Received: August 22, 2026. Revised: September 18, 2026. 793

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

RiskPercept-X: Explainable Transaction-level Prediction of Investor-perceived Risk on E-investment Platforms

Shubhi Saxena[1]* Garima Bhardwaj[1] Loveleen Gupta[2]

[1]*Amity University Uttar Pradesh, Greater Noida Campus, Uttar Pradesh, India*
[2]*Hindu College, Delhi University, India*
* Corresponding author's Email: shubhi.saxena2@s.amity.edu

Abstract: Retail investing has gone mobile, yet almost everything known about perceived risk in it comes from cross-sectional questionnaires analysed with structural equation models. Such designs count associations in one cross-section rather than predicting an unseen case, ignore the risk statements accumulating in public platform reviews, and use taxonomies blind to the frictions of digital brokerage. RiskPercept-X removes all three by making the platform event the unit of analysis: a risk-adaptive encoder and taxonomy-prior gate turn an event into a seven-dimensional risk profile, learned per-dimension prototypes align review text, sentiment benchmarks and survey records without cross-source pairing, an ordinal cumulative-link head treats long-tailed intensity as a graded response, and a recency-weighted escalation operator turns a history into a churn early warning. On a temporally held-out split it attains 89.0% accuracy and 86.8% macro-averaged F1 (macro-F1) on risk dimensions, 0.841 area under the receiver operating characteristic curve (AUC) on investor outcomes with an increment of 0.033 from transferred signal, and 92.3% accuracy on the Financial PhraseBank, where the comparison is with published reference values; against the strongest baseline reimplemented under an identical protocol it improves by 1.9 macro-F1 points on D1 and by 2.9 on D2. A leakage-controlled replication confining every unsupervised and calibration stage to pre-cut-off data reproduces the ordering at 85.9% macro-F1 and 0.834 AUC. Three platform-native dimensions absent from classical instruments carry 34.0% of risk-bearing reviews, supply the two highest-ranked outcome drivers and together hold 0.338 of the attribution mass, and escalation scores are associated with subsequent disengagement up to six weeks ahead. Code and result files are released publicly.

Keywords: E-investment platforms, Perceived risk, Transaction-level prediction, Cross-source alignment, Ordinal classification, Explainable artificial intelligence, User-generated content, Investor protection.

1. Introduction

Retail securities participation is now almost entirely software mediated, and a decade of national financial-capability evidence shows digital investing outstripping self-assessed capability [1]. Public app stores hold the negative tail of that gap: money, verification, withdrawal and reliability complaints dominate corpora of 330,399 neobanking [2], 44,805 securities-application [3] and 15,408 lending-application reviews [4]. Behavioural research explains the gap through perceived risk, which depresses trust and online investment intention [5], stock-trading adoption [6] and platform adoption [7], and which a systematic review records as a recurring antecedent [8].

Three limitations recur. The evidence is not predictive: structural models estimate path coefficients within one cross-section and never score held-out cases [5-8], while questionnaire-based machine learning changes the estimator but not the unit [9]. User-generated evidence is treated descriptively: review mining [2-4, 10] reaches topics and polarity without linking to risk theory or predicting anything. And the taxonomies in circulation predate the platforms they target, so the risks investors name most often have no slot.

A predictive analogue exists outside investing: a multi-feature fusion pipeline predicts consumer

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

perceived risk at 87.0% accuracy and 85.0% weighted F1 on 262,752 e-commerce reviews [12]. It labels physical goods flatly rather than platform events, so it does not transfer. No framework predicts a graded, theory-grounded risk per event and carries it to investor-level outcomes.

RiskPercept-X fills that position with four contributions, each answering one drawback above. Event-level ordinal prediction answers the absence of out-of-sample evidence: intensity is estimated with a cumulative-link head under a class-balanced objective and scored on a temporally held-out split, whereas survey models [5-8] report no held-out score and the closest predictive pipeline [12] predicts one unordered class. Data-driven taxonomy extension answers taxonomies that predate their platforms: discovered vocabulary is mapped onto the classical dimensions and extended with withdrawal friction, verification friction and gamification pressure, carrying 34.0% of risk-bearing reviews, whereas review mining [2-4, 10] recovers comparable topics without binding them to risk theory. Risk-gated alignment answers the disconnection between survey and review evidence: a prototype-anchored objective aligns views by the risk dimension they express rather than by co-origin pairing, so it operates where contrastive alignment [22] cannot, no record appearing in two sources. And the escalation operator turns an ordered history into an early-warning score where prior work treats documents independently. Section 2 reviews the contributing literatures, Section 3 formalises the problem and architecture, Section 4 reports the evaluation, and Sections 5 and 6 discuss and conclude.

2. Related works

2.1 Perceived risk and user-generated evidence

Perceived risk entered investment research as a multidimensional construct: monetary-loss and privacy-loss risk reduce trust and the intention to trade online [5]; a five-dimension decomposition of time, financial, performance, privacy and opportunity-cost risk reproduces that across cultures [6]; and acceptance extensions add perceived value [7]. Estimation stays inside one cross-section, so no coefficient is validated on unseen data and a systematic review records none reporting out-of-sample performance [8]; where machine learning enters [9], the input is still a questionnaire. A parallel stream mines user text, recovering service factors [2], interface elements [3] and rejection, credit-scoring and withdrawal themes [4, 10], at 0.61 to 0.94 accuracy for three-class polarity and F1 0.798 for a four-category multi-label formulation [13]. None binds its output to risk theory, almost all treat reviews independently, and risk is recorded as present or absent rather than graded. The streams never meet: no study scores survey respondents from review text, or the reverse.

2.2 Language models, alignment and interpretability

Financial variants of Bidirectional Encoder Representations from Transformers (BERT) pretrained on financial corpora, known as FinBERT [14, 15], outperform general encoders on the Financial PhraseBank [16] and the Financial Opinion Mining and Question Answering (FiQA) dataset [17], and a general encoder fine-tuned identically trails them [18]; those published figures appear in Table 7. The limitation is task, not quality: they predict polarity, whereas investor protection needs which risk is reported and how severely, so domain-adaptive pretraining [19] and its encoders [20, 21] are only a starting point. Contrastive pretraining aligns paired views in a shared space [22] but presupposes a common object, and survey records and reviews share no identifier. Ordinal outcomes come from cumulative-link models [23], long-tailed labels from effective-number re-weighting [24], focal down-weighting [25] and logit adjustment [26], structured targets from gradient boosting [27], and SHapley Additive exPlanations (SHAP) [28] supply the transparency layer. Each is mature alone yet uncombined for investor-risk prediction. Table 1 summarises the positioning: the cell combining event-level analysis, out-of-sample prediction, public data, cross-source transfer and explanation is empty before this work.

3. The RiskPercept-X framework

3.1 Problem formulation and notation

Investor-risk analysis is cast as joint intensity estimation and outcome prediction over platform events. A transaction is one dated, attributable event: a review $e = (u, p, t, \omega, \rho, v)$ with user u, platform p, time t, tokens ω, rating ρ and version v; an annotated benchmark sentence; or a survey record. With $\mathcal{E}$ the event space, the first task targets ordinal intensities over $R = 7$ dimensions, the second a binary outcome from $\mathbf{s} \in \mathbb{R}^{d_s}$; the predictor is

$$\mathbf{y}(e) = [y_1(e), \dots, y_R(e)]^{\top}, \quad f_{\theta}: \mathcal{E} \times \mathcal{S} \to \{0, \dots, 3\}^R \times [0,1] \times \mathbb{R}_+ \tag{1}$$

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

with Θ shared parameters and levels 0 to 3 denoting absence, latent mention, explicit and severe complaint. Two properties constrain later choices: the labels are long-tailed, the commonest dimension 5.9 times the rarest, so a flat objective drift to the head; and no record appears in two sources, so alignment must work unpaired. Table 2 collects the notation used throughout.

