0.3609	0.1004
XGBoost	0.3592	0.3301	0.8049	0.3634	0.1027
Linear SVM	0.3583	0.3309	0.8109	0.3560	0.0850
Random forest	0.3479	0.3189	0.7952	0.3634	0.0856
Naive Bayes	0.3063	0.2632	0.7456	0.2881	0.2227

5.2 Incremental value of LLM-extracted features (RQ2, H1)

Table 13 isolates the contribution of each feature family. On their own, the LLM features are weaker than TF-IDF (test PR-AUC 0.1949 against 0.2954 with LightGBM) and only slightly stronger than the structured fields (0.1852). Their value appears in combination with structured fields and company identity. Adding them to S+C raises test PR-AUC from 0.2874 to 0.3274 with LightGBM and from 0.2922 to 0.3386 with XGBoost. With 38 variables in place of a large TF-IDF vocabulary, the resulting S+C+G models match the corresponding S+C+W models on the test set (0.3274 and 0.3386 against 0.3281 and 0.3240) and outperform them on the validation set (0.3681 against 0.3452 with LightGBM).

Once all other families are present, the LLM features add little. Adding them to the full LightGBM model raises test PR-AUC from 0.3303 to 0.3326, a difference that is not statistically significant (0.0023, 95% CI −0.0062 to 0.0101, $p = 0.65$), although ROC-AUC improves by 0.0052 (DeLong $p = 0.004$). H1 is therefore not supported: LLM features do not significantly improve PR-AUC for a model that already contains all other text representations, and the gain of SCAM-XD over B1 and B3 (Table 10) reflects the ensemble as a whole rather than the LLM features alone. Their supported contribution is parsimony: with 286 features instead of 50,248, the S+C+G models match the corresponding lexical models on the test set and outperform them on the validation set, so the LLM features provide a compact, interpretable alternative to a high-dimensional lexical representation.

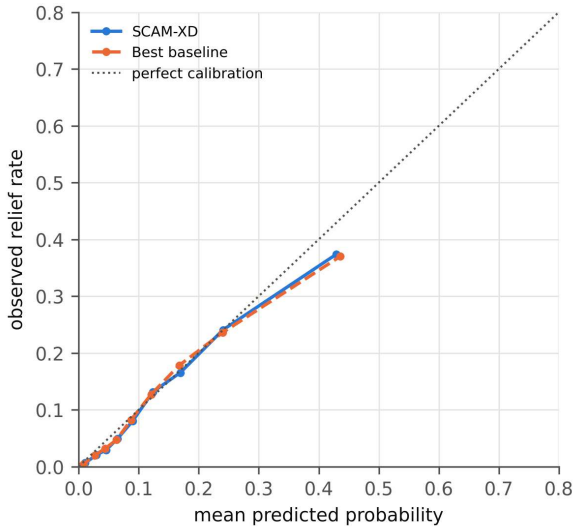


Fig. 3 Calibration of SCAM-XD and B1 on the test set (decile bins)

Table 13 Feature-family ablation (PR-AUC)

		LightGBM		XGBoost	
Families	Description	Val.	Test	Val.	Test
S	Structured fields only	0.2216	0.1852	0.2247	0.1877
G	LLM features only	0.2308	0.1949	0.2323	0.1982
W	TF-IDF only	0.2981	0.2954	0.2941	0.2870
S+C	Structured + company	0.3333	0.2874	0.3336	0.2922
S+C+W	+ TF-IDF	0.3452	0.3281	0.3478	0.3240
S+C+G	+ LLM features (no TF-IDF)	0.3681	0.3274	0.3692	0.3386
S+C+W+T+L	+ TF-IDF, text features, LDA	0.3442	0.3261	0.3431	0.3246
S+C+W+T+L+X	+ stacked text score	0.3397	0.3303	0.3427	0.3253
S+C+W+T+L+X+G	+ LLM features (all families)	0.3488	0.3326	0.3451	0.3352

5.3 Drivers of predicted redress and explanation stability (RQ3, H2)

Explanations are computed for the LightGBM model on structured, company and LLM features (Section 4.4). This model's test scores are closely aligned with those of SCAM-XD (Spearman correlation 0.949, Pearson 0.928), and it receives the largest coefficient in the metalearner, although the coefficients of highly correlated base models should not be interpreted as measures of importance. Company identity accounts for 41.68% of the total mean absolute SHAP value, LLM features for 31.97% and structured fields for 26.35% (Figure 4). The most important single feature is whether the complaint concerns Block, Inc. (mean |SHAP| 0.5947), which strongly lowers predicted relief, consistent with that company's relief rate of 0.10% (Table 7); the indicator for complaints against Bank of America (0.2009) raises it. The highest-ranked LLM features are a card payment rail (0.1232, third overall), the consumer's report of escalating the complaint within the institution (0.1003, fifth), an unknown amount lost

(0.0905) and the ordinal amount lost (0.0865). Other bank-response signals rank lower but act in plausible directions (Figure 5): prompt reporting by the consumer raises predicted relief, and account closure lowers it. Two further patterns are consistent with the liability rules. Card payments raise predicted relief, in line with the dispute rights attached to card payments (Consumer Financial Protection Bureau, 2026d,c), and narratives that the LLM codes as not describing a scam also raise it, which suggests that disputes outside the scam setting, such as billing or card disputes, are more readily remedied than authorised scam payments.

