

X-FCA: Diagnosing Explanatory Decoupling in Financial Information Systems via Counterfactual XAI and Conformal Prediction

Şeyda YONCACI

seydayoncaci@halic.edu.tr

Haliç University

Özal YILDIRIM

Firat University

Engin AVCI

Firat University

Research Article

Keywords:

Posted Date: October 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10555839/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

X-FCA: Diagnosing Explanatory Decoupling in Financial Information Systems via Counterfactual XAI and Conformal Prediction

Şeyda Yoncacı^{1,2*}, Özal Yıldırım² and Engin Avcı²

^{1*}Department of Software Engineering, Haliç University, Faculty of Engineering, İstanbul, Türkiye.

²Department of Software Engineering, Fırat University, Faculty of Technology, Elazığ, Türkiye.

*Corresponding author(s). E-mail(s): seydayoncaci@halic.edu.tr;

Abstract

Automated financial information systems increasingly delegate sentiment-driven decisions to transformer-based language models, yet whether those decisions rest on semantic content or on representational shortcuts is invisible to standard accuracy metrics. Analysing a domain-specific financial transformer, we identify a phenomenon we term *explanatory decoupling*: a robustness intervention can remove a model’s reliance on a spurious cue without altering what the model attends to, so explanation stability cannot be read as evidence of semantic grounding. We establish this with X-FCA (eXplainable Financial Counterfactual Analysis), a diagnostic framework coupling counterfactual perturbation, Conformal Prediction, and token-level attribution, applied to FinTwitBERT over a harmonized corpus of 177,965 microblogs spanning cryptocurrency, traditional finance, and institutional sources, so that any fragility observed reflects the architecture rather than a single dataset. Deterministic emoji ablation reverses predictions 2.77 times more often than a matched random-deletion control, providing causal evidence that visual tokens, not sequence-length sensitivity, drive the fragility. Integrated Gradients trace this to the tokenizer: all ten tested financial emojis, including semantically opposed bullish and bearish pairs, collapse to a single [UNK] token—a pre-inference information loss we term *tokenizer blindness* that leaves the model responding to visual density rather than polarity. Adversarial vaccination removes the coupling between emoji attention and prediction confidence (from $r=0.121$, $p=0.007$, to $r=-0.004$, n.s.) and significantly compresses prediction drift ($p < \mathbf{0.001}$), yet attributional consistency remains statistically unchanged: robustness without semantic regrounding. Conformal

prediction converts the residual overconfident errors into calibrated uncertainty sets. Robustness, uncertainty calibration, and attribution analysis are therefore distinct and non-substitutable axes of reliability assessment: evidence on any one licenses no conclusion about the others.

Keywords: Explainable AI, Conformal Prediction, Counterfactual Analysis, Epistemic Uncertainty, Explanatory Decoupling, Tokenizer Blindness

1 Introduction

As the scale of financial microblogging grows exponentially, social media sentiments have evolved into a high value leading indicator which is very effective for algorithmic trading, affecting volatility of stock prices [1] and market liquidity [2]. To effectively exploit and manage the colossal data, Transformer-based architectures and LLM-aided workflows become ubiquitous and are de facto standard achieving state-of-the-art performances with financial document processing [3]. Unfortunately, the above outstanding performance at a macro level masks the crucial limitation inherent in this data: financial tweets are very noisy, and the performance of models trained on this text needs robust noise eliminating mechanism in order to provide unbiased decisions [4]. It also contains severe structural limitations regarding sequence length (e.g., brevity), professional vocabulary (i.e. Jargon) as well as dependency of sentiments and trading sentiments, most specifically on emojis for expressing predictions (e.g. Rocket emojis indicates bulls or downward charts indicates bearishness [5]). Most conventional deep learning model evaluated on fixed validation dataset like F1, Accuracy fail to show model-driven insights or decision. Recent work integrates explainable AI directly into sentiment-driven financial forecasting to make model outputs interpretable to stakeholders [6], yet such explanations are typically presented as trustworthy without testing whether they remain faithful under perturbation. We consider the most essential and challenging questions: do the deep learning models truly grasp complex financial semantics, or are they just memorizing visual shortcuts? When models anchor and infer on emojis, using them as purely visual anchors and correlation elements, that would create and provide a deceptive sense of reliability. And due to their ambiguous semantics, consumer trust and model performances get confused [7], potentially creating enormous business capital risks and financial liquidity shocks [2] in an automated HFT setting with adversarial attacks (i.e. 'emoji-spam'), out-of-distribution dataset and data leakage into model predictions. In light of these concerns, we introduce X-FCA, an eXplainable Financial Counterfactual Analysis tool. The X-FCA aims at not only providing state-of-the-art results but also exposing and quantifying any reasoning collapse based on cognitive robust structure. We propose the use of a huge cross-domain master corpus with 177,965 harmonised microblogs which includes Crypto, Traditional Finance (TradFi) and Institution markets to achieve that objective. Through targeted, counterfactual probes that involve a tripartite contrastive attack - that is deterministically masking out emojis vs randomly deleting tokens, we provide an empirical verification on a causal basis that isolates dependence on

visual anchors by tracking uncertainties. We then apply Conformal Prediction (CP) [8, 9] and Integrated Gradients (IG) [10, 11] via the Kernel Density Estimation (KDE) on uncertainty shift, finally revealing and mathematically quantifying architecture deficiencies.

The main contributions of this paper are fourfold:

- **X-FCA Paradigm and Cross-Domain Harmonization:** We propose a diagnostic methodology that couples Conformal Prediction with counterfactual explainable AI to audit the epistemic reliability of financial information systems. We test it on a harmonized master corpus of 177,965 financial microblogs spanning Cryptocurrency, Traditional Finance, and Institutional sources, so that any fragility we observe reflects the architecture rather than a single dataset.
- **Causal Isolation of Visual Dependency and the “Threshold Effect”:** Through a tripartite control experiment, we demonstrate the brittleness of a domain-specific financial language model. Deleting a randomly selected alphanumeric token flips only 2.50% of predictions, whereas deterministically removing emojis flips 6.94%—a disproportionate collapse. The effect is also non-linear: instability rises to 10.0% once a sequence carries three or more emojis, a pattern we term the “Threshold Effect”.
- **Structural Evidence of “Tokenizer Blindness”:** Through global attribution-shift analysis (KDE plots of Integrated Gradients), we trace this fragility to the tokenizer itself: all ten tested financial emojis collapse *deterministically* into the [UNK] token. This pre-inference loss of information leaves the model responding to visual *density* rather than semantic polarity.
- **Definition of “Explanatory Decoupling”:** We document a previously uncharacterised phenomenon at the intersection of robustness and interpretability, which we call *explanatory decoupling*: adversarial vaccination removes the spurious coupling between emoji attention and prediction confidence, yet leaves token-level attribution patterns statistically unchanged. A model can thus become more robust without becoming more semantically grounded a foundational caution for the design of trustworthy knowledge-based systems.

2 Related Work

The intersection between model interpretability and financial sentiment analysis has achieved popularity in recent time. This chapter introduces the X-FCA model in context of previous works, identifying major insufficiencies in three basic perspectives namely, analyzing Transformer models in finance NLP, understanding semantics of visual anchor and the rising trend of counterfactual robustness in financial NLP task.

2.1 Transformer-based Financial Sentiment Analysis and Evaluation Constraints

The past few years have seen the overwhelming popularity of Transformers, specifically BERT and its variants like FinBERT and FinTwitBERT, and even newer generative

models like FinGPT [12], within the analysis of financial microblogging. These models outperform older lexicon and LSTM models at grasping contextuality [13], but the majority of their assessments have been limited to static macrometrics, such as F1 score or Accuracy. As recent studies have noted, these consolidated metrics represent a crucial problem: they tend to conceal the model’s susceptibility to out-of-distribution (OOD) inputs and adversarial perturbations. Nyong [14] argues that in the context of financial high-stake decision systems, models with high predictive power appear to have an opaque “black-box” nature, which limits regulation and trust. These models also exhibit what Sait and Alkhurayyif [15] calls “hallucinated sentiment”: a phenomenon in which models that seem to produce an overconfident prediction, lack any linguistic support and form a risky mirage of regulatory compliance and trust. Even with successful models that achieve cutting edge accuracy, there still exists a problematic “reasoning gap” that disassociates complex financial and linguistic volatility that characterizes the FinTwit space [12]. Most evaluation attempts have relied on fragmented, limited datasets and have been unable to capture the systemic threats posed by these models [16].

Evidence suggests that transformer-based classifiers learn surface level cues as opposed to learning the meaning behind words and concepts. This has been studied in shortcut learning literature [17]. Most importantly, text classifiers that analyze social media content experience dramatic drops in performance when exposed to naturally occurring and intentionally made shortcut cues. Sentiment normalization must be used to get misleading classifiers to learn deeper meaning in semantics. Sentiment normalization is similar to the emoji ablation strategy used in this research. Classifiers also learn surface level correlations Shaib et al. [18], which dominate the prompt semantics. This is true even for advanced models like OLMo-2, Llama-4, and GPT-4o. Our research assesses the concerns expressed by the above researchers. We validate these concerns on a large corpus containing 177,965 texts. This is a cross-domain corpus containing texts from the Cryptocurrency, Traditional Finance, and Institutional domains. This ensures that the concerns expressed by the researchers regarding lacking semantic learning and describing model architectures are not caused by the datasets used.

2.2 The Semantic Value of Emojis and “Tokenizer Blindness”

Emojis have shifted from mere ornamental symbols to highly-valuable semantically-weighted signals in social media discourse, especially in high-frequency financial communication channels in which short textual inputs are usually accompanied by images. Recent findings show that emojis convey considerable portions of the sentiment signal, which could be highly influential in shaping investors’ perception and future behavior [19]. Additionally, recent empirical findings show that visual signals in volatile markets (cryptocurrencies) could serve as powerful signals for traders, working as early predictors of crowd wisdom [20]. Last, but not least, state-of-the-art multimodal feature aggregation systems take these visual signals as key components of profile-level sentiment assessment, recognising that visual signals could reverse the meaning of the associated textual content in ironic or sarcastic context [21].

While the above-mentioned evidence highlights the importance of non-alphanumeric characters in conveying semantic value, most of contemporary NLP pipelines still rely on subword tokenisation techniques (e.g., BPE, WordPiece), that fail to properly represent the latter category. In the case of many transformer architectures, visual elements are usually either mapped to out-of-vocabulary (OOV) or [UNK] tokens, resulting in a disastrous loss of the semantic information in the early stage of pipeline [22]. This problem is especially acute in the specialised domain like Bitcoin sentiment analysis, where omission of visual tokens causes an observable reduction in intraday predictability and accuracy [23].

