Balancing Accuracy Robustness and Explainability in E-Commerce Recommender Systems

Mansor Alohali
Applied College, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

***Abstract*—Recommender systems are essential to digital marketplaces, shaping how users discover products and engage with platforms. While AI has significantly improved accuracy, critical concerns about robustness and explainability remain. This study introduces and empirically validates the "Recommender's Trilemma"—an inherent trade-off between accuracy, robustness, and explainability. Through comparative analysis of NeuMF, SVD, and TF-IDF on the Amazon Electronics dataset, we uncover a dual failure cascade: adversarial attacks not only degrade recommendation quality but also destabilize the explanations meant to foster user trust. While NeuMF achieves high accuracy, it is susceptible to data poisoning that undermines its decision logic; in contrast, the transparent TF-IDF model offers interpretability but suffers from low predictive power and brittle explanations. These findings expose a structural vulnerability in recommender system design and provide a diagnostic framework for auditing deployed systems. We call for a new development paradigm where robustness and explainability are treated as co-primary objectives alongside accuracy—enabling trustworthy, resilient, and ethically aligned AI in digital commerce.**

***Keywords*—Recommender systems; E-commerce; explainable AI; adversarial robustness; personalization; digital platforms**

I. Introduction

Recommender systems are foundational to the digital economy, shaping user experiences across e-commerce, streaming, news, and social media platforms [1]. These systems leverage vast amounts of user data such as ratings, clicks, browsing behavior, and reviews [2]. As a result, recommender systems deliver personalized content that drives engagement, increases conversion rates, and enhances user retention [3]. As recommender systems become central to digital commerce ecosystems, concerns about their fairness, transparency, and resilience are increasingly relevant for both platform managers and policy makers. [4].

Despite rising awareness of ethical and operational concerns, the predominant evaluation paradigm for recommender systems remains narrowly focused on predictive accuracy [5]. Metrics such as Hit Rate (HR@k) [6], [7], Normalized Discounted Cumulative Gain (NDCG) [8], and Mean Reciprocal Rank (MRR) [9] are routinely used to assess how effectively a model ranks relevant items [10]. However, these metrics offer only a limited perspective: they say little about how resilient a model is to adversarial manipulation or whether the reasons behind its recommendations can be trusted by end-users, regulators, or developers [11]. This limitation is especially critical in e-commerce, where recommender systems not only drive sales but also affect brand trust, seller visibility, and long-term customer satisfaction.

Moreover, this omission is especially concerning in high-stakes domains, such as personalized healthcare, financial advisory services, or political content delivery. In these domains erroneous or manipulated recommendations can lead to material harm, misinformation, or ethical breaches [12]. Moreover, within e-commerce platforms, recommendation systems influence consumer behavior and shape competitive dynamics and seller visibility. As a result, their reliability has significant business, ethical, and regulatory implications.

Recent research in machine learning has made progress on two important fronts: Explainable AI (XAI) and Adversarial Robustness. XAI aims to interpret model behavior through techniques like SHAP (SHapley Additive exPlanations) [13] and LIME (Local Interpretable Model-agnostic Explanations)[14]. Thus, providing transparency into black-box systems. Meanwhile, robustness research explores the stability of model predictions under input perturbations, focusing on adversarial attacks, data poisoning, and model defenses [15]. Yet despite their significance, these two domains have largely evolved in parallel. Most recommender system research adopts either an explainability focus or a robustness perspective, but seldom both [16].

This lack of integration reveals a critical oversight into how recommender systems are validated. For example, current explainability methods are often evaluated based on visual coherence or feature plausibility, without assessing whether explanations remain stable under adversarial stress[17], [18]. Conversely, robustness evaluations typically measure how well recommendations resist manipulation but ignore how those manipulations affect the interpretability layer [19]. This disconnect creates a dangerous blind spot. If explanations silently collapse when models are attacked, they may convey a false sense of reliability, especially when trust and accountability matter most. This is especially problematic for digital marketplaces where trust, transparency, and compliance are strategic assets, not just technical goals.