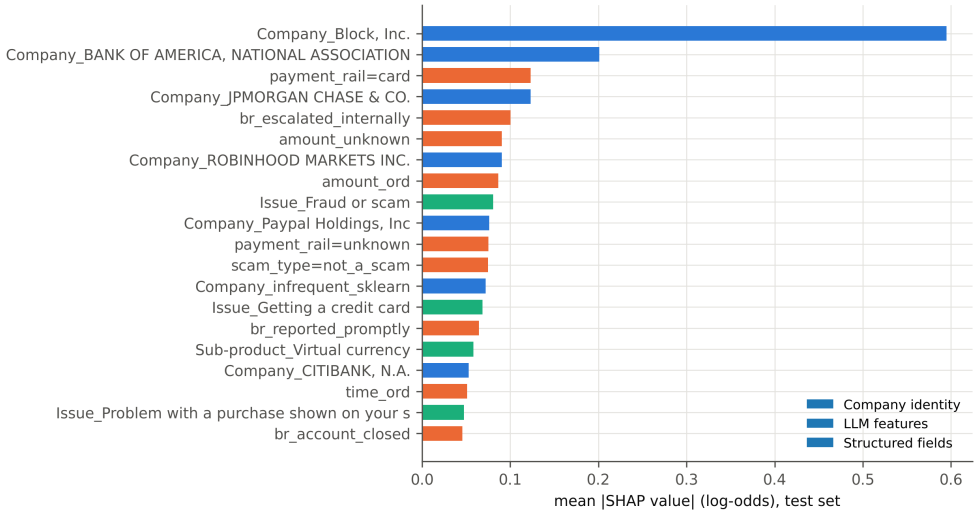


Fig. 4 Global SHAP feature importance (top 20)

H2 is partially supported. Bank-response cues (escalation within the institution, prompt reporting, account closure) and the payment rail rank among the top 20 contributors, so features beyond company identity matter. Scam typology contributes less: the highest-ranked scam-type indicator, whether the narrative describes a scam at all, ranks twelfth, and specific types such as investment or crypto scams fall outside the top 20. This is consistent with the composition of the modelling set, in which fewer than two in five complaints were coded as deception scams (Table 3). Company identity, rather than the narrative, is the largest single source of attributed importance. Company identity alone, however, is far from sufficient: a baseline that scores each complaint with its company's historical relief rate reaches a test PR-AUC of only 0.2378, and SCAM-XD exceeds it by 0.1173 (95% CI 0.1011–0.1350, $p < 0.001$; Table 10). Local explanations show how strongly company identity can dominate individual predictions. For one relief case scored at 0.0064, the Block, Inc. indicator contributed −2.884 log-odds and outweighed every narrative signal. For a relief case at the median score among positive cases (0.2217), a partial refund already given by the bank made the largest positive contribution (+0.465).

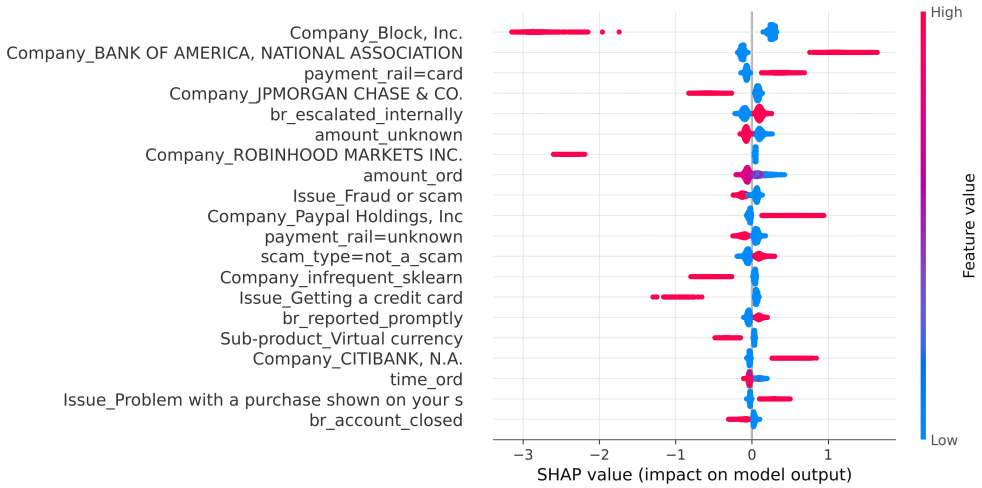


Fig. 5 SHAP summary (beeswarm) plot for the top 20 features on the test set

The explanations of the companion model are stable. Across seeds, the mean Kendall's τ between top-20 rankings is 0.7168 (minimum 0.6364), and on average 89.50% of the top-20 features are shared; between test quarters, the mean τ is 0.8805 (minimum 0.8095). The deletion test indicates that the top-ranked features carry predictive information for this model (Figure 6). Removing the top 5 SHAP features lowers test PR-AUC by 0.0463, against 0.0018 for 5 random features; removing the top 50 lowers it by 0.1306, against 0.0270. Removing only the top feature has no effect (-0.0001), which is consistent with correlated features carrying the information in the Block, Inc. indicator.