This structural limitation brings the weakness in modern language models, which we call “*Tokenizer Blindness*”—a situation when distributional hypothesis breaks down because the “oft-overlooked” tokenisation layer fails to make a connection between visual tokens and semantics [24]. Although the previous literature already acknowledges the general importance of emojis, the density-dependent effects of those elements remain understudied.

In our research, we introduce and validate the “*Threshold Effect*”, which claims that the stability of transformers’ predictions with respect to the presence of emojis is not proportional and undergoes dramatic changes at some specific density points. Specifically, when the number of visual tokens exceeds a certain limit, it provokes what we call “reasoning instability,” in which the model’s attention mechanisms lose their structural integrity under the stress of visual token sequences [25]. This way, we identify and empirically validate the threshold point, making it possible to understand how emoji-dense adversarial patterns can bypass the cognitive capacities of financial NLP systems.

2.3 Counterfactual Explainability Under Model Robustness

The paradigm shift toward Robust Artificial Intelligence has set up counterfactual analysis as a prominent tool to study “what-if” scenarios in ML systems. Counterfactual explanations have become a standard for analyzing model behaviour and auditing biases in general NLP tasks, starting from detecting demographic and racial biases and going up to neuro-symbolic reasoning audits [8, 26]. In financial applications specifically, counterfactual explanations have been used to derive actionable interventions for instance, identifying the feature changes that would reduce a bank’s predicted failure risk [27]. Recent research points out the significance of distribution shifts and test-time adaptation for the model to avoid catastrophic failure [28, 29].

Notwithstanding all the advancements, the use of counterfactual methods as a diagnostics approach for spotting spurious visual dependencies in financial sentiment models appears to be extremely scarce. Earlier works on financial NLP have paid much attention to hybrid architectures’ performance optimisation, such as Transformer + LSTM or attention mechanisms in RoBERTa model [30–32]. Even though these methods are very efficient and yield top macro-level metrics in benchmark corpora like S&P 500 or cryptocurrency signals [20, 33], the primary emphasis is made on achieving predictive efficiency rather than causal interpretation or robustness analysis under non-alphanumeric adversarial conditions.

Simultaneously, the field of Explainable Artificial Intelligence (XAI) has introduced several attribution-based methods, including SHAP, Integrated Gradients, and their specific modifications for financial audits [10, 11]. As shown by Rizinski et al. [34], even though state-of-the-art financial transformers perform quite well, their interpretability is poor and requires to go back to lexicon-based explanations. In turn, this calls for diagnostic framework like X-FCA. These methods, however, are mostly qualitative since they provide visual heatmaps the interpretations without any statistical justifications regarding explanation reliability [14, 35]. Most relevant to our research, Duell et al. [36] solve the issue of uncertainty in counterfactual explanations, proposing QUCE as a way of reducing out-of-distribution path traversal in IG-based counterfactuals an approach that we address in our tripartite perturbation design that ensures semantic validity through comparison of emoji ablation with random token deletion.

Recent research has attempted to tackle this issue via uncertainty quantification methods, including Conformal Prediction (CP) as an approach yielding statistically reliable confidence intervals [8, 9]. Campos et al. [37] prove CP to be the best theoretically justified framework for uncertainty quantification in NLP, due to the fact that it is model-agnostic and does not assume distribution-free statistical assumptions precisely the same properties that are at the basis of our tail-risk isolation approach. Giovannotti [38] also show that applying conformal methods to transformer-based classifiers results in more reliable uncertainty estimates than standard post-hoc calibration an effect that we observe when applying CP in our model. Vellore and Jha [39] propose CONFIDE, a new approach of applying CP directly to BERT and RoBERTa internal embeddings to yield statistically valid prediction sets the closest work to our application of CP in FinTwitBERT. Yet, the use of Conformal Prediction with attribution-based explainability methods to audit visual-linguistic dependencies in financial information systems has not been studied at all. Closest to our concern, Ayatollahi and Jafari [40] contrast predictive robustness with what they term *explainability stability* across machine-learning models forecasting equity returns, reporting a decoupling in which strong predictive performance coexists with highly unstable attributions. Their analysis operates on tabular market features at the model-comparison level; we instead isolate the inverse regime at the token level showing that a robustness intervention can leave attributions statistically unchanged while severing their coupling to prediction confidence and add distribution-free guarantees through Conformal Prediction.

By introducing two methodological novelties tripartite causal baseline, where deterministic visual ablation is contrasted with stochastic token deletion in order to spot spurious visual correlations (e.g., emoji-anchoring) in financial models and the integration of Conformal Prediction with global attribution shift analysis using Kernel Density Estimation (KDE) over IG distributions in order to quantify the epistemic uncertainty triggered by perturbations our framework, X-FCA fills this theoretical gap.

3 Methodology

Traditional financial information systems are making use of language models that are based on transformers for sentiment-driven decision-making, but epistemic validity of the knowledge representations constructed by these models is not well-understood. Static evaluation metrics such as aggregated F1-scores do not help us identify whether a model has indeed acquired semantic knowledge or overfit to the visual cues (e.g., emojis) in the input as spurious correlations. In order to address this crucial challenge, we introduce the X-FCA diagnostic framework. The entire pipeline is depicted in Fig. 1. This framework leverages CP and XAI techniques in a deterministic counterfactual setting for auditing the epistemic reliability of financial language models.

The X-FCA framework comprises four layers in a sequence, namely, (i) harmonization of data and hype-score filtering, (ii) tripartite counterfactual perturbation, (iii) conformal prediction for tail risk isolation, and (iv) cognitive diagnostics using XAI.

3.1 Task Formulation, Data Harmonization, and Hype Scoring

Financial sentiment analysis is formalized as a supervised classification task where an input text sequence $x \in \mathcal{X}$ is mapped to a discrete sentiment label $y \in \mathcal{Y} = \{0, 1, 2\}$, corresponding to Negative, Neutral, and Positive polarities. We deploy the FinTwit-BERT architecture, pre-trained specifically on financial microblogging corpora, as our baseline model f_θ . To ensure a rigorous evaluation without data leakage, input sequences are tokenized with a maximum length of 256 tokens and validated using a strict 5-Fold Cross-Validation protocol. The baseline softmax probability distribution is defined as:

$$\hat{p}(y|x) = \sigma(f_\theta(x))_y = \frac{e^{f_\theta(x)_y}}{\sum_{j=1}^K e^{f_\theta(x)_j}} \quad (1)$$

To achieve across-domain generality, while ensuring maximum exposure to extreme linguistic volatility, we combined 177,965 microblogs from Cryptocurrency, Traditional Finance, and Institutional corpora. The distribution of these data can be observed at Fig. 2. We applied a tight, stratified split (precisely 10% $N = 17,797$) as the pure hold-out test set. To find sequences in the test set most sensitive to disruption, we generated an algorithmic "Sarcasm and Hype Scoring" function ($Shype$) that searches for highly prevalent financial buzzwords (e.g., 'rekt', 'moon', 'ponzi') and visually dense clumps. Sequences with a $Shype$ score above the required threshold ($S_{hype} \geq 3.0$) were exclusively reserved to form the extremely volatile "Tail-Risk" candidate corpus ($N = 1,348$).

3.2 Tripartite Counterfactual Perturbation Strategy

To understand if the models' reasoning is truly driven by financial linguistics or by artificial hypes, we use a set of tripartite counterfactual testing, which can be viewed at the *Inference* lane of Fig. 1. For each hyperbolic original tweet x_{orig} , we generate two counterfactual versions: one neutralized twin x_{twin} and one random control x_{random} .

The deterministic counterfactual twin generation is a composite function defined as:

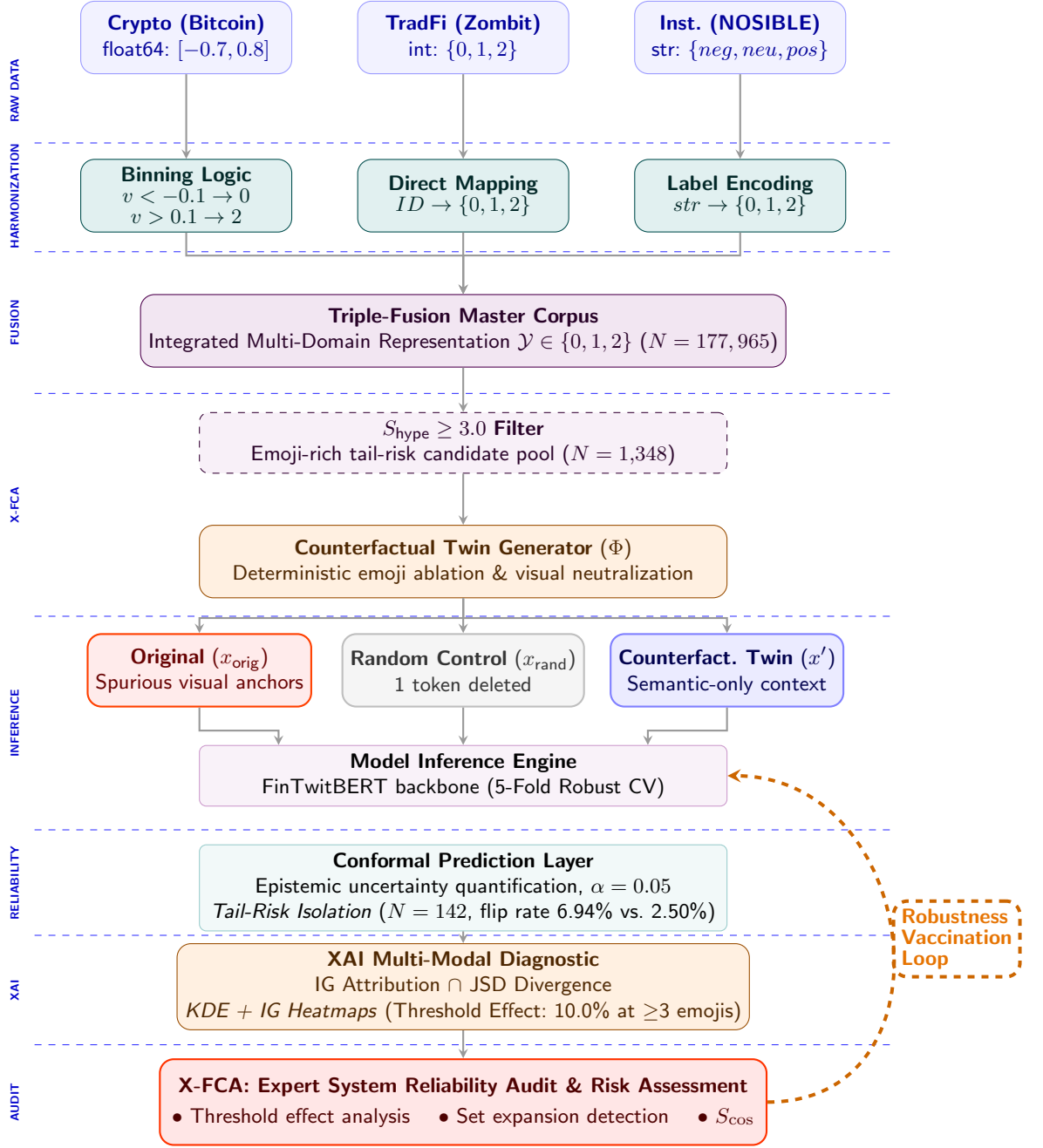


Fig. 1 The proposed X-FCA system architecture. The framework integrates a Triple-Fusion master corpus (Cryptocurrency, Traditional, and Institutional Finance) through a S_{hype} -based filtering stage into a **tripartite** counterfactual diagnostic pipeline (Original, Random Control, and Counterfactual Twin). It leverages Conformal Prediction for statistical tail-risk isolation ($N = 142$), XAI (KDE + IG Heatmaps) for quantifying reasoning shifts, and a dynamic feedback loop for iterative model reinforcement.