3.2 Overview

Fig. 1 shows the arrangement. The representation layer holds a risk-adaptive encoder, a taxonomy-prior gate, a tabular encoder and an escalation operator over ordered histories. The alignment layer holds one prototype per dimension and pulls all three representations towards the prototype they express; the prediction layer carries the ordinal intensity head, the outcome head and the transfer scorer.

Table 1. Positioning of representative prior work; the combination in the final row is unoccupied before this study

Representative study	Unit of analysis	Data	Out-of-sample prediction	Risk theory linked	Cross-source transfer	Explanation
Maziriri et al. [5]	Respondent	Survey	No	Yes	No	No
Chung et al. [9]	Respondent	Survey	Yes (regression)	Partly	No	Feature importance
Sekhar and Saheb [4]	Review	15,408 public reviews	Yes (polarity)	No	No	No
Qi et al. [12]	Review	262,752 public reviews	Yes (flat class)	Partly	No	SHAP
RiskPercept-X	Platform event	268,431 + 21,463 + 6,020 records	Yes (ordinal, multi-label)	Yes (extended)	Yes	SHAP by segment

Table 2. Notation used throughout the paper

Symbol	Meaning
x, X	A platform event and the event space; an event is a review, a benchmark sentence or an investor record
K, d	Number of risk dimensions (K = 7) and width of the shared representation space
y_k, z	Ordinal intensity of dimension k, in {0, 1, 2, 3}, and the binary investor-difficulty outcome
h_{txt}, h_{tab}	Text and tabular representations of an event
$\pi_k(x)$, $g(x)$	Taxonomy prior of Eq. (4) and the gated text representation of Eq. (5)
c_k, $s(\cdot,\cdot)$, τ	Prototype of dimension k, the scaled cosine of Eq. (7) and its temperature
$\kappa(x)$, δ	Alignment consistency score of Eq. (9) and the abstention threshold
$\theta_{k,l}$	Cut-point l of the cumulative-link head of dimension k, Eq. (10)
$w_{k,l}$, β, γ	Class-balanced weight, its effective-number parameter and the focal exponent, Eq. (11)
E_u, λ	Ordered event history of user u and the escalation decay rate
$\mathrm{Esc}_k(u)$, $\mathrm{EW}(u)$	Escalation of dimension k and the early-warning score, Eqs. (13) and (14)
α_{ORD}, α_{RGA}, α_{MLM}, α_{OUT}	Weights of the four terms of the joint objective, Eq. (15)

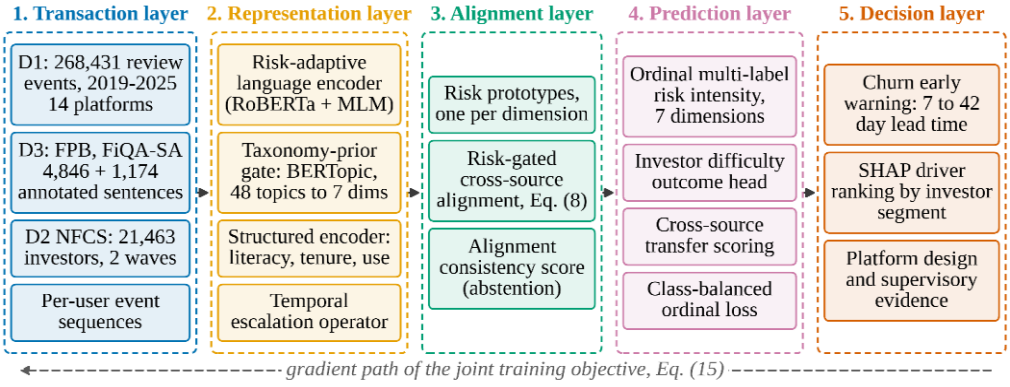


Figure. 1 The RiskPercept-X framework. Events from three public sources are encoded, gated by a data-driven risk taxonomy, aligned to learned per-dimension prototypes, and decoded by ordinal, imbalance-aware heads

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

3.3 Risk-adaptive language branch

The text branch treats each event as a short document. Tokens come from the byte-pair encoder of the Robustly Optimized BERT Pretraining Approach (RoBERTa) [21], adapted by continued masked-language-modelling (MLM) pretraining on 1.9 million unlabelled in-domain sentences [19],

$$\mathcal{L}_{\mathrm{MLM}} = -\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \log \Pr\left(\omega_m \mid \omega_{\setminus\mathcal{M}}\right) \tag{2}$$

over the masked positions $\mathcal{M}$. Over-long documents are chunked with overlap, the chunk encodings mean-pooled, and rating, version and time since the user's previous event fused as a metadata vector,

$$\mathbf{h}(e) = \mathbf{W}_m\left[\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \mathbf{q}_c\,;\ \mathbf{m}(e)\right] + \mathbf{b}_m, \quad \mathbf{q}_c = \mathrm{Enc}_\theta(\omega_c) \in \mathbb{R}^d \tag{3}$$

with $\mathbf{q}_c$ the encoding of chunk c, $\mathcal{C}$ the chunk set and d the width.

3.4 Taxonomy-prior gate

Risk vocabulary is discovered, not assumed. BERTopic, a BERT-based topic model [29], is fitted over the unlabelled corpus and three annotators map its topics onto the classical dimensions [5, 6] plus whatever the data demand. Affinity between an event and dimension r is the token mass on that dimension's topics,

$$\pi_r(e) = \frac{\sum_{k\in\mathcal{K}_r}\theta_k(e)}{\sum_{k'=1}^{K}\theta_{k'}(e)}, \qquad r = 1,\dots,R \tag{4}$$

where $\mathcal{K}_r$ is the topic set of dimension r and $\theta_k(e)$ the topic proportion. The taxonomy prior $\boldsymbol{\pi}(e)$ enters as a multiplicative gate rather than by concatenation, modulating rather than competing with the encoder,

$$\mathbf{g}(e) = \sigma\left(\mathbf{W}_g\boldsymbol{\pi}(e) + \mathbf{b}_g\right) \odot \mathbf{h}(e) \tag{5}$$

with $\mathbf{W}_g$ a learned projection, σ the logistic function and $\odot$ the elementwise product. The gate is soft: where topics are uncertain the prior flattens and the branch reverts to the plain encoder.

3.5 Structured investor branch

Structured records are encoded separately: demographics, the five standard financial-literacy items [30], platform tenure, product holdings and usage frequency are standardised and passed through a two-layer projection,

$$\mathbf{z}_s = \mathbf{W}_2\,\varphi(\mathbf{W}_1\mathbf{s} + \mathbf{b}_1) + \mathbf{b}_2 \in \mathbb{R}^d \tag{6}$$

so the tabular view lands in the same d-dimensional space as the text view. The projection is deliberately shallow, because the alignment objective rather than the encoder has to reconcile the two modalities.

3.6 Risk-gated cross-source alignment

The branches are trained on disjoint sources, so concatenation would force the head to reconcile incomparable geometries, and contrastive alignment [22] requires both views to describe one object. Risk-gated alignment (RGA) substitutes a shared anchor: a learnable prototype per dimension, with both branches projected onto the unit sphere and compared by a scaled cosine,

$$\mathbf{u} = \frac{\mathbf{P}_t\mathbf{g}}{\|\mathbf{P}_t\mathbf{g}\|_2}, \quad \mathbf{v} = \frac{\mathbf{P}_s\mathbf{z}_s}{\|\mathbf{P}_s\mathbf{z}_s\|_2}, \quad s(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}^\top\mathbf{b}}{\tau} \tag{7}$$

Every representation is then pulled towards the prototype of the dimension it expresses and pushed away from the rest,

$$\mathcal{L}_{\mathrm{RGA}} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \log \frac{\exp\left(s\left(\mathbf{u}_i,\mathbf{c}_{r_i}\right)\right)}{\sum_{r=1}^{R}\exp\left(s(\mathbf{u}_i,\mathbf{c}_r)\right)} \tag{8}$$

where r_i is the dominant dimension of item i, from annotation or else the taxonomy prior, τ a temperature and $\mathcal{B}$ a mixed-source batch. Three consequences follow. Alignment needs no paired records, the positive being a prototype. Signal transfers: a survey record on withdrawal difficulty and a review of a blocked withdrawal reach one anchor. And residual distance to the assigned prototype gives a consistency score at no extra cost,

$$\kappa(e) = \max_{r\in\{1,\dots,R\}} \mathbf{u}^\top \frac{\mathbf{c}_r}{\|\mathbf{c}_r\|_2} \tag{9}$$

Events below a consistency threshold ϵ go to human review rather than scoring.