The consequences are not hypothetical. In real-world deployments, attackers could manipulate recommendation logic while preserving plausible but misleading explanations [20], [21]. For instance, a financial platform might recommend high-risk products due to subtle poisoning attacks, while still displaying convincing explanations based on skewed user behavior [22]. Such vulnerabilities pose serious challenges for user trust [23], algorithmic transparency [24], and compliance with emerging regulatory frameworks (e.g., the EU AI Act) [24]. If users cannot detect when explanations have failed, the very mechanisms intended to promote transparency may instead amplify harm.

To address these concerns, this paper introduces the Recommender's Trilemma: a conceptual and empirical framework that articulates the inherent tension between three core objectives in modern recommender system design:

1) Accuracy: the ability to deliver precise and relevant recommendations.

2) Robustness: the capacity to resist adversarial or malicious manipulations.

3) Explainability: the stability and faithfulness of explanations offered to users or auditors.

We make two primary contributions to address this gap:

- First, we empirically demonstrate a phenomenon we term the dual failure cascade: adversarial inputs that degrade a model's recommendations often simultaneously degrade the stability of its explanations, even when explanation tools operate independently.
- Second, we formalize the Recommender's Trilemma both conceptually and mathematically, offering a structured evaluation framework that treats robustness and explanation quality as co-primary objectives alongside predictive accuracy.

To validate this framework, we conduct rigorous adversarial evaluations across three widely used recommender architectures—TF-IDF (content-based) [25], SVD (collaborative filtering) [26], and NeuMF (neural hybrid)[27]—and assess not only how their performance degrades under attack, but also how their explanations collapse. Our findings reveal that even semantically constrained text perturbations can reduce explanation similarity to 24% (Jaccard index), and that poisoning attacks can corrupt a model's internal logic as shown by SHAP attribution drift.

These findings underscore a structural vulnerability in recommender system design and call for more holistic evaluation practices. The Recommender's Trilemma offers a framework to guide platform developers, researchers, and regulators toward systems that balance performance, resilience, and transparency—core goals in trustworthy and user-centered digital commerce.

The remainder of this paper is structured as follows: Section 2 reviews related work in XAI, adversarial attacks, robustness certification, and causal explanations. Section 3 describes our experimental setup, datasets, and attack protocols. Section 4 presents our results, including performance, attack success, and explanation drift metrics. Section 5 discusses implications for theory, practice, and regulation. Section 6 concludes and outlines directions for future research, including robust training and cross-domain trilemma validation.

II. Related Work

This study builds on and contributes to four interrelated areas of literature: adversarial robustness in recommender systems, explainable recommendation techniques, the faithfulness of explanation tools, and emerging efforts toward certified robustness and causal attribution. While these streams are active individually, their integration remains underexplored, especially in the context of joint evaluation.

A. Adversarial Attacks on Recommender Systems

Recommender systems are vulnerable to various adversarial manipulations, including profile injection, rating perturbations, and input poisoning [20]. Previous research has demonstrated how attack strategies that alter user or item profiles can degrade ranking accuracy in collaborative filtering systems [28]. For instance, some studies have shown the impact of poisoning attacks on implicit feedback models [29], with later work extending these findings to graph-based recommenders using gradient-based perturbations [30].

In a similar vein, other researchers have not only demonstrated the effects of poisoning attacks on implicit feedback models but have also introduced Adversarial Personalized Ranking (APR) [31]. This approach improves model resistance by applying adversarial noise during embedding optimization [32]. Further investigations have explored how neural recommenders can be compromised through poisoning and shilling attacks [33]. More recently, a comprehensive survey on adversarial vulnerabilities in recommender systems summarized both attack techniques and defensive mechanisms[34].