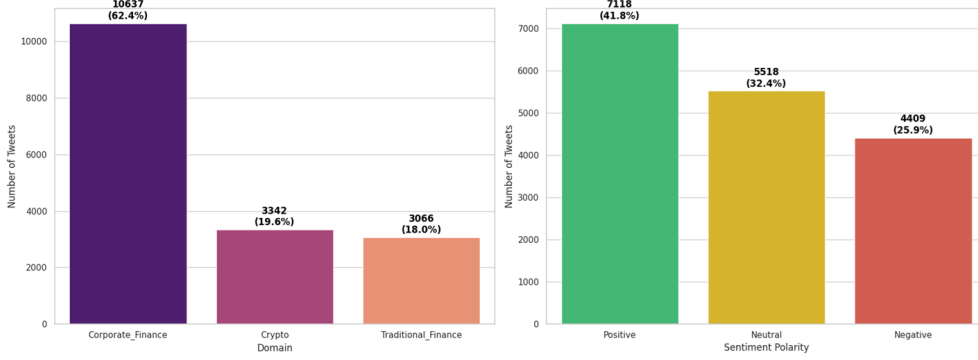


Fig. 2 Dataset diversity and domain distribution of the Triple-Fusion Master Corpus test set ($N = 17,797$). Left: domain distribution across Corporate Finance, Crypto, and Traditional Finance. Right: sentiment class distribution (0=Negative, 1=Neutral, 2=Positive).

$$\Phi_{\text{twin}}(x) = \Phi_{\text{emoji}}(\Phi_{\text{lexical}}(x)) \quad (2)$$

1. **Lexical Neutralization** (Φ_{lexical}): Maps hyperbolic financial jargon to formal semantic equivalents (e.g., “to the moon” $\rightarrow$ “expected to rise”, “rekt” $\rightarrow$ “liquidated”). This transformation is strictly governed by a domain-specific lookup dictionary to ensure that the underlying financial sentiment remains invariant while the “hype-signature” is removed.
2. **Visual Ablation** (Φ_{emoji}): Deterministically strips the sequence of all defined Unicode financial emojis (E), leaving a clean, formal text. This stage isolates the model’s reliance on visual anchors versus linguistic semantics.

To establish a causal baseline and control for general sequence-length sensitivity, we introduce the random token ablation function $\Phi_{\text{random}}(x)$, which removes a single, randomly selected alphanumeric token:

$$x_{\text{random}} = \Phi_{\text{random}}(x_{\text{orig}}) = x_{\text{orig}} \setminus \{t_{\text{rand}}\} \quad (3)$$

The divergence in prediction flip rates (Δ_{flip}) between x_{twin} and x_{random} serves as our primary metric for quantifying spurious dependency. This rigorous tripartite approach ensures that any observed “Decision Boundary Collapse” on the formal twins is exclusively attributable to the model’s over-reliance on noise and Tokenizer Blindness, rather than external semantic shifts.

3.3 Conformal Prediction (CP) for Tail-Risk Isolation

A core novel methodological aspect of X-FCA relies on the usage of CP to assess predictive epistemic uncertainty with strict statistical warranties. Whereas prediction based on softmax thresholds can be overly confident and poorly-calibrated in the OOD setting (due to reliance on ad-hoc heuristics), we make use of rigorously defined prediction sets. Based on a given significance level of $\alpha = 0.05$, (meaning a 95%

confidence level), we create prediction sets $C(x)$ built upon cumulated densities of sorted softmax scores.

For a given test sample x , the CP set satisfies the following marginal coverage guarantee:

$$\mathbb{P}(y \in C(x)) \geq 1 - \alpha \quad (4)$$

In the context of these analyses, we focus on the so called *Outliers* which correspond to instances where emoji ablation leads to an increase in the conformal prediction set size. Previous derivations limit tail-risk cases to singleton prediction sets ($|C(x_{\text{orig}})| = 1$) which our experiments showed to be less common for the FinTwitBERT model on large training set, that tends to return a large proportion of length-2 prediction sets ($|C(x_{\text{orig}})| = 2$; $N = 139$) even at a very high confidence as it captures more granular inter-class dependencies through its design for multilabel classification setup. A sequence is then strictly classified as a “*Tail-Risk*” one if it meets *Reasoning Collapse* on neutralization with the conditions:

$$(y_{\text{orig}} \neq y_{\text{twin}}) \vee (|C(x_{\text{twin}})| > |C(x_{\text{orig}})|) \quad (5)$$

Here, we define two types of reasoning failures: (i) A *complete prediction flip* ($N = 18$), in which the addition of visual anchors shifts the predicted class from what it was with their removal. (ii) A *major uncertainty expansion* ($N = 124$), wherein a model is still able to classify at its initially predicted class label, but it can no longer confirm the expected number of potential classes. Together these constitute the *Tail-Risk* cases that we consider of most concern ($N = 142$).

We inspect the generated Tail-Risk cases with Layer Integrated Gradients(IG) on the input embeddings for all identified cases, reasoning that the causes could be the underlying failures to reason through first-order logic in addition to failures in other parts of the Model-based reasoning architecture. The IG method is selected due to its axioms of *completeness* and *sensitivity* such that the full change of the prediction from baseline to model inference is the sum of attribution for all input elements. We derive input x_i based on an interpretation from the original question for each token x_i at each time step in N , with $m = 50$ integration steps.

$$\text{IG}_i(x) = (x_i - x'_i) \times \int_0^1 \frac{\partial f(x' + \beta(x - x'))}{\partial x_i} d\beta \quad (6)$$

where x' represents the baseline uninformative embedding vector. To rigorously quantify the *Attribution Shift* (cognitive displacement) caused by the perturbations, we normalize the raw attribution scores into probability distributions (P, Q) representing the original and twin inputs, respectively, and measure the Jensen-Shannon Divergence (D_{JS}). We prefer D_{JS} over Kullback-Leibler divergence due to its symmetric properties and its bounded nature in $[0, 1]$, which provides a more stable metric for knowledge-based system auditing:

$$D_{\text{JS}}(P \parallel Q) = \frac{1}{2} D_{\text{KL}}(P \parallel M) + \frac{1}{2} D_{\text{KL}}(Q \parallel M) \quad (7)$$

where $M = \frac{1}{2}(P + Q)$.

Beyond global attribution shift, we additionally compute the *Attributional Consistency* score S_{cos} as the cosine similarity between the IG attribution vectors of the original and twin inputs:

$$S_{\text{cos}}(x_{\text{orig}}, x_{\text{twin}}) = \frac{\mathbf{g}_{\text{orig}} \cdot \mathbf{g}_{\text{twin}}}{\|\mathbf{g}_{\text{orig}}\| \cdot \|\mathbf{g}_{\text{twin}}\|} \quad (8)$$

where $\mathbf{g}$ denotes the token-level IG attribution vector. A value of $S_{\text{cos}} \approx 1$ indicates that the model attends to the same tokens before and after perturbation (stable reasoning), while $S_{\text{cos}} \approx 0$ or negative values indicate a fundamental restructuring of the model’s attention mechanism (cognitive collapse).

By stratifying D_{JS} and S_{cos} distributions across *Flipped* and *Non-Flipped* Tail-Risk cases via Kernel Density Estimation (KDE), we provide definitive empirical evidence of “Cognitive Collapse,” as quantified in Section 4. This stark separation confirms that decision boundary collapse is not a marginal probabilistic shift but a complete internal reasoning restructuring: the model fails to maintain any semantic coherence once its visual anchors (emojis) are removed, validating the presence of Tokenizer Blindness within the architecture.

3.4 Tokenizer Blindness: Structural Evidence

To identify the architectural origin of this shortcut sensitivity of the model we look at the tokenizer process just before it enters the embedding layer projection. As seen in Fig. 3 each of the ten financially sensitive emojis (this includes contrasting pairs such as a rocket and bleeding drop etc) are all tokenized into a unique and deterministic [UNK] token (ID=2). This is a *pre-inference information collapse*: the model loses its ability to differentiate between an indicative bullish and its semantically opposing counterpart before even a single attention head has fired. Without the ability to extract a financially informative signal from different emojis, the attention mechanism only has the statistical *density* of uncategorized [UNK] tokens—the architectural source of all the false positives in 4.

Density-Dependent Consequence.

This pre-inference collapse implies that a model’s fragility does not grow linearly with the number of emojis, but a *Threshold Effect* of non-linear fragility: Since all emojis map to the same [UNK] token, the model is responding to the [UNK] *density* rather than their semantic meaning, leading to fragility spikes only at some clustering thresholds as observed. We show and measure this phenomenon in Section 4.3.