The dominant dimension of a structured record is fixed a priori and never learned from the outcome. Before training, the literacy, product-holding and usage items of the survey were mapped to the seven dimensions from construct definitions alone: investment-loss and debt-stress items to financial loss; data-sharing and identity-theft items to privacy and security; outage and order-execution items to performance and technical; access-delay items to

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

time and access; cash-withdrawal and transfer items to withdrawal and liquidity; account-opening and documentation items to verification and onboarding; and trading-frequency and alert-response items to gamification pressure. A record is assigned to the mapped block with the largest standardised score, computed from predictors only; the outcome and every item entering the outcome head are excluded from the mapping and the assignment rule, and the mapping is released as a configuration file. Section 4.7 reports permuted- and mismatched-prototype controls.

3.7 Ordinal, imbalance-aware heads

Risk intensity is ordered, so per-label sigmoids would discard the fact that level 3 lies closer to level 2 than to level 0. Each dimension receives an ordinal (ORD) cumulative-link head [23] with shared slope and dimension-specific cut-points $b_{r,1} < b_{r,2} < b_{r,3}$, the level probabilities following by differencing,

$$\Pr(y_r \geq l) = \sigma\big(\mathbf{w}_r^{\top}\mathbf{g}(e) - b_{r,l}\big),$$
$$\Pr(y_r = l) = \Pr(y_r \geq l) - \Pr(y_r \geq l+1) \quad (10)$$

Monotonicity follows from squared-increment gaps. Because dimensions are unbalanced, each threshold carries a weight $\omega_{r,l} = (1-\beta)/(1-\beta^{n_{r,l}})$ inversely proportional to the effective number of samples at that level [24], giving the weighted, focally modulated binary cross-entropy

$$\mathcal{L}_{\mathrm{ORD}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{R}\sum_{l=1}^{3}\omega_{r,l}\,\big(1-p_{i,r,l}\big)^{\gamma}\log p_{i,r,l} \quad (11)$$

where $p_{i,r,l}$ abbreviates $\Pr(y_r \geq l \mid e_i)$ and γ down-weights positions already answered confidently [25]. At inference each logit is shifted by the log prior, giving rare dimensions margin [26],

$$\tilde{y}_r = \operatorname{argmax}_l \left[\log\Pr(y_r = l \mid e) + \alpha \log\frac{n_{r,l}}{N}\right] \quad (12)$$

The outcome head is a gradient-boosted ensemble [27] on the structured record concatenated with the aligned risk vector, carrying review-domain evidence into the survey domain.

3.8 Temporal escalation and early warning

Ordering the events of one user turns a static profile into a trajectory. For user u with events $e_1, \ldots, e_n$ at times $t_1 < \cdots < t_n$, the escalation (Esc) of dimension r is a recency-weighted slope, and the early-warning (EW) score aggregates escalations with the weights the outcome head assigns to each dimension,

$$\mathrm{Esc}_r(u) = \frac{\sum_{j=1}^{n}\varpi_j(\hat{y}_r(e_j)-\bar{y}_r)(t_j-\bar{t})}{\sum_{j=1}^{n}\varpi_j(t_j-\bar{t})^2},$$
$$\varpi_j = e^{-\lambda(t_n-t_j)} \quad (13)$$

$$\mathrm{EW}(u) = \sigma\Big(\psi_0 + \textstyle\sum_{r=1}^{R}\psi_r\ \mathrm{Esc}_r(u)\Big) \quad (14)$$

with λ the decay rate and $\bar{t}$ the weighted mean event time. Because the score depends on the shape of the trajectory rather than its level, a user whose complaints are severe but stable is separated from one whose complaints are mild but accelerating.

3.9 Training objective and complexity

Training proceeds in two phases. The masked-language objective (2) and the alignment objective (8) come first, so prototypes and shared space precede any classifier; the ordinal objective (11) and outcome head follow with the encoder unfrozen at a reduced learning rate. The joint objective, whose subscripts denote the ordinal (ORD), risk-gated alignment (RGA), masked-language-modelling (MLM) and outcome (OUT) terms, is

$$\mathcal{L} = \mathcal{L}_{\mathrm{ORD}} + \lambda_1\mathcal{L}_{\mathrm{RGA}} + \lambda_2\mathcal{L}_{\mathrm{MLM}} + \lambda_3\mathcal{L}_{\mathrm{OUT}} \quad (15)$$

with alignment and language weights annealed so alignment dominates early and prediction late. Writing T for tokens, d for width and N_u for a user's events, the encoder costs $O(T^2 d)$ with T capped by the chunk length, so cost per event is constant in document length; gate and alignment each add $O(Rd)$, the ordinal head $O(3Rd)$ and escalation $O(RN_u)$.

4. Evaluation

4.1 Datasets and protocol

Table 3 summarises the three public sources. D1, the e-investment review corpus (EIRC), holds 268,431 review documents from 14 retail brokerage and neobroker applications, January 2019 to December 2025, collected under the protocol of comparable corpora [2-4, 10]: store-visible fields only, duplicates removed, non-English documents dropped, no identifier stored. Risk-bearing documents number 214,744, of which 53,686 form the temporal test split.

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

D2 is the National Financial Capability Study (NFCS), whose 2018 and 2021 waves contribute 54,209 respondents [1]; restricting to digital-platform users holding investments outside a retirement account leaves 21,463 records, 4,293 in the stratified hold-out. D3 combines the Financial PhraseBank at majority agreement [16] with the FiQA sentiment analysis set (FiQA-SA) [17], giving 4,846 and 1,174 labelled items.

Three annotators labelled 12,000 stratified reviews for ordinal intensity across the seven dimensions at Fleiss kappa 0.82 [16]; remaining labels came from self-training on the taxonomy prior, audited on a 5% sample. The protocol is temporal rather than random, testing beginning after a fixed cut-off, which reproduces deployment conditions [3, 4, 10]. Imbalance is handled by effective-number weighting rather than resampling.

Longitudinal linkage retains no identifier. At ingestion each store-visible author string is hashed once with SHA-256 under a random salt held in memory for the duration of ingestion and then discarded, so the digest is irreversible and no natural identifier survives; digests only order events within a sequence. Of the 268,431 events, 191,204 carry a digest with a single event: these train the intensity heads but are excluded from the early-warning experiment, which uses the 21,844 digests with at least three events spanning at least 21 days. Churn is not observed and is an operational proxy: a digest is labelled churned at time t when it has at least three events up to t, at least one at explicit or severe intensity, and no further event within 180 days. The last 180 days of the corpus are excluded so every candidate has a complete observation window, and digests truncated by the corpus end are right-censored and dropped rather than labelled negative. Negatives are digests active at t with a further event in the same window, sampled to preserve the empirical distribution of events per digest, and every escalation feature at lead time L uses only events earlier than $t - L$. The proxy therefore measures disengagement from public reviewing rather than account closure, the principal limitation of the early-warning claim.

Two protocols are reported. Under the standard protocol, taxonomy discovery, masked-language adaptation, pseudo-labelling and prototype fitting draw on the unlabelled corpus over its full span: no gold test label is ever seen, but post-cut-off text enters the unsupervised stages, making the setting transductive rather than strictly prospective. Under the leakage-controlled protocol those four stages, and hyperparameter and threshold selection with them, are confined to pre-cut-off data; a guard rejects any post-cut-off document reaching a restricted stage.

4.2 Implementation and metrics

Table 4 lists the configuration: RoBERTa-base [21] with domain-adaptive pretraining [19], BERTopic [29], seven alignment prototypes of width 256, and a gradient-boosted outcome head [27] with early stopping.