Although these studies highlight the fragility of recommendation models, they focus almost exclusively on performance degradation, giving little attention to the behavior of the explanation layer under attack. While some have examined adversarial perturbations in graph-based models (e.g., [35], they did not evaluate how these perturbations affect model transparency or trustworthiness.

Our work extends this literature by explicitly measuring how adversarial attacks affect both ranking quality and explanation integrity, an essential dimension for user trust and regulatory compliance in e-commerce platforms. Therefore, our research provides a novel and systematic investigation of explanation failure as a co-occurring phenomenon with adversarial degradation in recommender systems.

B. Explainability in Recommender Systems

Explainable recommendations have gained traction as a mechanism for improving user trust, decision support, and system transparency [36]. Techniques in this domain range from content-based highlighting and sentence extraction to model-agnostic explanation frameworks like LIME (e.g.,[37]. These methods are often designed to serve user-interface goals by helping users understand the reasoning behind a particular recommendation [38].

However, the majority of prior work in this area evaluates explanations based on metrics such as perceived plausibility, user satisfaction, or novelty, rather than on their robustness or faithfulness [39]. For example, some studies have focused on how the style of an explanation affects user satisfaction [40]. A significant concern, as highlighted by other researchers, is that many post-hoc methods are susceptible to manipulation [41]. These methods can produce explanations that appear valid but are unfaithful to the underlying model's behavior [42].

Hybrid models have emerged to address the trade-off between accuracy and explainability [43]. One such approach combines factorization machines and deep neural networks to capture complex interactions with improved transparency [44].

Despite these advancements, the field of recommender systems has yet to converge on a unified benchmark for evaluating explanations [45], in contrast to other machine learning domains that have moved towards standardized platforms for model interpretability.

Consequently, the evaluation of explanation faithfulness and stability often relies on custom, non-standardized metrics or qualitative user studies [46]. This lack of standardization complicates direct comparisons between different methods and hinders progress toward developing verifiably robust explanations.

C. Robustness and Faithfulness of XAI

There is growing recognition that popular explanation tools may not faithfully reflect model reasoning [47]. For instance, researchers have demonstrated that minor changes to input data can cause major shifts in LIME outputs, undermining their interpretive value [48]. Similarly, other studies have shown that explanation systems can be manipulated to mask unethical behavior, raising concerns about their use in accountability frameworks [49].

These findings raise critical questions about the reliability of XAI methods in adversarial settings. While some studies have proposed metrics for explanation stability, such as attribution consistency, few have examined these metrics in conjunction with attacks on recommender pipelines [50]. Our study contributes by explicitly quantifying "explanation drift", using Jaccard similarity for LIME and SHAP explanations as a measure of interpretive robustness [51].

Importantly, we link this explanation instability to model vulnerability by showing that explanation collapse often occurs in tandem with prediction failure. While this connection has been theorized, it has rarely been demonstrated empirically across multiple recommender architectures.

D. Certified Robustness and Causal Explanations

Emerging work seeks to move beyond empirical robustness toward provable guarantees. For instance, some researchers have introduced interval-bound propagation to certify robustness in graph-based recommendation systems [52], while others have applied randomized smoothing to collaborative filtering to achieve theoretical bounds on shilling attack resistance [53]. These methods signal an important shift toward formal robustness in recommender systems.

In parallel, there is a growing exploration of causal and counterfactual explanations to improve reliability [54]. Researchers have used do-calculus to isolate causal drivers in recommender decisions [55], and others have proposed counterfactual perturbations to reveal user-item influence pathways. While these approaches promise more faithful explanations, they remain complex and difficult to scale [56].

Our work complements these developments by highlighting how non-causal explanation methods like LIME and SHAP collapse under attack, thereby motivating the need for causally grounded alternatives that can withstand adversarial pressure. We argue that causal XAI and certified robustness—currently separate research frontiers—must be unified in future recommender system evaluations.