3.5 Data Leakage Prevention: The Counterfactual Barrier

For methodological purity, and to avoid having our robustness test undermined by data poisoning, we developed a sophisticated multi-tiered *Data Leakage Barrier*. As clearly illustrated in our overall architecture, our defense procedure enforces strict, full temporal and semantic isolation within two principal stages of operation:

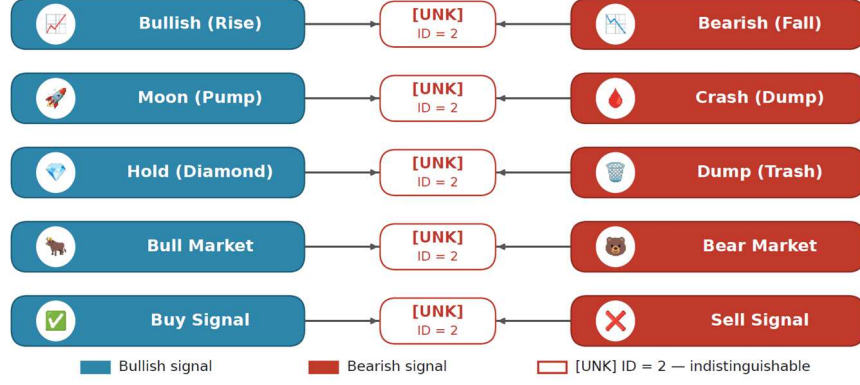


Fig. 3 Tokenizer Blindness: All ten tested financial emojis, including semantically opposite bullish–bearish pairs, collapse to the identical [UNK] token (ID=2), rendering the model structurally incapable of distinguishing directional market signals at the tokenization layer.

1. **Isolation at Source:** The Triple-Fusion Master Corpus ($N = 177,965$) is partitioned via stratified sampling into a 90% training set ($N = 160,168$) and a 10% hold-out test set ($N = 17,797$).
2. **Counterfactual Filtering:** Generated twins undergo a strict cross-partition audit. Only twins whose original counterparts (x_{orig}) reside in the training set are eligible for augmentation. Of 5,473 candidate twins, 4,463 are certified safe; 1,010 are permanently discarded as their originals belong to the hold-out set.

3.5.1 Balanced Adversarial Augmentation (Vaccination)

- **Corporate Finance:** Owing to the natural scarcity of high-hype expressions within this sector, all 139 empirically verified corporate twins were integrated into the robust training partition.
- **Crypto and TradFi:** To mitigate extreme semantic volatility in these domains, we enforced a strict *stratified-cap* threshold. Balancing constraints yielded 416 Crypto twins and 480 TradFi twins, mapped proportionally across the localized target sentiment label space $\mathcal{V} \in \{0, 1, 2\}$.

Final reinforced dataset The final reinforced data set ($N = 161,203$) has also been fully deduplicated and resulted in a net injection of 1,035 additional counterfactual twins (0.64% of the entire training set, see Table 1). We set these injections to a small ratio so that every robustness advantage we see in Section 4 comes directly from the targeted attack training and not from the distributional shift.

4 Quantitative Results

In this section we empirically validate our model, the FinTwitBERT model, transitioning to dynamic counterfactual evaluations away from static performance measures so as to isolate behavioural change across the original and hyper-neutral versions of the dataset, as the counterfactuals can distinguish where changes arise from their change

Table 1 Comprehensive Data Distribution: Training (Original vs. Twin) and Isolated Test Sets. ‘Twin Inj.’ represents the adversarial vaccination applied exclusively to the training partition to ensure zero leakage. Crypto and TradFi twin counts reflect empirical availability after stratified-cap filtering.

Category	Corporate	Crypto	Trad. Fin.	Total
<i>Training Set:</i>				
Original Samples	89,951	36,368	33,849	160,168
Twin Injections	139	416	480	1,035
Robust Train Total	90,090	36,784	34,329	161,203
<i>Testing Set:</i>				
Total Test (Isolated)	10,049	4,034	3,714	17,797

in form rather than solely their change in quantity. Two-sided Mann-Whitney U-tests are employed to determine statistical significance throughout at the level $p < 0.05$ unless indicated otherwise.

4.1 Baseline Model Performance and Stability

Prior to any counterfactual perturbation, the architecture’s predictive capability was evaluated on the fully isolated 10% hold-out test partition ($N = 17,797$) to establish a rigorous baseline. Table 2 presents the metrics across all five folds.

Table 2 Baseline Performance of FinTwitBERT via 5-Fold Cross-Validation on the isolated hold-out test set ($N = 17,797$).

Fold	Accuracy (%)	Macro F1 (%)	Neg F1 (%)	Val. Loss
Fold 1	79.29	78.84	76.45	0.1877
Fold 2	78.97	78.57	76.23	0.1901
Fold 3	79.58	79.12	76.80	0.1868
Fold 4	79.32	78.99	76.61	0.1634
Fold 5	79.63	79.27	77.13	0.1596
Mean $\pm$ SD	79.36 $\pm$ 0.26	78.96 $\pm$ 0.27	76.64 $\pm$ 0.34	0.1775 $\pm$ 0.014

The tight fold-level variance ($SD \approx 0.26\%$) confirms training stability, while the monotonically decreasing validation loss ($0.1877 \rightarrow 0.1596$) indicates progressive calibration improvement. Despite achieving a robust Macro F1 of $78.96 \pm 0.27\%$, these aggregate metrics are fundamentally insufficient as reliability indicators: a model exploiting visual shortcuts can produce identical scores while remaining critically fragile under adversarial perturbation. This diagnostic gap directly motivates the X-FCA counterfactual pipeline.

4.2 Causal Isolation: Tripartite Perturbation Analysis

To confirm that the decision boundary collapse is due specifically to the visual anchors and not sequence-length sensitivity we measure the prediction flip rates across the tripartite perturbation strategies over the emoji-rich set ($N = 1,348, S_{\text{hype}} \geq 3.0$) and the overall test set ($N = 17,797$). Table 3 shows that deterministic emoji ablation increases the flip rate by 6.94%, an increase of $2.77\times$ over the 2.50% random deletion control. The Traditional Finance domain shows the most flip rate (2.21% at BASE) while Corporate Finance shows the least (0.06% at BASE), revealing variations in hype-language density across market sectors. This susceptibility is tied to the architecture and specifically the *Tokenizer Blindness* (Section 3.4): since all finance-related emojis collapse to the single [UNK] token, the model bases its decision solely on visual *density* and not the semantic polarity of the emoji.

Table 3 Causal Isolation: Prediction Flip Rates Across Perturbation Strategies.

Strategy	Flip Rate	Ratio	Scope
Emoji Ablation (Twin, x')	6.94%	$2.77\times$	$S_{\text{hype}} \geq 3.0$ ($N = 1,348$)
Random Deletion (Control)	2.50%	$1.00\times$	$S_{\text{hype}} \geq 3.0$ ($N = 1,348$)
BASE (Global)	0.65%	—	Full set ($N = 17,797$)
ROBUST (Global)	0.63%	—	Full set ($N = 17,797$)

Applying the X-FCA diagnostic pipeline to the BASE ($N_{\text{BASE}} = 142$) and ROBUST ($N_{\text{ROBUST}} = 202$) Tail-Risk cohorts (Figs. 4–5) yields 3 main observations. First, the BASE tail-risk flip rate of 12.68% is decreased to 6.93% in the ROBUST model, for a relative improvement of 45.3%. Second, mean JSD goes down from 0.1529 to 0.0994 ($\downarrow 35.0\%$), reflecting the more invariant probabilities under perturbation. Third, mean attribution consistency S_{cos} goes down insignificantly ($0.6743 \rightarrow 0.6727$, $\Delta = -0.002$), but correct prediction retention goes up from 88.7% to 91.1%. This robustness-at-the-cost-of-attribution argument is discussed in Section 6.2.

4.3 The Threshold Effect: Density-Dependent Instability

Tokenizer Blindness not only does not scale linearly with the number of emoji, it produces a very specific *Threshold Effect* on model fragility, see Fig. 6: instability surges up to 10.0% within the 3-emoji-group region, considerably higher than the average 7.7%, proving that models respond to *density* and not semantic information. *ROBUST* addresses this by bringing it down to 5.0% while only needing 0.64% of data to be poisoned - an exponential increase in efficiency.

4.4 Tail-Risk Conformal Prediction Analysis

Reasoning Collapse criterion application ($y_{\text{orig}} \neq y_{\text{twin}} \vee (|C(x_{\text{twin}})| > |C(x_{\text{orig}})|)$) results in $N_{\text{BASE}} = 142$ samples (18 flips + 124 expansions of CP sets) and $N_{\text{ROBUST}} =$

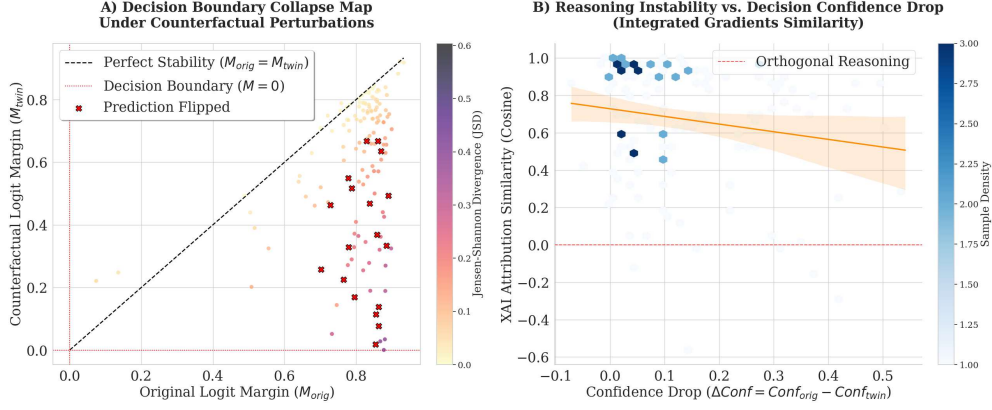


Fig. 4 X-FCA Diagnostic Analysis: BASE Model Tail-Risk Subset ($N = 142$). **Panel A:** Decision Boundary Collapse Map; colour encodes JSD and red crosses mark flipped predictions ($N = 18$). High-JSD cases cluster in the overconfident regime ($M_{orig} > 0.80$). **Panel B:** Attribution Consistency (S_{cos}) vs. Confidence Drop ($\Delta Conf$); the negative regression slope confirms that larger confidence loss corresponds to deeper internal reasoning disruption.

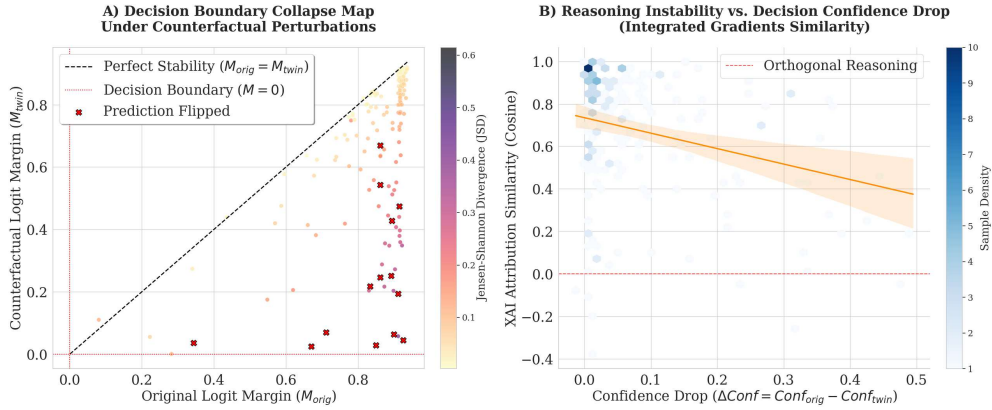


Fig. 5 X-FCA Diagnostic Analysis: ROBUST Model Tail-Risk Subset ($N = 202$). **Panel A:** Red crosses mark flipped predictions ($N = 14$), a 22.2% reduction vs. BASE. **Panel B:** The negative S_{cos} - $\Delta Conf$ coupling is preserved with a wider calibration interval, reflecting the ROBUST model's more conservative uncertainty posture.