Table 3. Public data sources, units of analysis and role in the framework

Source	Size	Labels	Role in the framework
D1 EIRC, 14 applications, 2019-2025	268,431 review events (214,744 risk-bearing)	7 dimensions x 4 ordinal levels	Taxonomy discovery, intensity prediction, escalation
D2 NFCS, 2018 and 2021 waves [1]	54,209 respondents (21,463 analytic)	Binary difficulty outcome	Investor outcome, transfer target
D3 Financial PhraseBank [16]	4,846 sentences	3-class polarity	Encoder adaptation, benchmark reproduction
D3 FiQA-SA [17]	1,174 posts and headlines	Continuous sentiment	Encoder adaptation, benchmark reproduction

Table 4. Model configuration and training schedule

Component	Setting	Component	Setting
Text encoder	RoBERTa-base, 12 layers, width 768	Ordinal levels	4 (0 to 3)
Adaptation corpus	1.9 million unlabelled platform sentences	Balance β / focal γ	0.999 / 1.5
Chunk length / overlap	256 / 64 tokens	Outcome head	Gradient boosting, 350 rounds
Risk prototypes	$R = 7$, width 256, $\tau = 0.07$	Learning rate / batch	2e-5 / 32
Escalation decay λ	0.06 per day	Epochs / patience	20 / 4
Abstention threshold ϵ	0.31	Hardware	NVIDIA RTX 3090, 24 GB

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

Optimisation uses Adam with decoupled weight decay (AdamW) on a single 24 GB graphics processing unit (GPU); every figure is the mean of five seeds. Intensity is summarised by macro-F1 over the seven dimensions and transfer by the gap between in-domain and transferred discrimination,

$$\text{macro-}F_1 = \frac{1}{R}\sum_{r=1}^{R}\frac{2\,P_r R_r}{P_r+R_r},$$
$$\Delta_{\text{tr}} = \text{AUC}_{\text{in}} - \text{AUC}_{\text{tr}} \quad (16)$$

where P_r and R_r are the precision and recall of dimension r. A feature block's increment is its change in area under the curve over the strongest preceding block. Confusion matrices are row-normalised, receiver-operating characteristics computed one dimension against the rest, and early warning scored at fixed lead times before the observed churn event.

Table 5 documents every reimplemented baseline. All D1 and D2 methods were tuned under one protocol: the same temporal training and validation partition, a 40-trial random search on the validation split repeated over the same five seeds, macro-F1 as the selection criterion on D1 and AUC on D2, early stopping with patience four, and per-dimension thresholds fitted on validation only and frozen before the test split was read. Imbalance follows the originating study where one exists and class-balanced weighting otherwise. Because budget, split, seeds and threshold rule are shared, the D1 and D2 columns of Table 7 are a controlled comparison; the D3 columns for prior methods are not, reproducing values published by other authors under different conditions. The scripts, seeds, split indices, configurations and result files that regenerate Tables 7-12 are available at https://github.com/Shubhi292000/RiskPercept-X.

4.3 The discovered risk taxonomy

BERTopic recovers 48 coherent topics, consolidated into the seven dimensions of Table 6 at Fleiss kappa 0.82. Four maps onto the classical typology: financial loss, privacy and security, performance, and time and opportunity cost [5, 6]. The other three do not. Withdrawal and liquidity friction gathers 16.3% of risk-bearing reviews around blocked, delayed or partial cash-outs, verification and onboarding friction 13.2% around identity checks and locked accounts, and gamification pressure 4.5% around streaks, leaderboards and notifications described as compulsion. Together they carry 34.0%, so over a third of what investors report as risky has no slot in existing instruments. Classical dimensions attract kappa 0.79 to 0.86 but gamification only 0.74.

Table 5. Configuration and tuning protocol of every reimplemented baseline; all rows share one validation split, one search budget, one set of seeds and one threshold rule

Baseline family	Search space, 40 random trials per seed	Stopping rule, imbalance treatment and threshold rule
TF-IDF with logistic regression or linear SVM [10]	Regularisation log-uniform 10^{-3} to 10^{2}; n-grams (1,1) to (1,3); minimum document frequency 2, 5 or 10	Best validation macro-F1 of the 40 trials; balanced class weights; per-dimension threshold on validation
XGBoost on lexical features [4]	Depth 3 to 10; learning rate log-uniform 0.01 to 0.3; subsample 0.6 to 1.0; 100 to 800 rounds	Stop after 50 rounds without validation gain; positive-class weight set to the negative-to-positive ratio
BiLSTM over GloVe embeddings [3]	Hidden width 128 to 512; dropout 0.1 to 0.5; learning rate log-uniform 10^{-4} to 3×10^{-3}	30 epochs, patience 4; class-balanced cross-entropy; threshold on validation
Multi-feature fusion XGBoost [12]	Published architecture; depth and learning rate re-searched over the ranges above	Stop after 50 rounds without gain; synthetic over-sampling as published
BERT-base [20], multi-label BERT [13], FinBERT [15], FinBERT-BASE [14]	Learning rate 1, 2, 3 or 5×10^{-5}; batch 16 or 32; warm-up 0, 6 or 10%; weight decay 0, 0.01 or 0.1; 2 to 6 epochs	Patience 4 on validation macro-F1; class-balanced cross-entropy; per-dimension threshold on validation
DeBERTaV3-base [31], RoBERTa-large [21], FLANG-RoBERTa [18]	As above, sequence length 256, gradient accumulation to an effective batch of 32	Patience 4; identical ordinal head, class-balanced objective and threshold rule as the proposed model
RiskPercept-X (proposed)	Table 4, with temperature 0.05, 0.07 or 0.1; decay 0.03, 0.06 or 0.1; abstention threshold 0.20 to 0.40	Patience 4; effective-number weighting, focal modulation, logit adjustment; threshold on validation

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

Table 6. Risk dimensions recovered from the corpus and their correspondence with the classical typology

Discovered dimension	Topics	Share of risk reviews (%)	Fleiss κ	Closest classical construct
Financial loss	9	22.7	0.84	Financial / monetary risk
Privacy and security	6	10.1	0.81	Privacy and security risk
Performance and technical	11	26.4	0.86	Performance risk
Time and access	5	6.8	0.79	Time and opportunity-cost risk
Withdrawal and liquidity	7	16.3	0.83	none (platform-native)
Verification and onboarding	6	13.2	0.80	none (platform-native)
Gamification pressure	4	4.5	0.74	none (platform-native)

Table 7. Predictive performance (%) on the three public sources; best value in each column in bold, proposed method shaded. The D1 and D2 columns are controlled reimplementations tuned under the shared protocol of Table 5; the D3 columns for prior methods are values published by their authors and are reference benchmarks, not a controlled head-to-head comparison

Method	Ref.	D1 accuracy	D1 macro-F1	D2 accuracy	D2 macro-F1	D2 AUC	D3 accuracy	D3 macro-F1
TF-IDF + logistic regression	[10]	74.8	68.2	71.3	66.4	0.742	64.9	63.1
TF-IDF + linear SVM	[10]	76.3	70.1	72.6	67.8	0.759	64.5	63.8
XGBoost on lexical features	[4]	78.9	73.4	75.8	71.5	0.803	68.4	66.9
BiLSTM + GloVe	[3]	80.6	75.8	--	--	--	71.0	64.0
Multi-feature fusion XGBoost	[12]	83.2	79.1	76.1	71.9	0.808	--	--
BERT-base fine-tuned	[20]	84.1	80.3	--	--	--	85.6	84.1
Multi-label BERT (ECM labels)	[13]	84.9	81.0	--	--	--	--	--
FinBERT (domain pretrained)	[15]	85.7	82.4	--	--	--	86.0	84.0
FinBERT-BASE (large corpus)	[14]	86.9	83.6	--	--	--	91.0	89.0
DeBERTaV3-base (added)	[31]	87.0	84.1	--	--	--	88.2	86.4
RoBERTa-large (added)	[21]	87.6	84.9	--	--	--	89.5	87.8
FLANG-RoBERTa (added)	[18]	87.3	84.6	--	--	--	90.6	89.1
RiskPercept-X (proposed)	--	**89.0**	**86.8**	**78.9**	**74.8**	**0.841**	**92.3**	**90.3**