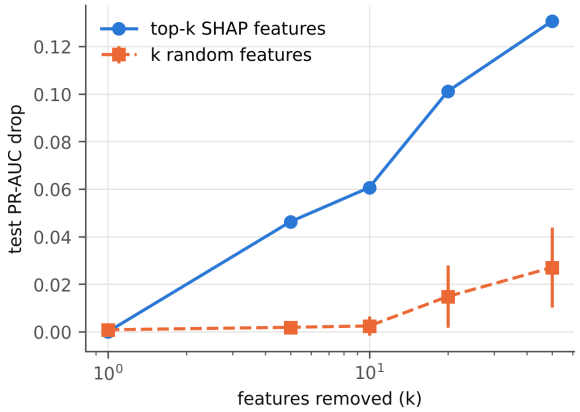


Fig. 6 Drop in test PR-AUC when removing top-k SHAP versus random features

5.4 Decision value of cost-sensitive triage (RQ4, H3)

Table 14 and Figure 7 report the expected cost per complaint under each policy. When a missed relief case costs no more than a review ($r = 1$), no policy improves materially on reviewing nothing (0.1090 against 0.1096). As r increases, the cost-optimal threshold chosen on the validation set becomes clearly preferable. At $r = 10$, SCAM-XD with this threshold reviews 46.1% of complaints, identifies 88.9% of relief cases and has an expected cost of 0.4861. This compares with 0.9276 for the fixed 0.5 threshold, which identifies only 17.2% of relief cases, 0.6015 for reviewing the top 20% and 1.0958 for reviewing nothing. At $r = 20$, the corresponding costs are 0.6009, 1.8345, 1.0661 and 2.1915. H3 is therefore supported under the specified cost model: validation-selected thresholds substantially lower modelled expected cost relative to threshold-only classification. Because predicted probabilities rarely exceed 0.5, the fixed threshold flags only 3.9% of complaints and is an undemanding comparator. The result also holds against a stronger threshold-only policy: the threshold that maximises F1 on the validation set reviews 23.2% of test complaints and identifies 63.5% of relief cases, but its expected cost is higher than that of the cost-optimal threshold at every cost ratio (0.5626 against 0.4861 at $r = 10$, and 0.9624 against 0.6009 at $r = 20$). The cost-optimal threshold is also cheaper than reviewing the top 10% or the top 20% of complaints at every cost ratio (Table 14).

Table 14 Expected cost per complaint by triage policy and cost ratio r

Model	r	Cost-opt.	Fixed 0.5	F1-opt.	Top 10%	Top 20%	None	All
SCAM-XD	1	0.1090	0.1113	0.2028	0.1327	0.1834	0.1096	0.8904
SCAM-XD	2	0.2014	0.2020	0.2428	0.2038	0.2298	0.2192	0.8904
SCAM-XD	5	0.3616	0.4741	0.3627	0.4173	0.3692	0.5479	0.8904
SCAM-XD	10	0.4861	0.9276	0.5626	0.7730	0.6015	1.0958	0.8904
SCAM-XD	20	0.6009	1.8345	0.9624	1.4843	1.0661	2.1915	0.8904
B1	1	0.1084	0.1105	0.1647	0.1356	0.1884	0.1096	0.8904
B1	2	0.2046	0.2080	0.2212	0.2081	0.2374	0.2192	0.8904
B1	5	0.3771	0.5004	0.3905	0.4258	0.3844	0.5479	0.8904
B1	10	0.4967	0.9878	0.6728	0.7886	0.6293	1.0958	0.8904
B1	20	0.6001	1.9627	1.2374	1.5141	1.1192	2.1915	0.8904

Expected cost = $(\mathrm{FP}+r\cdot\mathrm{FN})/N$; reviewing a complaint that merits relief costs nothing. Cost-opt. and F1-opt., thresholds that minimise expected cost or maximise F1 on the validation set; None and All, reviewing no or all complaints.

Most of the decision value comes from the policy rather than from the choice of model. With cost-optimal thresholds, SCAM-XD has a lower expected cost than B1 at $r = 2$, 5 and 10 (at $r = 10$, 0.4861 against 0.4967), but a slightly higher one at $r = 1$ (0.1090 against 0.1084) and $r = 20$ (0.6009 against 0.6001). The difference is clearer under fixed review capacity: when the top 20% of complaints are reviewed, SCAM-XD identifies 57.6% of relief cases and B1 55.3%. On the test set, this amounts to about 41 additional relief cases identified for the same review effort, or about 25 per 10,000 complaints at the test base rate. The decision curves of the two models largely

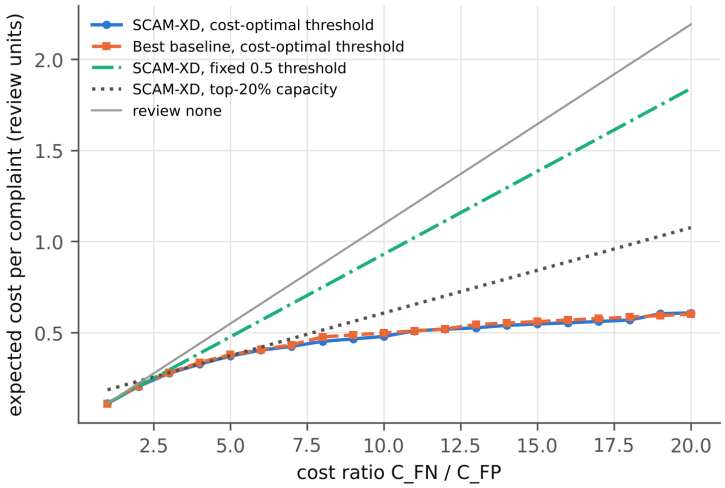


Fig. 7 Expected cost per complaint across cost ratios

overlap (Figure 8), and both lie above the "review all" and "review none" strategies over most of the threshold range.