202 tail-risk cases. The ROBUST case pool is larger due to the fact that the vaccinated predictor is more conservative in uncertainty estimation: by reducing its decision confidence in the case of uncertainty, the model detects a higher number of sequences which should be verified in downstream processes an advantageous property for a safety-critical deployment rather than a degradation of performance. The calibrated APS predictor provides empirical 99.72% coverage (target: 95.0%) with mean prediction set size 2.54, 96.9% of errors being turned into uncertainty sets.

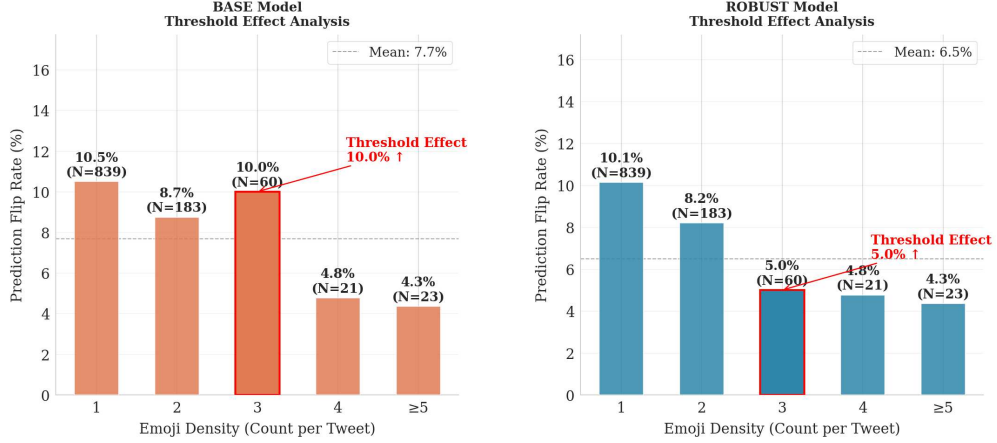


Fig. 6 The Threshold Effect: Prediction Instability vs. Emoji Density ($S_{\text{hype}} \geq 3.0$ subset; $N_{\text{bin}} = [839, 183, 60, 21, 23]$). BASE model (left): non-linear instability spike at 3-emoji clusters (**10.0%**), above global mean of 7.7%. ROBUST model (right): spike suppressed to 5.0% ($\downarrow 50\%$), below ROBUST global mean of 6.5%.

The results shown in Table 4 constitute the complete matrix of robustness comparisons. Apart from the increase in the flip-rate and in the JSD score mentioned above, the ROBUST model also yields an improvement of 34.8% in the compression of confidence drop ($p < 0.001$), whereas S_{cos} stays unchanged (n.s.), a crucial result that is discussed further in Section 6.2. As shown in Fig. 7, the ROBUST model’s more balanced CP set-size distribution ($|C(x_{\text{orig}})|$: 1.98 $\rightarrow$ 1.45) additionally signals improved baseline calibration independent of perturbation.

Table 4 Comparative Robustness Matrix on Model-Specific Tail-Risk Subsets (** $p < 0.001$; n.s. not significant, Mann-Whitney U).

Metric	BASE	ROBUST	Δ	Status
Tail-Risk N	142	202	+60	—
Adversarial Flips (N)	18	14	−4	$\downarrow 22.2\%$
Flip Rate (%)	12.68	6.93	−5.75	$\downarrow 45.3\%$
Mean JSD	0.153	0.099	−0.053	$\downarrow 35.0\%^{***}$
Mean S_{cos} (%)	67.4	67.3	−0.1	n.s.
Mean Conf. Drop (%)	13.1	8.6	−4.6	$\downarrow 34.8\%^{***}$
Mean Orig. Conf. (%)	88.4	91.8	+3.4	Improved
Mean $ C(x_{\text{orig}}) $	1.98	1.45	−0.53	Calibrated

4.5 Cognitive Collapse: Mechanistic Failure Evidence

To advance from behavioural tracking to mechanistic validation, we analyse the causal antecedents of decision boundary failure via token-level Integrated Gradients (IG). As

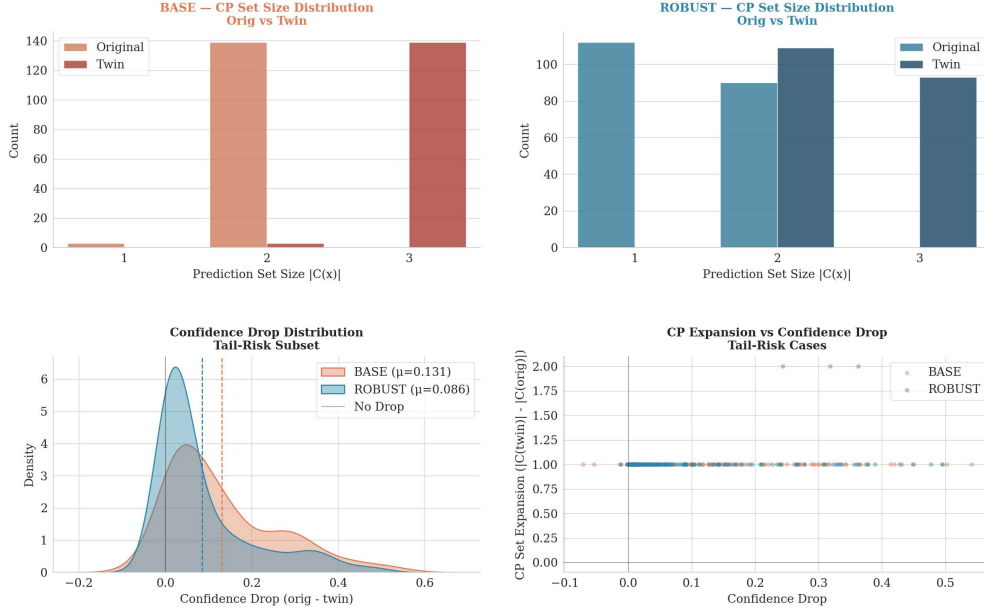


Fig. 7 X-FCA Conformal Prediction Analysis ($\alpha = 0.05$, $S_{\text{hype}} \geq 3.0$). **Top Row:** CP set-size distributions for BASE ($N = 142$) and ROBUST ($N = 202$); all cases exhibit set expansion. The ROBUST model shows a more balanced size-1/size-2 distribution, indicating improved baseline calibration. **Bottom-Left:** Confidence drop KDE; BASE ($\mu = 0.131$) is shifted rightward relative to ROBUST ($\mu = 0.086$). **Bottom-Right:** CP expansion vs. confidence drop scatter; uniform magnitude of +1 confirms one-step epistemic uncertainty escalation.

demonstrated in Fig. 8 and quantified in Table 5, flipped predictions exhibit $4.65\text{--}5.96\times$ higher JSD divergence than non-flipped counterparts ($p < 0.001$), empirically confirming that emoji over-reliance is a strong antecedent of not merely correlated with decision collapse, and that adversarial vaccination reduces the magnitude of visual anchoring rather than merely shifting decision thresholds.

Table 5 Cognitive Collapse Diagnostics: JSD and Emoji Attribution in Flipped vs. Not-Flipped Tail-Risk Cases (Mann-Whitney U).

Model	Mean JSD		Ratio	Mean Emoji Attn.		p
	Flipped	Not-Flipped		Flipped	Not-Flipped	
BASE	0.486 ($N = 18$)	0.105 ($N = 124$)	$4.65\times$	23.4%	7.6%	<0.001
ROBUST	0.441 ($N = 14$)	0.074 ($N = 188$)	$5.96\times$	16.0%	6.5%	<0.001

To control for the sample composition bias among the tails, we test the ROBUST model only in these specific 18 cases where BASE failed. ROBUST still achieves right classification on 7 of these 18 failed cases (38.89% of recovered), and the 11 shared

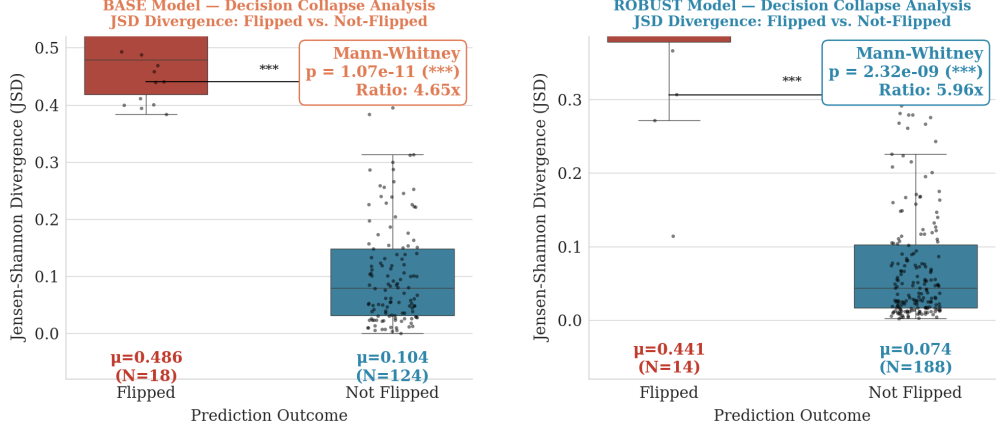


Fig. 8 Cognitive Collapse Under Emoji Ablation: JSD Divergence in Flipped vs. Not-Flipped Tail-Risk Cases ($p < 0.001$, Mann-Whitney U). Flipped predictions exhibit 4–6 $\times$ higher JSD, constituting definitive empirical evidence that boundary collapse is preceded by complete internal reasoning restructuring. The higher ratio in the ROBUST model (5.96 $\times$ vs. 4.65 $\times$) reflects beneficial polarisation: non-flipped ROBUST cases are more stable ($\mu_{\text{JSD}} = 0.074$), sharpening the diagnostic contrast.

failures where it again fail, shows a less confidence fall to collapse (0.248 compared with 0.274, $\downarrow$ 9.3%), meaning the effect of vaccination works not only to prevent collapses overall, but in lessening the severity of collapses.