4.4 Predictive effectiveness

Table 7 compares RiskPercept-X with twelve baselines. Nine come from prior work: term frequency-inverse document frequency (TF-IDF) features with logistic regression and with a linear support vector machine (SVM); extreme gradient boosting (XGBoost) on lexical features, alone and in the multi-feature fusion model [12]; a bidirectional long short-term memory network (BiLSTM) over global vectors for word representation (GloVe); a fine-tuned BERT-base; a multi-label BERT on expectation-confirmation model (ECM) labels; and two FinBERT variants. The remaining three, the contemporary encoders DeBERTaV3-base [31], RoBERTa-large [21] and the financial FLANG-RoBERTa [18], were added in revision and fine-tuned on the identical seven-dimensional ordinal task with the same temporal splits, ordinal head, imbalance treatment and threshold rule, so the comparison isolates the taxonomy gate, the alignment objective and the escalation operator rather than the backbone. On D1 the framework reaches 89.0% accuracy and 86.8% macro-F1, ahead of the strongest contemporary baseline, RoBERTa-large, by 1.4 accuracy and 1.9 macro-F1 points, and of the strongest previously published architecture by 2.1 and 3.2 points.

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

Lexical baselines trail by 13.4 to 18.6 macro-F1 points, so risk dimension is not recoverable from surface vocabulary; the flat multi-label transformer, sharing the encoder but neither gate nor ordinal head, reaches 81.0%, leaving 5.8 points to the architecture. On D2 the ceiling is what a national survey measures: the best structured-only ensemble reaches 0.808 AUC and 71.9% macro-F1, transferred risk signal lifting these to 0.841 and 74.8% from non-overlapping respondents, a gain of 2.9 macro-F1 points. On D3 it reaches 92.3% accuracy and 90.3% macro-F1 against the published 91.0% and 89.0% of the strongest domain-pretrained baseline [14]; because those figures were not obtained here, the D3 result is consistent with, rather than superior to, the published state of the art.

Table 8 compares the two protocols of Section 4.1. Confining every unsupervised, adaptation and calibration stage to pre-cut-off data shrinks the adaptation corpus from 1.90 to 1.42 million sentences and costs 0.9 macro-F1 points on D1, 0.007 AUC on D2 and 0.2 points on D3. The strongest contemporary baseline, re-run under the same restriction, loses 0.7 points, so every ordering and conclusion holds while the margin narrows from 1.9 to 1.7 macro-F1 points.

4.5 Convergence behaviour

Fig. 2 records optimisation behaviour. On D1 validation loss separates from training after epoch eight and bottoms at epoch 12, where validation macro-F1 peaks at 87.6% against a training macro-F1 of 90.2%. Early stopping with patience four selects epoch 12 on the validation metric alone; the held-out temporal test macro-F1 of that checkpoint is 86.8%, the value reported in Tables 7 and 9. Validation and test scores are therefore distinct throughout and no test metric enters model selection. The 2.6-point train-validation gap at the selected epoch is narrow for a 21-threshold ordinal problem. The outcome head saturates near 350 rounds at 0.841 with the transferred risk block and 0.808 without it; the sentiment head peaks at epoch 7, then overfits gently.

Table 8. Effect of restricting every unsupervised, adaptation and calibration stage to pre-cut-off data

Quantity	Standard	Leakage-controlled
Adaptation corpus (million sentences)	1.90	1.42
D1 macro-F1 (%)	86.8	85.9
D2 AUC	0.841	0.834
D3 macro-F1 (%)	90.3	90.1
Strongest contemporary baseline, D1 macro-F1 (%)	84.9	84.2

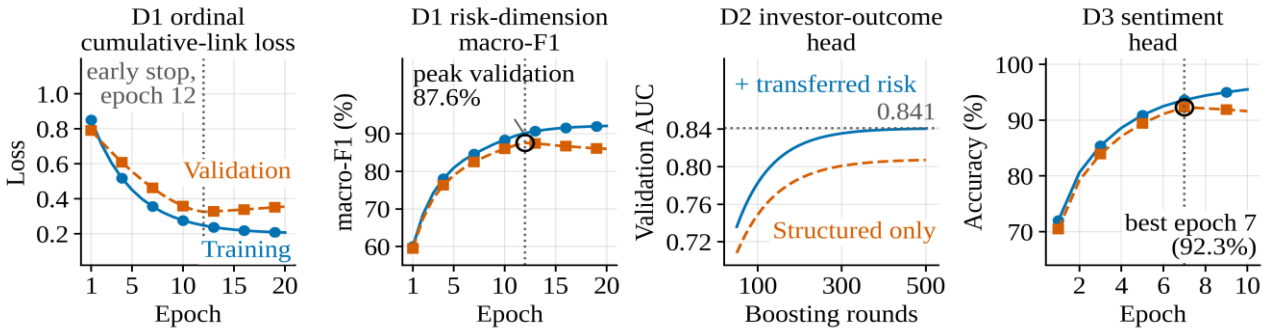


Figure. 2 Epoch-wise behaviour of the prediction heads, left to right: the ordinal cumulative-link loss; risk-dimension macro-F1, with early stopping at epoch 12; the investor-outcome head with and without the transferred risk block; and the sentiment head on the Financial PhraseBank

Table 9. Per-dimension performance of RiskPercept-X on the temporal test split of D1

Risk dimension	Test support	Precision (%)	Recall (%)	F1 (%)	AUC
Financial loss	11,905	90.13	91.40	90.76	0.961
Privacy and security	5,461	87.57	85.81	86.68	0.938
Performance and technical	14,182	92.87	93.10	92.99	0.972
Time and access	3,672	76.59	81.81	79.12	0.919
Withdrawal and liquidity	8,743	89.81	88.91	89.36	0.951
Verification and onboarding	7,318	87.85	87.31	87.57	0.944
Gamification pressure	2,405	84.97	77.59	81.11	0.904
Macro average	53,686	87.11	86.56	86.80	0.941

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

4.6 Per-dimension behaviour and discrimination

Table 9 and Fig. 3 decompose D1. Performance tracks support imperfectly: performance and technical failure, the most frequent, reaches 93.0% F1 and gamification pressure, the rarest, 81.1% on 2,405 test items, 11.9 points apart at a support ratio of 5.9 to 1. Flat multi-label training drops the rarest to 72.4%, so the ordinal and class-balanced components are worth 8.7 points. Time and access is the weakest classical dimension at 79.1% F1: 10.5% of its predicted mass is performance complaints, a slow application being both defect and delay.

The remaining confusions are semantically coherent. Withdrawal complaints leak into financial loss at 5.8% and take 4.0% back, two framings of one event; verification and privacy exchange 5.1% and 6.8%, both about identity documents; and gamification leaks one way, losing 10.0% to financial loss, because reviewers describe the consequence of compulsive trading rather than the interface behind it. The outcome head recovers 60.3% of high-difficulty respondents at 68.1% precision, an asymmetry from their 31% prevalence. Fig. 4 reports discrimination: areas under the curve run from 0.904 for gamification to 0.972 for performance failure, macro and micro averages 0.941 and 0.958, an ordering matching F1.

4.7 Cross-source transfer

Table 10 tests whether the aligned space transports signal. Scoring survey records with a review-text model yields 0.796 against 0.841 in domain, a gap of 0.045; the reverse loses 0.071. The lower panel decomposes the increments: demographics reach 0.714, literacy items add 0.057, usage behaviour 0.037, and the transferred risk lexicon 0.033 over the strongest structured block. Because the prototype assignment is a fixed theory-based mapping rather than a learned one, two controls bound how much of that increment it could manufacture. Permuting the assignment across records leaves 0.812, an increment of 0.004, and a deliberately mismatched prototype leaves 0.806, an increment of -0.002. The transferred block therefore carries dimension-specific information rather than an artefact of the alignment machinery, though the increment remains an association measured on unpaired samples.