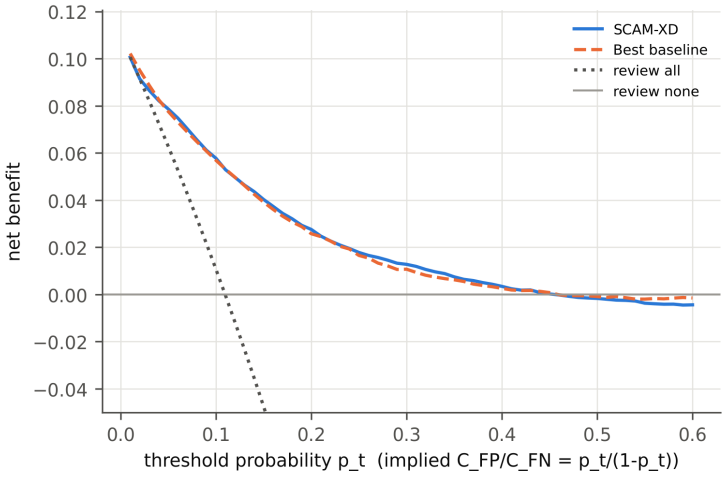


Fig. 8 Decision curves for SCAM-XD and B1 on the test set

5.5 Temporal and subgroup robustness (RQ5)

Table 15 reports the performance of the explanation model, retrained on all earlier data, over rolling six-month windows. PR-AUC declines from 0.3892 in the first half of 2024 to 0.3315 in the second half of 2025, then recovers to 0.3551 in 2026. This pattern partly follows the base rate, which falls from 0.1347 to 0.1028 over the same

period, while ROC-AUC rises from 0.7776 to 0.8227. Explanations shift more between retrained models than within a fixed model: Kendall's τ between consecutive windows ranges from 0.5290 to 0.6710, below the 0.8805 observed between test quarters for a single model.

Table 15 Performance over rolling six-month windows

Window	n_{train}	n_{window}	Base rate	PR-AUC	95% CI	ROC-AUC	τ (prev.)
2024 H1	11,288	5,228	0.1347	0.3892	[0.3539, 0.4258]	0.7776	–
2024 H2	16,516	5,704	0.1327	0.3825	[0.3503, 0.4182]	0.7986	0.5290
2025 H1	22,220	8,387	0.1059	0.3735	[0.3431, 0.4060]	0.8356	0.6364
2025 H2	30,607	8,148	0.1028	0.3315	[0.3017, 0.3619]	0.8153	0.5968
2026 Jan–Jul	38,755	8,060	0.1164	0.3551	[0.3274, 0.3847]	0.8227	0.6710

Performance also differs by product. Table 16 reports subgroup results for the explanation model. PR-AUC is highest for credit cards (0.4116), where the base rate is also highest (0.1892), and lowest for money transfers (0.2956), where the base rate is lowest (0.0798). Relative to its base rate, however, the money-transfer model performs best, with a PR-AUC lift of 3.71 and a ROC-AUC of 0.8537. This matters for review coverage: when the global top 20% of scores is reviewed, 81.62% of credit-card relief cases are flagged but only 40.86% of money-transfer relief cases. Across the ten largest states, PR-AUC ranges from 0.2652 (Georgia) to 0.4393 (Illinois), with wide and overlapping CIs. The same pattern holds for SCAM-XD, which outperforms B1 in every product (credit cards 0.437 against 0.414; checking or savings 0.345 against 0.323; money transfers 0.312 against 0.280) and whose PR-AUC across the ten largest states ranges from 0.288 (Georgia) to 0.438 (Illinois).

Table 17 reports performance by consumer tag. SCAM-XD's PR-AUC is similar across groups and at least as high as B1's for older Americans and untagged complaints. For servicemembers, however, relief cases are less often among SCAM-XD's highest-ranked 20% of complaints (0.508) than among B1's (0.564), compared with 0.586 for untagged relief cases under SCAM-XD. That said, the PR-AUC CIs overlap and the group is small.

Finally, rerunning the comparison on the full sample without scope exclusions preserves the ordering (Table 18). On 17,570 test complaints with a base rate of 0.1012, the LightGBM S+C+G model reaches a PR-AUC of 0.3288, against 0.3191 for B3.

6 Discussion

This section interprets the results in light of the research questions and hypotheses. It first summarises the principal findings, then considers what they imply for research on complaint outcomes and LLM-based text coding, and for banks and regulators managing scam redress. It closes with the limitations of the study and directions for future research.

Table 16 Test performance of the explanation model by state (ten largest) and product

Group	n	Base	PR-AUC	95% CI	Lift	ROC-AUC	Rec.@20%
State							
CA	2,271	0.1193	0.3114	[0.2664, 0.3681]	2.61	0.7872	0.5867
FL	1,563	0.1094	0.3573	[0.2995, 0.4324]	3.27	0.8141	0.5848
TX	1,483	0.1065	0.3685	[0.2939, 0.4515]	3.46	0.8264	0.5696
NY	1,075	0.1079	0.3041	[0.2368, 0.3988]	2.82	0.7851	0.5000
GA	780	0.0872	0.2652	[0.1963, 0.3681]	3.04	0.8152	0.4265
PA	554	0.1065	0.3207	[0.2299, 0.4474]	3.01	0.8049	0.5593
IL	543	0.1197	0.4393	[0.3346, 0.5591]	3.67	0.8672	0.6769
NC	497	0.0885	0.3190	[0.2138, 0.4697]	3.60	0.8296	0.5000
NJ	487	0.1437	0.4196	[0.3195, 0.5340]	2.92	0.7912	0.5286
VA	454	0.1167	0.2773	[0.1958, 0.3968]	2.38	0.7839	0.4906
Product							
Checking/savings	5,429	0.1160	0.3199	[0.2853, 0.3585]	2.76	0.7710	0.5175
Money transfer	8,163	0.0798	0.2956	[0.2688, 0.3281]	3.71	0.8537	0.4086
Credit card	2,616	0.1892	0.4116	[0.3687, 0.4550]	2.18	0.7639	0.8162

Table 17 Test performance by consumer tag for SCAM-XD and B1

			SCAM-XD			B1		
Group	n	Base	PR-AUC	[95% CI]	Rec.@20%	PR-AUC	[95% CI]	Rec.@20%
Older American	1,824	0.138	0.368	[0.313, 0.429]	0.574	0.340	[0.291, 0.404]	0.570
Servicemember	1,707	0.106	0.336	[0.277, 0.404]	0.508	0.340	[0.280, 0.406]	0.564
Both tags	346	0.133	0.383	[0.265, 0.517]	0.587	0.419	[0.282, 0.548]	0.587
No tag	13,023	0.107	0.357	[0.334, 0.382]	0.586	0.338	[0.316, 0.362]	0.550

Older American and Servicemember rows include complaints with both tags. Base, group base rate; Rec.@20%, share of the group's relief cases among the global top 20% of scores. CIs from 1,000 bootstrap resamples.