4.6 Multi-Dimensional Robustness Comparison

The combined effects of different kinds of perturbing a system are presented in Fig. 9 with 95% bootstrap confidence interval ($B = 2,000$). The effects on JSD reduction and reduction in confidence can be considered statistically significant ($p < 0.001$), while the flip rate reduction ($\downarrow$ 45.3%) is not statistically significant probably due to low number of flipped outputs ($N=18$ versus $N=14$). This is more a power concern than evidence of absence of an effect. In contrast, to keep $Scos$ unaffected is not statistically significant (n.s.) but is arguably the most important from a conceptual point of view of this paper and discussed in Section 6.2.

4.7 CP+XAI Symbiotic Mechanism

CP+XAI heatmap in Fig. 10 highlights a key synergistic connection. The BASE model’s increase in size of the CP set from $N = 1$ (emoji attention: 7.51%, JSD 0.03) to $N = 139$ (JSD: 0.16, confidence 0.88) shows how an increase in conformal uncertainty successfully blocks the failures of the BASE model due to dependence on emojis. In the ROBUST model, emoji attention remains near-constant across CP set sizes (7.25% vs. 7.14%), confirming that adversarial vaccination structurally eliminates the emoji–confidence dependency identified in Section 4.2.

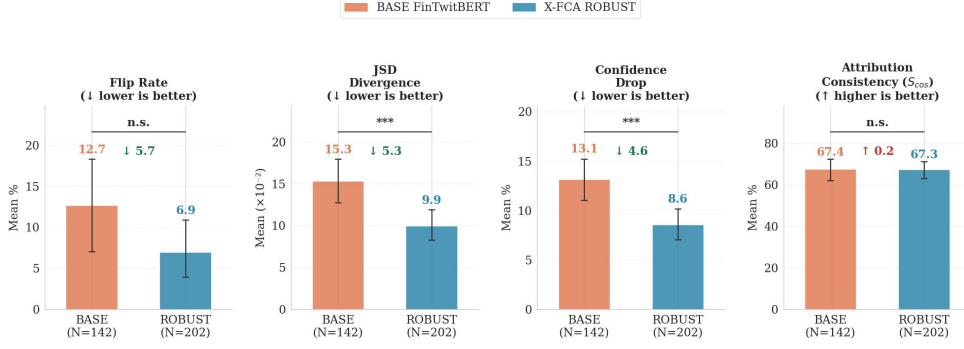


Fig. 9 Multi-Dimensional Robustness Comparison: BASE vs. ROBUST ($N_{BASE} = 142$, $N_{ROBUST} = 202$). Error bars: 95% bootstrap CI ($B = 2,000$). Significance: Mann-Whitney U (***) $p < 0.001$; n.s.). JSD (↓ 35.0%) and Confidence Drop (↓ 34.8%) are highly significantly compressed. Flip Rate (↓ 45.3%) is substantial but non-significant due to low absolute counts. S_{cos} is statistically unchanged, confirming robustness gains derive from decision boundary stabilisation.

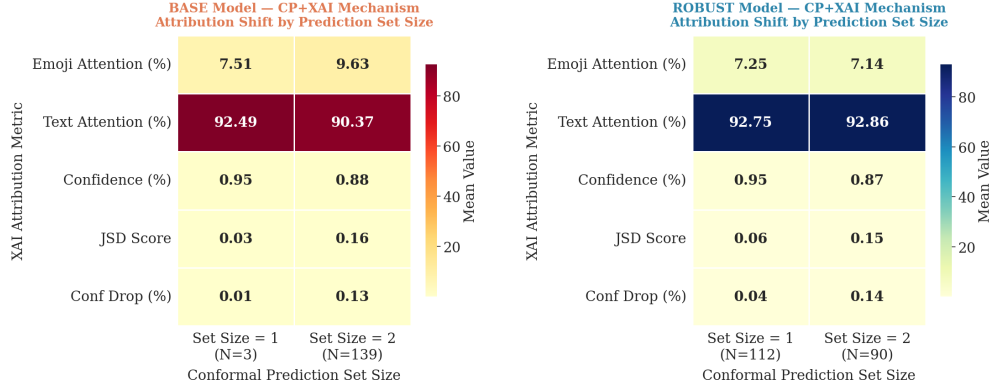


Fig. 10 X-FCA CP+XAI Composite Heatmap on Tail-Risk Subset. Metrics stratified by CP set size (Size=1: high confidence; Size=2: uncertainty zone). **BASE** (left): CP expansion is coupled to elevated emoji reliance and JSD. **ROBUST** (right): emoji attention near-constant (7.25% vs. 7.14%) across CP set sizes, confirming full decoupling of uncertainty from visual anchor reliance.

4.8 Spurious Coupling and Sarcasm Sensitivity

Spurious Coupling.

An even more convincing statistical corroboration for shortcut-learning: as illustrated in Fig. 11, the BASE-model features a strong positive correlation between emoji-attention and prediction confidence ($r = 0.121$, $p = 0.007$) - vaccination leads to full decarrelation ($r = -0.004$, $p = 0.925$, n.s.). Crucially, vaccination does not modify the emoji-attention distribution (Section 6.2) at all.

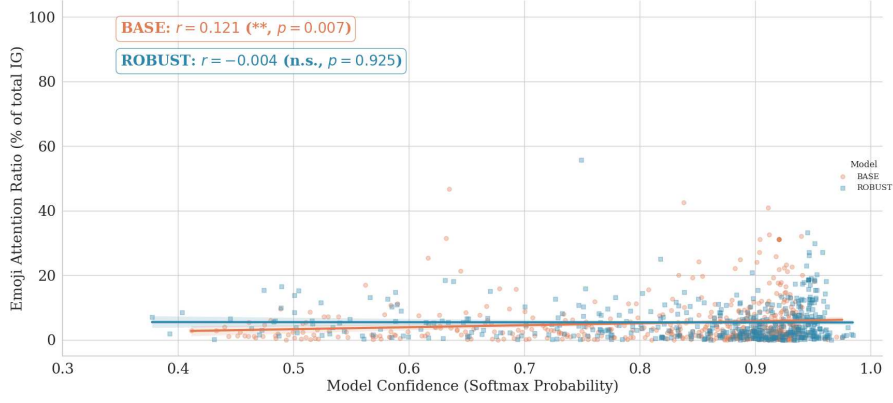


Fig. 11 Spurious Coupling: Emoji Attention vs. Prediction Confidence ($N = 500$; 250 Sarcasm, 250 Normal). **BASE**: $r = 0.121$, $p = 0.007$ (**) confirms shortcut learning. **ROBUST**: $r = -0.004$, $p = 0.925$ (n.s.) demonstrates complete decoupling (Confidence $\perp$ Emoji Attention).

Sarcasm Sensitivity.

As showed in Table 6, for balanced sub-population ($N = 500$, 250 sarcasm instances detected by S_{hype} , 250 negative controls) sarcasm leads to a $176\times$ increased prediction instability (100% flip vs. 0.57%), proving it as the strongest adversarial stressor by far. Sarcasm is unique as it has genuine semantic polarity conflict instead of token density dependence as in Threshold Effect. Not only does ROBUST detect more sarcasm traps (21 vs. 17 in BASE), but also it also reduces the gap of sarcasm vs normal attention (from +2.64% to +1.53%, $\downarrow$ 42%), meaning that vaccination increase sensitivity in contradictory signals, not uncertainty as a whole.

Table 6 Sarcasm vs. Normal: Emoji Attention Ratio Comparison (Mann-Whitney U , $N = 500$).

Model	Sarcasm (%)	Normal (%)	Δ	p -value	Sig.
BASE	6.56	3.92	+2.64	< 0.001	***
ROBUST	6.16	4.63	+1.53	0.0004	***

5 Qualitative XAI Case Studies

In this section, we move beyond macro-statistical observations and enter the micro-level mechanistic explanations. By probing individual token-level, highly adversarial sequences, we reveal the fine-grained causal mechanism of visual shortcut dependence for the BASE model and show how X-FCA vaccine enforces semantic anchors in adversarially-manipulated situations.

5.1 Case Study: The Illusion of Sarcasm Comprehension (ID: 4055)

Tweet \$EEENF my squad and I did [water wave /rocket]—chosen as the Top Attribution Emoji Case for the BASE Tail-Risk sub-example ($D_{JS} = 0.486$, emoji attention 25%)—fits the typical pattern of X-FCA failure perfectly. BASE predicts Class 1 (Neutral) with 91.5% probability and is “91.5% sure”—and as per our emoji-basedIGatt the token-level, this certainty rests solely on emoji presence and can be dissected as in Fig. 12.

Upon emoji removal (x'), the emoji cluster’s attribution support collapses and redistributes onto the \$-sign token, which alone cannot sustain the Neutral decision—triggering a Class 1 $\rightarrow$ 0 flip with a 49.6% confidence loss. In contrast, the ROBUST model processes the same sequence with 95.1% confidence but allocates only 7.9% to emoji tokens. Following removal, it maintains Class 1 (95.1% $\rightarrow$ 67.3%, stable), demonstrating that vaccination teaches the model to maintain financially grounded attention regardless of visual cue availability.

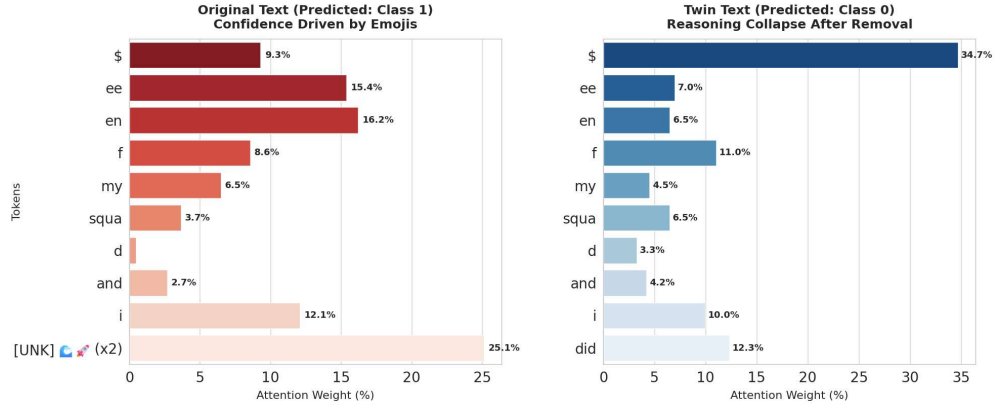


Fig. 12 XAI Diagnostic, Case ID 4055 (\$EEENF my squad and I did [water wave / rocket]). **Left** (Original, BASE Class 1, 91.5%): emoji cluster carries **25.0%** of total attribution—the highest of any feature. **Right** (Twin, BASE Class 0, 41.9%; $\downarrow 49.6\%$ confidence loss): attribution collapses onto the \$ token ($\approx 35\%$), insufficient to sustain the original prediction ($S_{cos} = 0.861$, $D_{JS} > 0.4$).