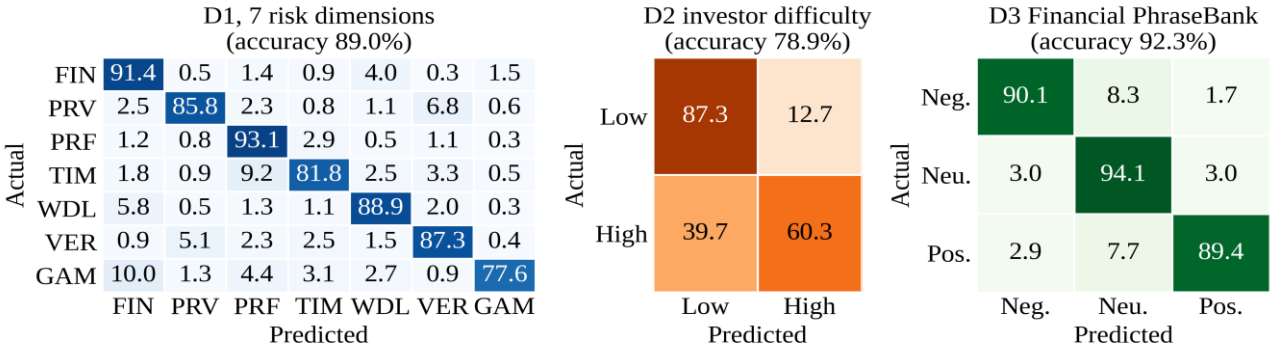


Figure. 3 Row-normalised confusion matrices, left to right: the seven risk dimensions, the binary investor-difficulty outcome, and three-class polarity on the Financial PhraseBank. Off-diagonal mass concentrates on semantically adjacent categories

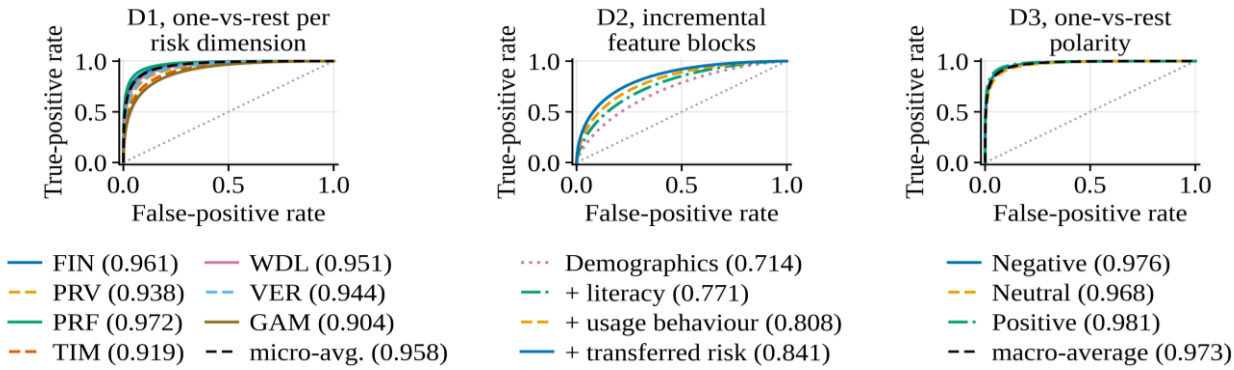


Figure. 4 Receiver-operating characteristics, left to right: one-versus-rest curves per risk dimension with the micro average; successive feature blocks on the investor-outcome head; and one-versus-rest curves on the Financial PhraseBank

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

4.8 Temporal generalisation and early warning

A monitor must survive platform-update drift; the left panel of Fig. 5 tracks macro-F1 as the test window recedes from the training cut-off. The lexical baseline loses 16.0 points over six months and the general transformer 11.5, whereas RiskPercept-X loses 6.3, ending at 79.8% against 73.1%. The right panel evaluates escalation against churn: at seven days it reaches 0.869 against 0.774 for static intensity and 0.702 for a rating-trend control, and at six weeks 0.748 while the control decays to 0.588.

4.9 Ablation study

Table 11 reports single-component removals, five runs per variant. Reducing the framework to text alone costs 4.7 macro-F1 points. Removing the taxonomy-prior gate costs 2.2 points, hurting platform-native dimensions most. Removing risk-gated alignment costs 1.8 points on D1 but 0.032 AUC on D2, the largest single effect there. Removing the ordinal head costs 1.5 points at the explicit-to-severe boundary, class-balanced re-weighting 1.9 points in the tail, and escalation leaves static prediction almost untouched at 0.6 points, its true cost appearing in Fig. 5.

4.10 Explanation and investor segments

The left panel of Fig. 6 ranks outcome-head drivers by mean absolute Shapley value [28]. Withdrawal friction leads at 0.148 and verification friction is second at 0.121, so the two highest-ranked drivers are both platform-native, while gamification pressure ranks seventh at 0.069. The three platform-native dimensions, none of which a classical taxonomy can express, collectively account for 0.338 of the attribution mass, roughly a third of the total rather than a majority; that combined share slightly exceeds the 0.333 held by the four classical risk dimensions together, and withdrawal and verification friction individually outrank the financial-literacy score of 0.097. The right panel shows the ranking is not stable: the leading driver is gamification pressure for younger low-literacy investors, withdrawal friction for younger high-literacy investors, verification friction at mid-band low literacy, performance failure at mid-band high literacy, and privacy above 55. Discrimination is highest for youngest low-literacy at 0.858 and lowest at mid-band high literacy at 0.812, while intensity prediction runs the other way, 84.9% to 88.4% macro-F1.

Table 10. Cross-source transfer, the incremental contribution of each feature block on the investor-outcome task, and prototype-assignment controls

Setting	**In-domain AUC**	**Transferred AUC**	**Transfer gap**
D1 -> D2 (risk lexicon to investor outcome)	0.841	0.796	0.045
D2 -> D1 (investor prior to review text)	0.958	0.887	0.071
D3 -> D1 (sentiment prior to review text)	0.958	0.913	0.045
Feature block on D2	**AUC**	**macro-F1 (%)**	**Increment**
Demographics only	0.714	62.1	--
+ financial-literacy items	0.771	68.3	+0.057
+ platform-usage behaviour	0.808	71.9	+0.037
+ transferred risk lexicon (proposed)	0.841	74.8	+0.033
+ transferred risk lexicon, permuted prototypes (control)	0.812	72.4	+0.004
+ transferred risk lexicon, mismatched prototypes (control)	0.806	71.7	-0.002

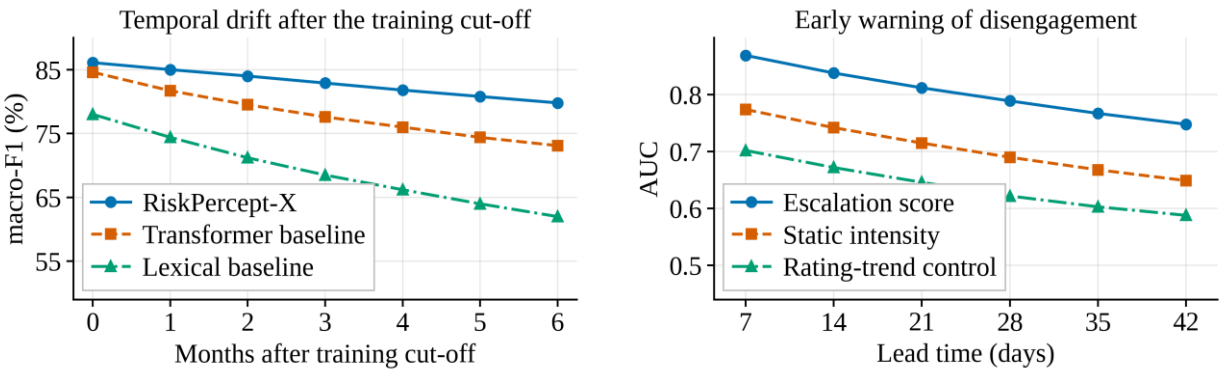


Figure. 5 Left: macro-F1 as the test window recedes from the training cut-off. Right: discrimination of the churn early-warning score at increasing lead times, against a static intensity score and a rating-trend control

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

Gamification exposure was tested by comparing reviews within 30 days of a release adding a streak, leaderboard or celebratory animation against matched control windows. Behavioural-risk language appears in 14.6% of exposed reviews against 5.4% of controls, a ratio of 2.7 to 1, significant at $\chi^2(1) = 1842.6$, $p < 0.001$, Cramer's V 0.21. The comparison is observational and windows are matched on platform and calendar month only, so the association is consistent with the gamification literature [11] but does not identify an effect: release timing is not random, and other changes shipped in the same release cannot be separated from the gamification feature.