Table 18 Test performance on the full sample without scope exclusions

Model	Test n	Base rate	PR-AUC	ROC-AUC	Recall@10%
LightGBM, S+C+G	17,570	0.1012	0.3288	0.8256	0.3498
B3 (Wang et al. method)	17,570	0.1012	0.3191	0.8106	0.3487

6.1 Principal findings

Table 19 summarises the results of the hypothesis tests. With respect to RQ1, monetary redress for complaints flagged as fraud- or scam-related can be predicted substantially above the base rate using only information available at the time of submission. The gain of the full ensemble over a well-specified linear baseline is statistically significant and concentrated among the highest-ranked complaints. LLM-extracted features are most useful as a compact, interpretable representation that performs almost as well as a large lexical one (RQ2). Predictions are driven first by company identity and then by payment-rail and bank-response cues, and the companion-model

explanations are stable across seeds and test periods and pass a deletion test (RQ3). Under assumed costs, validation-selected thresholds substantially reduce modelled expected cost (RQ4). Performance holds over time and across subgroups, but the share of relief cases flagged for review differs between products and is lower for servicemembers (RQ5).

Table 19 Summary of hypothesis tests

Hypothesis	Result	Evidence
H1	Not supported; LLM features add parsimony, not accuracy	LLM features do not significantly improve the full LightGBM model ($p = 0.65$); with 286 instead of 50,248 features, S+C+G models match the corresponding TF-IDF models on the test set.
H2	Partially supported	Payment-rail and bank-response cues rank in the top 20; company identity accounts for 41.68% of SHAP importance, and scam typology ranks lower.
H3	Supported	At $r = 10$, the cost-optimal threshold yields an expected cost of 0.4861, against 0.9276 for a fixed 0.5 threshold and 0.5626 for the F1-optimal threshold; differences between models at the decision level are small.

6.2 Theoretical implications

These findings extend research on complaint outcomes (Siering, 2022) to scams and treat the redress stage as a decision problem in its own right, distinct from the detection problems that dominate scam research (Gopal et al, 2022; Gunathilaka et al, 2024; Tang et al, 2024; Ye et al, 2026; Liu et al, 2026). Three implications apply beyond this setting. First, LLM-extracted features should be judged by what they add to strong existing representations, not only by comparison with weak baselines. Here, their main value was interpretability at near parity with lexical models of up to 50,248 features, rather than additional accuracy, so studies that use LLMs as coders should report ablations against conventional text representations as a matter of course (Ziems et al, 2024). Second, the dominance of company identity indicates that redress depends heavily on the institution, not only on the merits of the complaint as the consumer describes it. Among the largest firms, relief rates range from 0.00% (Robinhood) to 26.70% (Bank of America). Firm identity alone is nonetheless far from sufficient: scoring complaints by their company's historical relief rate reaches a PR-AUC of 0.2378, against 0.3552 for SCAM-XD, so the content of the complaint matters as well. A predictive model can expose this heterogeneity but cannot establish whether it reflects differences in customers, products, policies or compliance, which would require causal designs. The direction of the narrative effects is nonetheless consistent with the legal asymmetry described in Section 1: complaints whose facts fall under established dispute rights, such as card payments and disputes that do not involve a scam, are more likely to be remedied than authorised payments, for which US consumers have no

comparable protection (Consumer Financial Protection Bureau, 2026c,d). LLM coding therefore makes visible how liability rules shape redress, which lexical features cannot show directly. Third, the evaluation design matters. Out-of-time testing, a metric suited to class imbalance, repeated seeds with paired tests, and checks of explanation stability and deletion behaviour give a more conservative picture than accuracy on a single balanced split. Validation rankings among close competitors did not always carry over to the test period, which underlines the importance of keeping the test set out of model selection.

6.3 Practical and regulatory implications

For banks, the results support a triage design in which complaints are ranked by predicted relief and reviewed under an explicit cost ratio. The choice of policy matters far more than the choice among competitive models. At an assumed cost ratio of 10:1, moving from a fixed 0.5 threshold to a validation-selected threshold roughly halves modelled expected cost, whereas replacing B1 with SCAM-XD reduces it by 0.0106. The model matters more when review capacity is fixed: reviewing the highest-ranked 20% of complaints, SCAM-XD identifies 57.6% of relief cases against 55.3% for B1, or about 25 additional relief cases per 10,000 complaints. Where simplicity and exact explanations matter more than the last increment of accuracy, the LightGBM explanation model offers a practical alternative: with a wider hyperparameter search than the S+C+G model in Table 13 (0.3274), it reaches a test PR-AUC of 0.3364, close to that of the 13-model stack. Banks should also monitor coverage by product, because a single global threshold flags a far smaller share of money-transfer relief cases than credit-card relief cases, and money transfers are the rail most associated with authorised scams. Coverage should be monitored by consumer group as well: servicemember relief cases are ranked lower by SCAM-XD than by B1, which a bank would need to investigate before relying on the model. For regulators, the explanations provide a supervisory lens in the spirit of RegTech (Siering, 2022). Large and stable differences in reported relief rates between firms may warrant supervisory attention, although they do not by themselves establish that comparable complaints were treated differently. Under regimes that make reimbursement mandatory, such as the UK requirement for APP scams (Payment Systems Regulator, 2023), or that impose prevention and response obligations, such as Australia's Scams Prevention Framework (Commonwealth of Australia, 2025), similar models could monitor whether redress decisions are consistent with firms' obligations. Transferring the models themselves would require local data, because US redress outcomes reflect US rules on authorised and unauthorised electronic transfers (Consumer Financial Protection Bureau, 2026c).