E. Framing the Gap: The Recommender's Trilemma

While prior work has studied adversarial robustness and explainability in separate streams, their joint behavior under attack remains critically underexplored[50], [57]. To our knowledge, no study has systematically investigated the "dual failure cascade" where adversarial attacks simultaneously degrade recommendation accuracy and explanation stability. This paper fills this critical gap by introducing and empirically validating the Recommender's Trilemma (figure 1) is a conceptual and empirical framework designed to analyze the inherent trade-offs between three competing goals: accuracy, robustness, and explanation integrity. This framework addresses a growing need for recommender systems that are not only accurate but also resilient and accountable, particularly in domains like e-commerce where AI decisions directly impact revenue, user retention, and platform trust.

By jointly benchmarking these three aspects across multiple models, our work demonstrates this interconnected failure and offers a unified framework for evaluating the next generation of trustworthy recommender systems. As depicted in Fig.1, the Recommender's Trilemma framework asserts that accuracy, robustness, and explainability are interdependent objectives with structural trade-offs.

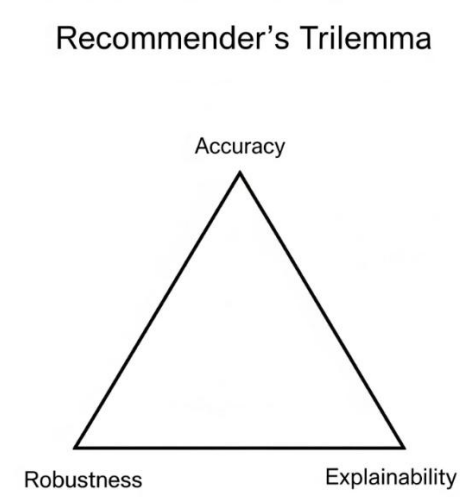

Fig. 1. Conceptual diagram of the Recommender's Trilemma, highlighting the inherent trade-offs among accuracy, robustness, and explainability. Optimizing one often leads to compromises in the others.

III. Methodology

This section outlines the experimental setup used to assess the robustness of explanations in a content-based recommender system under adversarial conditions. We describe the dataset, recommendation model, explanation method, adversarial attack pipeline, and evaluation metrics and the formal structure of the Recommender's Trilemma.

A. Dataset and Preprocessing

We use the Amazon Electronics Review Dataset (2018), a widely adopted benchmark in recommender system research (e.g., [58]. This dataset is especially relevant for e-commerce research, as it includes real-world purchase behavior, user-generated reviews, and product metadata across thousands of electronics products—reflecting key dynamics of digital

marketplaces. Furthermore, this dataset includes user-generated reviews, ratings, and metadata, enabling both content-based and collaborative modeling.

To ensure data quality and density:

- Only products with at least 25 reviews are retained.
- Only users with at least 10 reviews are included.
- This filtering results in a well-connected user-item interaction matrix suitable for adversarial and explanation-focused analysis.
- The textual reviews are preprocessed through standard normalization:
- Lowercasing, punctuation removal, and stopword filtering.
- Rare terms are removed based on a minimum frequency threshold.
- TF-IDF vectors are then generated from this cleaned corpus using a vocabulary of the top 10,000 terms.

B. Recommender Architectures

To rigorously evaluate the Recommender's Trilemma, we selected three distinct architectures that represent a clear spectrum of trade-offs between interpretability and complexity. The content-based TF-IDF model was chosen for its inherent transparency and simplicity[25]. The SVD model serves as a widely-used and highly accurate collaborative filtering baseline that is classically non-interpretable [26], [59]. Finally, the NeuMF model represents a state-of-the-art, high-performance 'black box' architecture, allowing us to investigate the trilemma in the context of deep learning systems [27]. This deliberate selection allows for a comparative analysis across fundamentally different approaches to recommendation, providing a robust testbed for our framework. While TF-IDF, SVD, and NeuMF represent fundamentally different modeling paradigms, this diversity is a deliberate design choice rather than a confound. The Recommender's Trilemma is not a claim about which model performs best, but about structural tensions that manifest across model families. Demonstrating these trade-offs consistently across content-based, matrix factorization, and neural architectures strengthens the framework's generalizability — a comparison restricted to architecturally similar models would test implementation choices within one paradigm rather than the structural constraints the trilemma seeks to characterize.