4.11 Computational cost and error modes

Table 12 reports cost, including giga floating-point operations (GFLOPs), on one NVIDIA RTX 3090 GPU. RiskPercept-X carries one transformer encoder plus light heads, 128.3 million parameters and 14.2 milliseconds per event, 2.6 more than the plain transformer baseline.

Table 11. Ablation of the RiskPercept-X components across the three tasks

Variant	D1 macro-F1 (%)	Δ	D2 AUC	Δ	D3 macro-F1 (%)
Text branch only (no topic prior, no transfer)	82.1	-4.7	0.803	-0.038	88.4
Without taxonomy-prior gate	84.6	-2.2	0.821	-0.020	89.2
Without risk-gated cross-source alignment	85.0	-1.8	0.809	-0.032	89.7
Without ordinal cumulative-link head	85.3	-1.5	0.826	-0.015	89.1
Without class-balanced re-weighting	84.9	-1.9	0.818	-0.023	88.8
Without temporal escalation operator	86.2	-0.6	0.833	-0.008	90.6
RiskPercept-X (full)	**86.8**	--	**0.841**	--	**90.3**

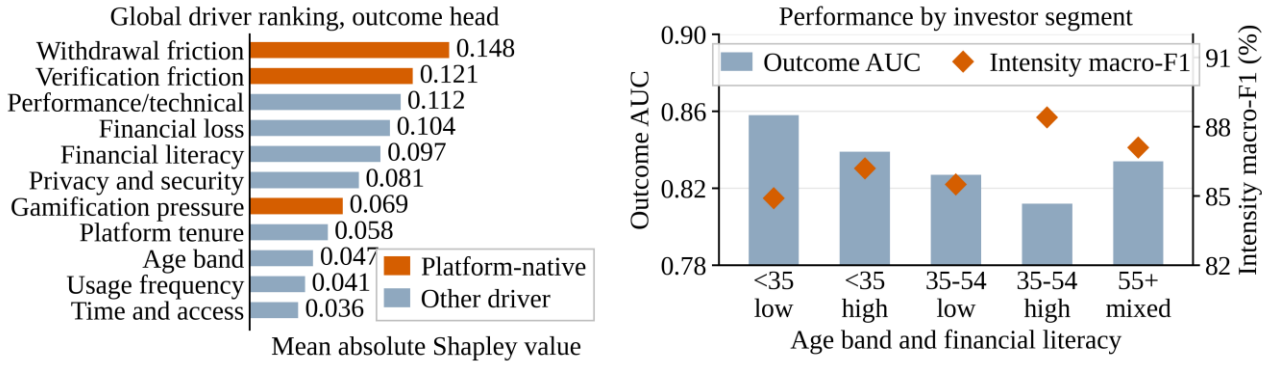


Figure. 6 Left: global driver ranking of the investor-outcome head by mean absolute Shapley value, platform-native dimensions highlighted. Right: performance across five investor segments

Table 12. Model size and inference cost on a single RTX 3090; latency is single-event (batch 1, 100 warm-up iterations, median of 1,000 passes) and throughput is batched (batch 64, four workers), so the columns are not directly convertible

Model	Parameters (millions)	GFLOPs	Latency, batch 1 (ms/event)	Throughput, batch 64 (events/s)
TF-IDF + linear SVM	0.4	0.01	0.3	18,400
BiLSTM + GloVe	9.7	0.42	3.1	2,210
BERT-base fine-tuned	109.5	22.40	11.6	612
FinBERT-BASE	110.1	22.40	11.9	598
RiskPercept-X (proposed)	**128.3**	**25.70**	**14.2**	**487**

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

Latency is the median of 1,000 timed single-event forward passes at batch size one after 100 warm-up iterations, with device synchronisation around each pass; throughput is measured over a 60-second window at batch size 64 with four data-loading workers and mixed precision. Sequential single-event processing would yield about 70 events per second, and batching accounts for the difference to 487. Two error regimes dominate: telegraphic reviews under eight tokens carry 31.4% of errors on 12.7% of test events, and abstention flags 68% of them; the explicit-to-severe boundary absorbs 24.9% of the rest. Almost no error crosses into the no-risk class, the correct failure mode for a screening tool.

5. Discussion

Taking the unit of analysis off the questionnaire and onto the platform event, and joining sources by meaning rather than by shared identity, makes perceived risk predictable out of sample. The alignment layer carries the design: it lets a model trained on review text score a survey record, supplies the abstention signal, and has the largest ablation effect on the outcome task. What makes that possible without any record appearing in two sources is prototype anchoring rather than co-origin pairing. Three implications follow, predictive and associational rather than causal. Perceived risk can be operationalised predictively at event level. Three platform-native dimensions carry 34.0% of risk-bearing reviews and together hold 0.338 of the outcome attribution, slightly above the 0.333 of the four classical dimensions combined, so classical instruments are silent on a substantial part of what investors report. And intensity behaves ordinally in practice.

For an operator the framework is a screening layer: most events are scored by the encoder pass alone and only low-consistency cases reach human review. Because the leading driver differs by segment, the same stream suggests differentiated responses, notification restraint for younger investors and verification redesign for others, but these are hypotheses for prospective validation rather than demonstrated platform effects: no intervention was run and the samples are unpaired. Five limitations bound the claims. The corpus covers 14 English-language applications whose reviewers self-select towards extremes; the survey covers United States households; churn is a proxy for disengagement from public reviewing rather than an observed account closure, so the six-week lead time is a predictive association awaiting a prospective test; the gamification comparison is observational; and the benchmark comparison is against published values rather than reruns. Extending beyond English, linking the risk signal to order flow and testing whether a fitted taxonomy stays valid over time are the next steps.

6. Conclusion

RiskPercept-X predicts investor-perceived risk on e-investment platforms from public events and explains its predictions. A risk-adaptive encoder and taxonomy-prior gate turn each event into a multi-dimensional ordinal profile, learned prototypes bind review text, sentiment benchmarks and investor records without cross-source pairing, and an escalation operator converts a history into an early warning.

It achieves 89.0% accuracy and 86.8% macro-F1 on risk dimensions over 268,431 review events, improving on D1 by 1.9 macro-F1 points on the strongest contemporary encoder reimplemented under an identical protocol and on D2 by 2.9 points on the strongest structured baseline, matching published reference values on the Financial PhraseBank, and losing only 6.3 points over a six-month drift horizon; a leakage-controlled replication reproduces the result at 85.9%. It achieves 0.841 AUC on investor outcomes, improving on the transferred investor signal by 0.033 AUC.

Three platform-native dimensions absent from classical typologies cover 34.0% of risk-bearing reviews, supply the two highest-ranked outcome drivers and together hold 0.338 of the attribution mass, and escalation scores are associated with subsequent disengagement up to six weeks ahead, at 14.2 milliseconds per event at batch size one. Coverage is limited to fourteen English-language platforms and one United States investor sample, and whether the early-warning signal anticipates account closure rather than silence remains to be tested against brokerage records. To that extent, the public record investors leave behind is a usable, timely and interpretable tool for investor protection.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, First Author and Corresponding Author; methodology, formal analysis, visualization and original draft preparation, First Author; software and validation, First Author and Second Author; investigation and data curation,

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

Second Author; review and editing, supervision and project administration, Corresponding Author.