6.4 Limitations and directions for future research

Several limitations qualify these findings. First, the data are not representative of all scam victims: they cover only consumers who complained to a single US regulator and consented to publication of their narratives (Consumer Financial Protection Bureau, 2026a), and narratives are available only for complaints received up to 29 July 2026.

Second, the scam filter is broad: human coding indicates that about half of the modelling set (47.3%) concerns fraud or a scam, and only 34.3% of those admitted by narrative keywords alone do, so the findings describe complaints flagged as potentially fraud-related rather than scam complaints alone. The LLM codes agree closely with human coding for scam type but less closely for payment rail and authorisation, and the escalation flag records the consumer's escalation within the institution rather than escalation by the bank. LLM coding was checked against two human coders on 200 complaints, and all codes depend on one model and prompt version (Claude Haiku 5.5); they are therefore signals of what the consumer reports, not verified facts. Third, the outcome is monetary relief as reported by the company rather than verified reimbursement; partial or nonmonetary remedies are not captured, and no escalation label exists from 2023 onwards. Narrative cues such as a partial refund already given may partly anticipate this outcome, and an ablation of the bank-response cues would quantify their contribution.

Fourth, the explanations are computed for a LightGBM model rather than the 13-model stacking ensemble. The two models' test scores are closely aligned (Spearman correlation 0.95), but the explanations describe a related model rather than SCAM-XD itself, and company identity, the strongest predictor, may encode unobserved case mix and firm policies that change over time; performance on companies unseen in training was not evaluated. Model selection also proceeded over rounds in which earlier test results were visible (Section 4.3), so the test estimates may be slightly optimistic, and the cost model is illustrative and omits staffing constraints. Fifth, computing resources limited the transformer baselines: DistilRoBERTa was fine-tuned on a central processing unit (CPU) with 256 tokens, two epochs and one seed. Sixth, the findings are associational and do not explain why firms grant redress.

Future research should refine the scam filter, for example by combining the issue rule with LLM screening, extend human validation to more coders and fields, compare LLMs and prompts, and test larger transformer models on full-length narratives. Further work could broaden the fairness analysis, including by examining why servicemember relief cases are ranked lower; involve complaint-handling experts in reviewing the explanations; apply the framework in other jurisdictions, including those with mandatory reimbursement; and use causal designs to separate firm effects from case mix. Finally, because LLM-assisted complaints have been found to be more likely to obtain relief (Shin et al, 2026), the effect of AI-assisted writing on both redress outcomes and the narrative signals used by SCAM-XD deserves study.

7 Conclusion

This study developed SCAM-XD, an explainable, LLM-augmented framework that predicts whether a fraud- or scam-related complaint to the CFPB will close with company-reported monetary relief and converts that prediction into a cost-sensitive triage decision. In an out-of-time evaluation on complaints from the CFPB narratives archive flagged as fraud- or scam-related, about half of which human coding confirmed as fraud or scams, SCAM-XD achieved a statistically significant and consistent improvement over strong benchmark models, concentrated among the highest-ranked

complaints. The LLM-extracted features offered an interpretable representation that performed almost as well as a high-dimensional lexical model but did not significantly improve the full model, so their principal value lies in interpretability rather than predictive accuracy. Company identity was the dominant driver of predicted redress, yet firm identity alone predicted far less well than the full model, and payment-rail and bank-response cues acted in directions consistent with the liability rules that govern authorised and unauthorised payments. Explanations of a companion LightGBM model were stable across random seeds and test periods and passed a deletion test. Under assumed costs, validation-selected thresholds reduced modelled expected cost substantially relative to threshold-only policies, whereas the choice among competitive models had a much smaller effect, which suggests that in complaint triage the decision rule matters more than the model. As regulators shift liability for scam losses towards banks and the CFPB's complaint narratives move into an archive, the benchmark and evaluation design reported here provide a foundation for research and practice on scam redress.

Abbreviations

Abbreviation	Definition
AI	Artificial intelligence
APP	Authorised push payment
AUC	Area under the curve
CFPB	Consumer Financial Protection Bureau
CI	Confidence interval
CPU	Central processing unit
DeFi	Decentralised finance
FN	False negative
FP	False positive
FT	Fine-tuned
JSON	JavaScript Object Notation
LDA	Latent Dirichlet allocation
LLM	Large language model
LR	Logistic regression
PR-AUC	Area under the precision–recall curve
RegTech	Regulatory technology
ROC-AUC	Area under the receiver operating characteristic curve
RQ	Research question
SD	Standard deviation
SHAP	SHapley Additive exPlanations
SVM	Support vector machine
TF-IDF	Term frequency–inverse document frequency
TP	True positive
UK	United Kingdom
US	United States
XAI	Explainable artificial intelligence
Feature families	
S	Structured complaint fields
C	Company identity
W	TF-IDF word unigrams and bigrams
T	Handcrafted text features
L	Latent Dirichlet allocation topics
X	Out-of-fold stacked text score
G	LLM-extracted features
Benchmark models	
B1	Logistic regression on S+C+W
B2	TF-IDF with XGBoost
B3	Reimplementation of Wang et al (2026)
B4a	MiniLM sentence embeddings with logistic regression
B4b	Fine-tuned DistilRoBERTa