5.2 CP+XAI Symbiotic Case Study: BASE Fails, ROBUST Recovers

Tweet ID 12299 as one example among the 7 cases discovered in Section 4.5 exhibits the “Uncertain Recovery” pattern, depicted in Fig. 13. Despite a complete collapse of the emoji attention to 0%, the decision is flipped due to lack of grounding in text from the start. In the ROBUST case, however, the decision remains intact in spite of a larger absolute drop in confidence because this model had performed a large part of the attribution in the text before perturbation was applied. This increased level

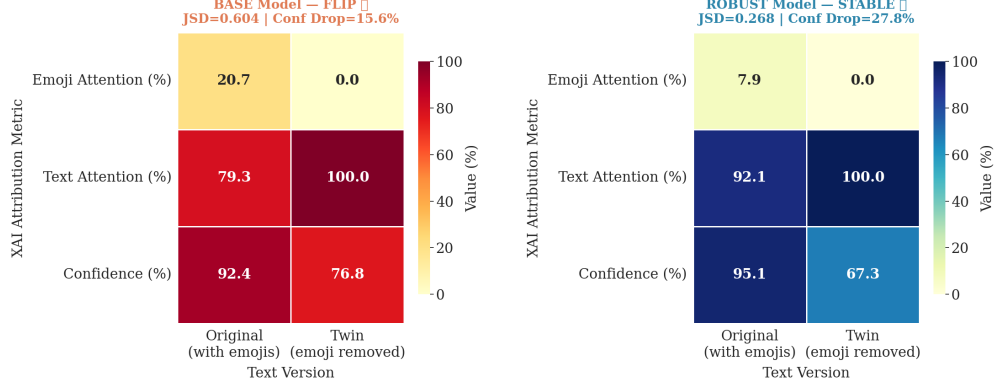


Fig. 13 CP+XAI Case Study: BASE Flips, ROBUST Stays Stable (Tweet ID 12299). **BASE** (left, FLIP): emoji attention 20.7% $\rightarrow$ 0%; decision flips (1 $\rightarrow$ 0; $D_{JS} = 0.604$, $\Delta\text{Conf} = 15.6\%$). **ROBUST** (right, STABLE): higher confidence drop ($\Delta\text{Conf} = 27.8\%$) but prediction maintained (1 $\rightarrow$ 1; $D_{JS} = 0.268$), owing to lower pre-ablation emoji reliance (7.9%) and stronger text grounding (92.1%). This exemplifies *epistemic honesty*: the ROBUST model acknowledges uncertainty without breaking its semantic bounds.

of uncertainty is epistemically *honest*, in that the model does not pretend certainty through a shortcut of the visual modality, but rather correctly identifies an ambiguity which is the intended effect of the CP layer.

5.3 The Bias-Dominance Pattern: Polarized Emoji Conflicts ($N = 17$)

The analysis of 17 BASE Tail-Risk cases ($N = 142$) presenting contradictory emoji polarity (e.g., bullish rocket alongside bearish skull) provides clear evidence of the *Bias-Dominance* property. Due to the established Tokenizer Blindness Section 3.4, the model is not capable of composing the contradictory visual cues and returns a verdict dictated by the dominant pre-training distribution—silencing the ironic language expressed by pairs of counter-poised indicators. It thus appears that the Threshold Effect (Section 4.3) is an architectural incapacity rather than a mere density artifact, thereby posing a security threat in any algorithmic trading applications susceptible to “emoji spam” type adversarial manipulations.

6 Discussion

6.1 Causal Isolation: What the Tripartite Design Proves

The $2.77\times$ differential established in Section 4.2 constitutes strong empirical evidence of a causal relationship, beyond mere correlation: model fragility is specifically attributable to visual [UNK] token density, not to sequence-length compression per se. The domain-level stratification sharpens this interpretation: Traditional Finance is approximately $37\times$ more susceptible than Corporate Finance (BASE flip rates: 2.21%

vs. 0.06%; Table 3), implying that retail-driven hype-language density is the proximate mediator of shortcut learning rather than a uniform architectural property. For automated risk systems this carries an immediate operational implication: TradFi and Crypto text streams require elevated adversarial scrutiny at ingestion.

The composition asymmetry between tail-risk subsets is a methodological *feature*: the ROBUST model’s larger pool reflects its more conservative uncertainty calibration, which flags a broader set of ambiguous predictions for downstream verification—a desirable property in safety-critical deployment. The controlled intersection analysis in Section 4.5, which confirms a 38.89% recovery rate on identical failure instances, provides the anchor that validates robustness gains as structurally genuine.

6.2 Stability Without Understanding: Explanatory Decoupling

The statistical invariance of attribution consistency (S_{cos} , n.s.; see Table 4) under adversarial vaccination constitutes the most theoretically significant finding of this study. This invariance reveals a precise behavioural paradox: X-FCA does not alter *what* the model attends to, but rather *how much it trusts* what it attends to. The spurious coupling analysis (Section 4.8) confirms this mechanistically: before vaccination, emoji attention *drives* prediction confidence ($r = 0.121$, $p = 0.007$); after vaccination, the identical attention pattern is fully decoupled from confidence ($r = -0.004$, $p = 0.925$). The model masters a state of *attention-without-anchoring*.

This finding introduces a foundational warning for the XAI literature. We term this phenomenon “*explanatory decoupling*”: robustness interventions can enforce predictive invariance without yielding more faithful internal representations. A model operating under robustness without understanding therefore remains vulnerable to out-of-distribution adversarial inputs structurally different from its training distribution (e.g., novel emoji types introduced post-vaccination). The CP layer addresses this residual risk by generating calibrated uncertainty signals precisely when internal representations are most destabilised, as demonstrated in Section 4.7. We propose that future robustness evaluations should report S_{cos} invariance explicitly as a diagnostic of whether an intervention genuinely improves semantic grounding or merely suppresses visual shortcuts at the decision layer.

6.3 The Threshold Effect: Implications for Operational Defence

The non-linear density-dependent vulnerability mentioned earlier has practical implications for our operations. By using a simple emoji count filter looking for three or more visual tokens—we can easily identify high-risk inputs for a closer look before they go into our model. This method is quick and doesn’t require us to retrain anything. More importantly, focusing on targeted vaccination with 3-cluster density twins can reduce spikes by 50% while only using 0.64% of our training data. This shows that the X-FCA method is much better at making use of data compared to random augmentation. This efficiency is crucial for deployment, especially when we have budget limits for retraining and need to maximize our training sample coverage.

6.4 Sarcasm as a Qualitatively Distinct Adversarial Stressor

The extreme instability amplification in response to sarcastic inputs (Section 4.8) indicates that sarcasm represents a qualitatively new failure mode, distinct from the Threshold Effect. Where the Threshold Effect is an artefact of high density solvable via counting filters, sarcasm exists in the realm of *semantic polarity conflict* a situation where positive visual information is pitted against negative textual information in a way that even a well-anchored system struggles to handle. Importantly, there is no easy filter solution to sarcasm; the dual performance enhancement provided by the ROBUST model (reduced attention gap *and* increased trap detection) shows that vaccination increases semantic sensitivity instead of simply reducing uncertainty.

6.5 Implications for Trustworthy Algorithmic Trading Systems

The structural fragility isolated in our tail-risk cohorts carries direct financial liability exposure. As automated trading systems increasingly ingest FinTwit streams as momentum signals, architectures over-sensitive to visual shortcuts are susceptible to deliberate “*Emoji Spam*” manipulation appending dense bullish emoji clusters to semantically neutral or bearish text to induce false positive momentum signals. The three-emoji threshold means this attack vector is practically trivial to execute at scale. Without a dedicated semantic auditing layer, such manipulation can propagate into automated execution, causing unintended capital exposure, order-book imbalances, and market liquidity shocks. The complete X-FCA pipeline combining adversarial vaccination, conformal calibration, and token-level XAI monitoring establishes a three-layer defence against this threat class.

6.6 Limitations and Future Directions

Some limitations should be mentioned. In the controlled intersection analysis, while the subset composition difference is covered, tail-risk protocols on a matched sample would deliver better head-to-head guarantees. The set of vaccinations only covers part of the entire emoji language; a methodical analysis of the OOV visual coverage should be conducted. Invariance with respect to *Scos* verifies that the explanation shapes remain the same (Section 6.2), but does not verify whether explanations are more “correct”, which can be examined by using human-annotated ground-truth attribution benchmarks in future work.

Three promising research trajectories follow directly from these boundaries:

1. **Sentiment-Aware Tokenizers:** Pre-inference tokenization schemes that encode emoji semantic polarity as first-class feature vectors, eliminating the [UNK] abstraction that is the structural root of Tokenizer Blindness.
2. **Multimodal Financial Architectures:** Extending counterfactual twin generation to benchmark models that process text alongside embedded images, chart graphics, and visual memes.
3. **Generalised Shortcut Auditing:** Deploying the tripartite counterfactual design to audit other social media shortcut phenomena such as hashtag clustering, URL density manipulation, and ticker-mention spam.

7 Conclusion

This paper introduced X-FCA (eXplainable Financial Counterfactual Analysis), a diagnostic framework coupling counterfactual perturbation, Conformal Prediction, and token-level attribution to audit the epistemic reliability of transformer-based language models deployed in financial information systems. Applied to a domain-specific financial transformer over a harmonized cross-domain corpus, the framework shows that high predictive accuracy may conceal reliance on visual shortcut representations rather than on semantic evidence. Its principal outcome is not the framework itself but the phenomenon the framework exposes: a robustness intervention removed the model’s dependence on a spurious visual cue while leaving its token-level attribution patterns statistically unchanged. We term this dissociation *explanatory decoupling*. Its consequence for reliability assessment is direct explanation stability cannot be read as evidence of semantic grounding, and a model may therefore become measurably more robust without becoming more semantically grounded.