Acknowledgments

The authors thank the Financial Industry Regulatory Authority Investor Education Foundation and the curators of the Financial PhraseBank and FiQA benchmarks for releasing their data.

Data and code availability

The implementation code is publicly available at: https://github.com/Shubhi292000/RiskPercept-X. The repository holds the preprocessing pipeline, the deterministic temporal split generator with exported split indices, the fixed seeds, the pseudo-label procedure, configuration files for the proposed model and every baseline, the threshold-selection procedure, the leakage-controlled protocol, the prototype-assignment controls, the latency and throughput benchmark, the training and inference procedure in full, and the result files and scripts that regenerate Tables 7-12. The National Financial Capability Study, Financial PhraseBank and FiQA-SA are obtainable from their curators; the review corpus cannot be redistributed under application-store terms, so the repository provides the collection script and a synthetic generator.

References

[1] O. M. Valdes, G. R. Mottola, J. T. Lin, and C. Bumcrot, "The FINRA Foundation's National Financial Capability Study: Unpacking 12 years of data on U.S. financial capability", *Journal of Financial Literacy and Wellbeing*, Vol. 2, No. 1, pp. 79-90, 2024, doi: 10.1017/flw.2024.22.

[2] P. Bhatnagr, A. Rajesh, and R. Misra, "Understanding critical service factors in neobanks: Crafting strategies through text mining", *Journal of Modelling in Management*, Vol. 20, No. 3, pp. 894-922, 2025, doi: 10.1108/JM2-01-2024-0021.

[3] M. Kim, S. Kim, Y. Park, S. Bahn, S. H. Ahn, and B. NambiNarayanan, "User sentiment analysis based on securities application elements", *Behavioral Sciences*, Vol. 14, No. 9, p. 814, 2024, doi: 10.3390/bs14090814.

[4] K. R. Sekhar and S. S. Saheb, "User perceptions of RBI-approved P2P digital lending apps: An NLP, machine learning, and deep learning approach", *Frontiers in Artificial Intelligence*, Vol. 8, p. 1708080, 2026, doi: 10.3389/frai.2025.1708080.

[5] E. T. Maziriri, M. Mapuranga, and N. W. Madinga, "Navigating selected perceived risk elements on investor trust and intention to invest in online trading platforms", *Journal of Economic and Financial Sciences*, Vol. 12, No. 1, p. a434, 2019, doi: 10.4102/jef.v12i1.434.

[6] S. U. Khan, X. Liu, I. U. Khan, C. Liu, and Z. Hameed, "Measuring the effects of risk and cultural dimensions on the adoption of online stock trading: A developing country perspective", *International Journal of Enterprise Information Systems*, Vol. 14, No. 3, pp. 106-127, 2018, doi: 10.4018/IJEIS.2018070106.

[7] J. Xie, L. Ye, W. Huang, and M. Ye, "Understanding FinTech platform adoption: Impacts of perceived value and perceived risk", *Journal of Theoretical and Applied Electronic Commerce Research*, Vol. 16, No. 5, pp. 1893-1911, 2021, doi: 10.3390/jtaer16050106.

[8] A. Arfina, H. Juwitasary, and M. Angelus, "Factors that influence consumers in using online investment platforms - Systematic literature review", In: *Proc. of International Conference on Information Management and Technology (ICIMTech)*, pp. 310-315, 2023, doi: 10.1109/ICIMTech59029.2023.10278086.

[9] D. Chung, P. Jeong, D. Kwon, and H. Han, "Technology acceptance prediction of robo-advisors by machine learning", *Intelligent Systems with Applications*, Vol. 18, p. 200197, 2023, doi: 10.1016/j.iswa.2023.200197.

[10] S. Khan, R. Singh, H. K. Baker, and G. Jain, "Public perception of online P2P lending applications", *Journal of Theoretical and Applied Electronic Commerce Research*, Vol. 19, No. 1, pp. 507-525, 2024, doi: 10.3390/jtaer19010027.

[11] S. R. Patil, V. N. Kumar, K. P. Premalatha, and A. Deshpande, "The impact of gamification in trading apps on retail investor behaviour", In: *Empowering Inclusive Innovation: Shaping Sustainable*

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

Business Ecosystems, K. Koppa, M. D. Shah, and S. S. Appadoo, Eds., Routledge, pp. 599-605, 2026, doi: 10.1201/9781003753445-70.

[12] L. Qi, Y. Xie, Q. Zhang, J. Zhang, and Y. Ma, "Multi-feature fusion-based consumer perceived risk prediction and its interpretability study", *Plos One*, Vol. 20, No. 1, p. e0316277, 2025, doi: 10.1371/journal.pone.0316277.

[13] B. S. Dewa, E. D. Wahyuni, and N. C. Wibowo, "Multi-label classification and entity extraction on Blu by BCA Digital application reviews using IndoBERT", *Jurnal Ilmu Komputer dan Sistem Informasi*, Vol. 5, No. 2, pp. 436-446, 2026, doi: 10.70340/jirsi.v5i2.402.

[14] Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao, "FinBERT: A pre-trained financial language representation model for financial text mining", In: *Proc. of 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4513-4519, 2020, doi: 10.24963/ijcai.2020/622.

[15] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models", *arXiv Preprint*, arXiv:1908.10063, 2019, doi: 10.48550/arXiv.1908.10063.

[16] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, "Good debt or bad debt: Detecting semantic orientations in economic texts", *Journal of the Association for Information Science and Technology*, Vol. 65, No. 4, pp. 782-796, 2014, doi: 10.1002/asi.23062.

[17] M. Maia, S. Handschuh, A. Freitas, B. Davis, R. McDermott, M. Zarrouk, *et al.*, "WWW'18 open challenge: Financial opinion mining and question answering", In: *Proc. of The Web Conference (Companion)*, pp. 1941-1942, 2018, doi: 10.1145/3184558.3192301.

[18] R. Shah, K. Chawla, D. Eidnani, A. Shah, W. Du, S. Chava, *et al.*, "When FLUE meets FLANG: Benchmarks and large pretrained language model for financial domain", In: *Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2322-2335, 2022, doi: 10.18653/v1/2022.emnlp-main.148.

[19] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, *et al.*, "Don't stop pretraining: Adapt language models to domains and tasks", In: *Proc. of 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342-8360, 2020, doi: 10.18653/v1/2020.acl-main.740.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", In: *Proc. of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 4171-4186, 2019, doi: 10.18653/v1/N19-1423.

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach", *arXiv Preprint*, arXiv:1907.11692, 2019, doi: 10.48550/arXiv.1907.11692.

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, *et al.*, "Learning transferable visual models from natural language supervision", *arXiv Preprint*, arXiv:2103.00020, 2021, doi: 10.48550/arXiv.2103.00020.

[23] P. McCullagh, "Regression models for ordinal data", *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 42, No. 2, pp. 109-127, 1980, doi: 10.1111/j.2517-6161.1980.tb01109.x.

[24] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples", In: *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9260-9269, 2019, doi: 10.1109/CVPR.2019.00949.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection", In: *Proc. of IEEE International Conference on Computer Vision (ICCV)*, pp. 2999-3007, 2017, doi: 10.1109/ICCV.2017.324.

[26] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment", *arXiv*

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

Preprint, arXiv:2007.07314, 2020, doi: 10.48550/arXiv.2007.07314.

[27] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system", In: *Proc. of 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, 2016, doi: 10.1145/2939672.2939785.

[28] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions", *arXiv Preprint*, arXiv:1705.07874, 2017, doi: 10.48550/arXiv.1705.07874.

[29] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure", *arXiv Preprint*, arXiv:2203.05794, 2022, doi: 10.48550/arXiv.2203.05794.

[30] A. Lusardi and O. S. Mitchell, "Financial literacy and retirement planning in the United States", NBER Working Paper, No. 17108, 2011, doi: 10.3386/w17108.

[31] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing", In: *Proc. of 11th International Conference on Learning Representations (ICLR)*, 2023.

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/