Declarations

- **Funding:** No funding was received for conducting this study.
- **Conflict of interest:** The authors declare that they have no competing interests.
- **Ethics approval and consent to participate:** Not applicable. The study uses publicly released, deidentified complaint records published by the US Consumer Financial Protection Bureau.
- **Consent for publication:** Not applicable.
- **Clinical trial number:** Not applicable.
- **Data availability:** The complaint data analysed in this study are publicly available from the Consumer Financial Protection Bureau (CFPB) Consumer Complaint Database narratives archive, https://www.consumerfinance.gov/foia-requests/foia-electronic-reading-room/cfpb-consumer-complaint-database-narratives-archive/ (Consumer Financial Protection Bureau, 2026a). The study uses archive exports 4 to 21, each of which can be downloaded directly as a ZIP file:
  - https://files.consumerfinance.gov/f/documents/CCDB_Export_4_November_2022_through_August_2023.zip
  - https://files.consumerfinance.gov/f/documents/CCDB_Export_5_September_2023_through_March_2024.zip
  - https://files.consumerfinance.gov/f/documents/CCDB_Export_6_April_2024_through_July_2024.zip
  - https://files.consumerfinance.gov/f/documents/CCDB_Export_7_August_2024_through_October_2024.zip
  - https://files.consumerfinance.gov/f/documents/CCDB_Export_8_November_2024_through_December_2024.zip
  - https://files.consumerfinance.gov/f/documents/CCDB_Export_9_January_2025_through_February_2025.zip
  - https://files.consumerfinance.gov/f/documents/CCDB_Export_10_March_2025_through_April_2025.zip
  - https://files.consumerfinance.gov/f/documents/CCDB_Export_11_May_2025_through_June_2025.zip
  - https://files.consumerfinance.gov/f/documents/CCDB_Export_12_July_2025_through_August_2025.zip
  - https://files.consumerfinance.gov/f/documents/CCDB_Export_13_September_2025_through_October_2025.zip
  - https://files.consumerfinance.gov/f/documents/CCDB_Export_14_November_2025_through_December_2025.zip
  - https://files.consumerfinance.gov/f/documents/CCDB_Export_15_January_2026_through_February_2026.zip
  - https://files.consumerfinance.gov/f/documents/CCDB_Export_16_March_2026.zip
  - https://files.consumerfinance.gov/f/documents/CCDB_Export_17_April_2026.zip
  - https://files.consumerfinance.gov/f/documents/CCDB_Export_18_May_2026.zip
  - https://files.consumerfinance.gov/f/documents/CCDB_Export_19_June_2026.zip
  - https://files.consumerfinance.gov/f/documents/CCDB_Export_20_July_2026.zip

– https://files.consumerfinance.gov/f/documents/CCDB_Export_21_August_2026.zip

The complaint identifiers of the analytical sample and the filtering rules are available from the corresponding author upon reasonable request.

- **Code availability:** The code, the LLM extraction prompt and coding schema, and full implementation details, including hyperparameter search spaces, are available from the corresponding author upon reasonable request.
- **Use of generative AI:** Claude Haiku 5.5 (Anthropic) was used as a research instrument to code complaint narratives (Section 4.2). Claude (Anthropic) assisted with analysis code and writing. The authors verified all content and take full responsibility for the manuscript.
- **Author contributions:** T.A.-Q.: conceptualisation, methodology, software, data curation, formal analysis, investigation, visualisation, writing – original draft, writing – review and editing. R.N. and U.R.S.: validation (independent manual coding of complaint narratives), writing – review and editing. O.S.A., N.A.H., I.M.T. and J.A.: writing – review and editing. All authors read and approved the final manuscript.

References

Alvarez-Melis D, Jaakkola TS (2018) On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049, URL https://arxiv.org/abs/1806.08049

Ayres I, Lingwall J, Steinway S (2014) Skeletons in the database: An early analysis of the CFPB's consumer complaints. Fordham Journal of Corporate & Financial Law 19(2):343

Bastani K, Namavari H, Shaffer J (2019) Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Systems with Applications 127:256–271. https://doi.org/10.1016/j.eswa.2019.03.001

Bera D, Ogbanufe O, Kim DJ (2023) Towards a thematic dimensional framework of online fraud: An exploration of fraudulent email attack tactics and intentions. Decision Support Systems 171:113977. https://doi.org/10.1016/j.dss.2023.113977

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. Journal of Machine Learning Research 3:993–1022

Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 785–794, https://doi.org/10.1145/2939672.2939785

Commonwealth of Australia (2025) Scams Prevention Framework Act 2025 (Cth), no. 15, 2025. Federal Register of Legislation, URL https://www.legislation.gov.au/C2025A00015

Consumer Financial Protection Bureau (2024) Complaint, Consumer Financial Protection Bureau v. Early Warning Services, LLC, Bank of America, N.A., JPMorgan Chase Bank, N.A., and Wells Fargo Bank, N.A., no. 2:24-cv-03652 (D. Ariz.). United States District Court for the District of Arizona, filed 20 December 2024, URL https://files.consumerfinance.gov/f/documents/cfpb_Zelle-Complaint_2024-12.pdf

Consumer Financial Protection Bureau (2026a) CFPB consumer complaint database narratives archive. FOIA Electronic Reading Room, URL https://www.consumerfinance.gov/foia-requests/foia-electronic-reading-room/cfpb-consumer-complaint-database-narratives-archive/

Consumer Financial Protection Bureau (2026b) Consumer complaint database: Release notes. URL https://cfpb.github.io/api/ccdb/release-notes.html

Consumer Financial Protection Bureau (2026c) Electronic fund transfers (Regulation E), 12 CFR part 1005. Electronic Code of Federal Regulations, URL https://www.ecfr.gov/current/title-12/chapter-X/part-1005