The diagnostic pathway yielding this finding is causally and mechanistically continuous. Each of the 10 test financial emojis all paired, including opposite polarity bullish/bearish pairs are collapsed to the same [UNK] token at the tokenizer, representing a pre-inference information loss that we call tokenizer blindness. Since polarity is lost but occurrence may be preserved, the model learns to be sensitive to visual density and not semantics, resulting in the non-linear Threshold Effect by which instability increases sharply at clusters of emojis rather than linearly with the number. A deterministic ablation of emojis causes reversal far more often than an equally sized random ablation, giving causal proof that visual tokens are the primary cause of the fragility, not sequence length sensitivity. Vaccination of the model with an ad hoc set of counterfactual twins decreases the instability while de-correlating emoji attention from prediction confidence. Conformal Prediction converts these residual shortcut induced errors into calibrated uncertainty sets, providing distribution-free uncertainty estimates on top of an explanation-based auditing method. These findings indicate that robustness, calibration, and attribution are themselves independent, non-substitutable axes for model reliability. None of these three can imply the presence of another. These results also indicate the need for sentiment-aware tokenizers which treat emoji polarity as a first-class token instead of losing it at the vocabulary boundary. The domain of these results is specific to the single domain-specific model for financial microblog sentiment tested; whether the phenomenon of explanatory decoupling is applicable across tokenizer schemas, architectures, and languages is the biggest open question for future research. Instead of offering alternative modeling to Robust Model Design, X-FCA presents a stand-alone auditing layer to detect unintended shortcuts before they deploy in critical financial information systems.

References

- [1] Yahaya, O.A.: Does ceo social media activity influence stock price volatility in nigeria. *International Journal of Accounting and Finance* **24**(1), 16–44 (2026)

- [2] Garcia, J., Lu-Andrews, R.: Cultural buffers against social-media noise: religiosity and the sentiment–liquidity link. *Review of Behavioral Finance* **18**(1), 55–73 (2026)
- [3] Lawson, R., Nicole, J., Jones, F.: LLM-Enhanced NLP Pipelines for Automated Financial Document Processing and Market Signal Extraction. Available on ResearchGate (2024). <https://www.researchgate.net>
- [4] Sarder, A., Mondal, R.K.: A brief introduction to elimination of noise from big data in social media context: A review. In: *Recent Advancements in Computational Intelligence and Design Engineering*, pp. 159–176. Springer, ??? (2025)
- [5] Narayanan, P.L., Kunal, K., Ganesan, V., Ganesan, S., Jaganathan, A., Madeshwaren, V.: A review on emoji entry prediction for future finance market analysis using convolutional neural network. *International Journal of Computational and Experimental Science and Engineering* **11**(2) (2025)
- [6] Sravankumar, B., Bala Krishna, G., Vishnu Murthy, G., Ramana Kumar, D., Sridevi, M., Varaprasad Rao, M.: Integration of explainable AI with deep learning for real-time sentiment-driven stock price prediction. *Computational Economics* (2025) <https://doi.org/10.1007/s10614-025-11159-w>
- [7] Sun, Q., Rajamma, R.K., Lu, L.: Consumer confusion: examining emoji use in online communication. *Internet Research*, 1–18 (2025)
- [8] Roohi, S., Skarbez, R., Nguyen, H.D.: Reliable uncertainty estimation in emotion recognition in conversation using conformal prediction framework. *Natural Language Processing* **31**(5), 1163–1186 (2025)
- [9] Chankaev, S., Ilyushin, E.: Calibration of large language models based on the conformal prediction. *International Journal of Open Information Technologies* **13**(6), 10–15 (2025)
- [10] Ihejirika, C.J.: Explainable artificial intelligence in deep learning models for transparent decision-making in high-stakes applications. *Int J Res Publ Rev* **6**, 1924–40 (2025)
- [11] Müller, R., Schreyer, M., Sattarov, T., Borth, D.: Reshape: explaining accounting anomalies in financial statement audits by enhancing shapley additive explanations. In: *Proceedings of the Third ACM International Conference on AI in Finance*, pp. 174–182 (2022)
- [12] Djagba, P., Odinakachukwu, C.A.: Assessing the capabilities and limitations of fngpt model in financial nlp applications. *arXiv preprint arXiv:2507.08015* (2025)
- [13] Tian, G., Feng, D., Wang, R., Wen, Z.: A social financial text classification method

- based on dynamic word embeddings and semantic features fusion. *Concurrency and Computation: Practice and Experience* **37**(27–28), 70402 (2025) <https://doi.org/10.1002/cpe.70402>
- [14] Nyong, E.: Explainable ai (xai) for high-stakes decision systems. *International Journal of Technology, Management and Humanities* **11**(02), 118–129 (2025)
 - [15] Sait, A.R.W., Alkhurayyif, Y.: Hallucination-aware interpretable sentiment analysis model: A grounded approach to reliable social media content classification. *Electronics* **15**(2), 409 (2026) <https://doi.org/10.3390/electronics15020409>
 - [16] Khalil, R.: Ai driven sentiment analysis in financial markets: using transformer base models and social media signals for stock market predictions. *Journal of Modelling in Management*, 1–25 (2026)
 - [17] Du, Y., Yan, J., Chen, Y., Liu, J., Zhao, S., She, Q., Wu, H., Wang, H., Qin, B.: Less learn shortcut: Analyzing and mitigating learning of spurious feature-label correlation. In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, p. 560 (2023)
 - [18] Shaib, C., Suriyakumar, V., Wallace, B., Ghassemi, M.: Learning the wrong lessons: Syntactic-domain spurious correlations in language models. In: *Advances in Neural Information Processing Systems*, vol. 38 (2025)
 - [19] Carrillo, H.E., Pennington, R., Zhang, Y.: Is an emoji worth a thousand words? the effect of emoji usage on nonprofessional investors’ perceptions. *Journal of Information Systems* **36**(2), 1–15 (2022)
 - [20] Haase, F., Celig, T., Rath, O., Schoder, D.: Wisdom of the crowd signals: Predictive power of social media trading signals for cryptocurrencies. *Electronic Markets* **35**(1), 64 (2025)
 - [21] Gorraiz-Bengoechea, M., Gonzalez-Santocildes, A., Pastor, I., Estévez, A., Aonso-Diego, G.: Profile-level positivity assessment using multimodal feature aggregation from social media content. *IEEE Access* (2026)
 - [22] Moon, S., Okazaki, N.: Effects and mitigation of out-of-vocabulary in universal language models. *Journal of Information Processing* **29**, 490–503 (2021)
 - [23] Kürzinger, L.: Social media sentiment analysis, cryptocurrency returns and intraday bitcoin predictability. PhD thesis, Dissertation, Düsseldorf, Heinrich-Heine-Universität, 2024 (2024)
 - [24] Zimmerman, J.W., Hudon, D., Cramer, K., Ruiz, A.J., Beauregard, C., Fehr, A., Fudolig, M.I., Demarest, B., Bird, Y.M., Trujillo, M.Z., et al.: Tokens, the oft-overlooked appetizer: Large language models, the distributional hypothesis, and meaning. *arXiv preprint arXiv:2412.10924* (2024)

- [25] Chan, D.M., Corona, R., Park, J., Cho, C.J., Bai, Y., Darrell, T.: Analyzing the language of visual tokens. arXiv preprint arXiv:2411.05001 (2024)
- [26] Pattnaik, A.: Neuro-symbolic ai: Bridging reasoning and learning. *Next-Gen Machine Learning: Algorithms for Adaptive Intelligence*, 107 (2025)
- [27] Gunonu, S., Altun, G., Cavus, M.: Explainable bank failure prediction models: Counterfactual explanations to reduce the failure risk. *Computational Economics* (2026) <https://doi.org/10.1007/s10614-026-11353-4>
- [28] Zeng, Y., Liu, J., Lam, H., Namkoong, H.: Llm embeddings improve test-time adaptation to tabular $y|x$ -shifts. arXiv preprint arXiv:2410.07395 (2024)
- [29] Smith, A., Martinez, W., Garcia, S., Thomas, B., Davis, O., Ye, W.: Understanding distribution shift in llms: Methods, evaluations, and challenges. Preprint on ResearchGate (02 2024). doi **10** (2024)
- [30] Gona, N.S.R., Aunugu, D.R., Methuku, V., Agrawal, M., Myakala, P.K.: Hybrid lstm-transformer model for stock market prediction: A deep learning approach. In: 2025 IEEE International Conference on Artificial Intelligence Testing (AITest), pp. 87–93 (2025). IEEE
- [31] Kim, J., Kim, H.-S., Choi, S.-Y.: Forecasting the s&p 500 index using mathematical-based sentiment analysis and deep learning models: A finbert transformer model and lstm. *Axioms* **12**(9), 835 (2023)
- [32] Mahdi, M.A., Fati, S.M., Ragab, M.G., Hazber, M.A., Ahamad, S., Saad, S.A., Al-Shalabi, M.: A novel hybrid attention-based roberta-bilstm model for cyberbullying detection. *Mathematical and Computational Applications* **30**(4), 91 (2025)
- [33] Chantrasmee, C., Jaiyen, S., Chaikhan, S., Wattanakitrungrroj, N.: Stock trading signal prediction using transformer model and multiple indicators. In: 2024 28th International Computer Science and Engineering Conference (ICSEC), pp. 1–6 (2024). IEEE
- [34] Rizinski, M., Peshov, H., Mishev, K., Jovanovik, M., Trajanov, D.: Sentiment analysis in finance: From transformers back to eXplainable lexicons (XLex). *IEEE Access* **12** (2024) <https://doi.org/10.1109/ACCESS.2024.3349970>
- [35] Mishra, S., Dutta, S., Long, J., Magazzeni, D.: A Survey on the Robustness of Feature Importance and Counterfactual Explanations. Workshop on Explainable AI in Finance (XAI-FIN), JPMorgan Chase & Co. (2023)
- [36] Duell, J., Seisenberger, M., Fu, H., Fan, X.: Quce: The minimisation and quantification of path-based uncertainty for generative counterfactual explanations. In: 2024 IEEE International Conference on Data Mining (ICDM), pp. 693–698

- (2024). <https://doi.org/10.1109/ICDM59182.2024.00078>
- [37] Campos, M., Farinhas, A., Zerva, C., Figueiredo, M.A.T., Martins, A.F.T.: Conformal prediction for natural language processing: A survey. Transactions of the Association for Computational Linguistics **12**, 1497–1516 (2024) https://doi.org/10.1162/tacl.a_00715
 - [38] Giovannotti, P.: Conformal prediction methods for reliable language processing. PhD thesis, Royal Holloway, University of London (2026)
 - [39] Vellore, A., Jha, N.K.: Uncertainty-aware transformers: Conformal prediction for language models. arXiv preprint arXiv:2604.08885 (2026)
 - [40] Ayatollahi, M., Jafari, S.M.: Robustness vs. explainability stability: A comparative XAI analysis of machine learning models in financial market anomalies. Computational Economics (2026) <https://doi.org/10.1007/s10614-025-11307-2>