Consumer Financial Protection Bureau (2026d) Truth in lending (Regulation Z), 12 CFR part 1026. Electronic Code of Federal Regulations, URL https://www.ecfr.gov/current/title-12/chapter-X/part-1026

DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 44(3):837–845. https://doi.org/10.2307/2531595

Efron B, Tibshirani RJ (1993) An Introduction to the Bootstrap. Chapman & Hall, New York

Elkan C (2001) The foundations of cost-sensitive learning. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), pp 973–978

Federal Trade Commission (2026) FTC testifies before Joint Economic Committee on agency's efforts to combat fraud. Press release, URL https://www.ftc.gov/news-events/news/press-releases/2026/03/ftc-testifies-joint-economic-committee-agencys-efforts-combat-fraud

Flesch R (1948) A new readability yardstick. Journal of Applied Psychology 32(3):221–233. https://doi.org/10.1037/h0057532

Gilardi F, Alizadeh M, Kubli M (2023) ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences 120(30):e2305016120. https://doi.org/10.1073/pnas.2305016120

Gopal RD, Hojati A, Patterson RA (2022) Analysis of third-party request structures to detect fraudulent websites. Decision Support Systems 154:113698. https://doi.org/10.1016/j.dss.2021.113698

Grinsztajn L, Oyallon E, Varoquaux G (2022) Why do tree-based models still outperform deep learning on typical tabular data? In: Advances in Neural Information Processing Systems, pp 507–520

Gunathilaka M, Wickramanayake S, Bandara HMND (2024) DeFiTrust: A transformer-based framework for scam DeFi token detection using event logs and sentiment analysis. Expert Systems with Applications 251:123913. https://doi.org/10.1016/j.eswa.2024.123913

Heiding F, Mayrink Verdun C, Lermen S, et al (2027) Evaluating AI models' capability to automate voice phishing attacks. Expert Systems with Applications 332:133620. https://doi.org/10.1016/j.eswa.2026.133620

Hooker S, Erhan D, Kindermans PJ, et al (2019) A benchmark for interpretability methods in deep neural networks. In: Advances in Neural Information Processing Systems

Ke G, Meng Q, Finley T, et al (2017) LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems

Kendall MG (1938) A new measure of rank correlation. Biometrika 30(1/2):81–93. https://doi.org/10.1093/biomet/30.1-2.81

Liu S, Yang J, Yu X, et al (2026) Ethereum phishing scam detection via multi-view contrastive learning on higher-order networks. Expert Systems with Applications 321:132189. https://doi.org/10.1016/j.eswa.2026.132189

Liu Y, Ott M, Goyal N, et al (2019) RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692, URL https://arxiv.org/abs/1907.11692

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems

Lundberg SM, Erion G, Chen H, et al (2020) From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence 2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9

National Anti-Scam Centre (2026) Targeting scams report: Continued action critical to combat fraud as annual scam losses exceed $2 billion. Australian Competition and Consumer Commission, URL https://www.accc.gov.au/media-release/continued-action-critical-to-combat-fraud-as-annual-scam-losses-exceed-2-billion

Pangakis N, Wolken S, Fasching N (2023) Automated annotation with generative AI requires validation. arXiv preprint arXiv:2306.00176, URL https://arxiv.org/abs/2306.00176

Payment Systems Regulator (2023) Fighting authorised push payment scams: Final decision (PS23/4). Payment Systems Regulator, London, URL https://www.psr.org.uk/media/kwlgyzti/ps23-4-app-scams-policy-statement-dec-2023.pdf

Reimers N, Gurevych I (2019) Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp 3982–3992, https://doi.org/10.18653/v1/D19-1410

Reiss MV (2023) Testing the reliability of ChatGPT for text annotation and classification: A cautionary remark. arXiv preprint arXiv:2304.11085, URL https://arxiv.org/abs/2304.11085

Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432

Sanh V, Debut L, Chaumond J, et al (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, URL https://arxiv.org/abs/1910.01108

Shin M, Kim J, Shin J (2026) The adoption and efficacy of large language models in US consumer financial complaints. Nature Human Behaviour 10(4):669–680. https://doi.org/10.1038/s41562-026-02409-4

Siering M (2022) Explainability and fairness of RegTech for regulatory enforcement: Automated monitoring of consumer complaints. Decision Support Systems 158:113782. https://doi.org/10.1016/j.dss.2022.113782

Tang M, Ye M, Chen W, et al (2024) BiLSTM4DPS: An attention-based BiLSTM approach for detecting phishing scams in ethereum. Expert Systems with Applications 256:124941. https://doi.org/10.1016/j.eswa.2024.124941

UK Finance (2026) Annual fraud report 2026. UK Finance, London, URL https://www.ukfinance.org.uk/system/files/2026-06/UK%20Finance%20Fraud%20Report%202026.pdf

Vickers AJ, Elkin EB (2006) Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making 26(6):565–574. https://doi.org/10.1177/0272989X06295361

Wang W, Wei F, Dong L, et al (2020) MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In: Advances in Neural Information Processing Systems, pp 5776–5788

Wang Z, Zhu S, Chen X (2026) From complaint narratives to monetary relief: A hybrid machine learning framework for CFPB consumer complaints. arXiv preprint arXiv:2606.22664, URL https://arxiv.org/abs/2606.22664

Wolpert DH (1992) Stacked generalization. Neural Networks 5(2):241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Ye M, Han D, Tang M, et al (2026) HyperSeq-LLM: Evolving hypergraph sequence learning with LLMs for phishing scam detection on Ethereum. Expert Systems with Applications 308:131101. https://doi.org/10.1016/j.eswa.2026.131101

Ziems C, Held W, Shaikh O, et al (2024) Can large language models transform computational social science? Computational Linguistics 50(1):237–291. https://doi.org/10.1162/coli_a_00502