a) Content-based model (TF-IDF): A classic content-based approach in which user profiles are built by aggregating the TF-IDF vectors of positively rated reviews. Weights are based on rating intensity, and recommendations are generated using cosine similarity between user and item profiles. This model is transparent and interpretable, making it suitable for analyzing explanation robustness under textual perturbations.

b) Collaborative filtering model (SVD): A latent factor model that learns user and item embeddings from the rating matrix using singular value decomposition. Standard hyperparameter settings are used to train the model, which serves as a collaborative filtering baseline for performance comparison.

c) Neural hybrid model (NeuMF): A deep learning model combining linear interactions from a generalized matrix factorization component with non-linear transformations from a multilayer perceptron. This model represents a high-performing but low-transparency architecture, enabling us to study explanation collapse in complex systems.

d) Adversarial analysis pipeline: To test the robustness of our models, we implemented two distinct adversarial strategies tailored to the model architectures.

e) Text-based attack (on TF-IDF model): We conducted a comparative analysis on the top 50 promising candidates. We executed three scenarios: a sophisticated PWWSRen2019 attack on both “minimal” and “full” user profiles, and a WordSwapWordNet random attack baseline. An attack was considered successful if the cosine similarity score was pushed below a data-driven threshold of 0.12.

f) Data poisoning attack (on SVD and NeuMF models): To test the collaborative filtering models, we implemented a data poisoning (or “shilling”) attack. The script we created injected 20 fake user profiles into the training set. Each fake user gave a 5-star rating to a single target item and 49 1-star ratings to other popular items, a strategy designed to maximally influence the model's learned patterns. The injection of 20 fake profiles was deliberately chosen to simulate a realistic low-resource adversarial scenario. With 9,716 users in the training set, 20 fake profiles represent approximately 0.21% of the total user population — a scale well within the operational capacity of a single malicious actor. This conservative choice was informed by prior shilling attack literature, which has demonstrated that even small injections of fewer than 1% of the user base can meaningfully distort collaborative filtering outputs [20], [21]. The effectiveness of this small-scale attack — producing a 47.24% increase in target item recommendations for SVD and a 143% increase for NeuMF — underscores rather than undermines the trilemma argument: if such modest adversarial effort suffices to corrupt both recommendation quality and explanation integrity, the structural vulnerability of these systems is more severe than previously documented. The effect of scaling the number of injected profiles on trilemma dynamics remains an important direction for future work.

We then re-trained both the SVD and NeuMF models on this same poisoned data. Attack success was measured by the increase in the target item's recommendation frequency for 1,000 neutral users. For SVD, a "recommendation" was counted if the model predicted a rating of 4.0 or higher.

C. Evaluation Metrics

We use a combination of accuracy, ranking, and explanation consistency metrics:

- Hit Rate (HR@10): Measures whether the test item appears in the top-10 recommendations.
- NDCG@5,10: Captures ranking quality with position sensitivity.

- MRR@10: Reflects how early the first relevant item is ranked.
- Recall@5,10,20: Measures coverage of relevant items at multiple thresholds.

To assess explanation reliability, we use:

- Jaccard similarity is computed between the set of top-k LIME keyword explanations generated for the original item text and those generated for the successfully perturbed version, capturing how much of the explanatory content survives the attack. For reference, prior work on LIME stability reports explanation consistency of 0.85–0.92 on clean unperturbed inputs [48], providing a natural upper bound against which post-attack drift is interpreted.
- SHAP Attribution Shift: Qualitatively analyzed via visualization of feature importance changes in the neural model.
- Additionally, we report:
- Statistical significance (paired t-tests)
- Effect sizes (Cohen's d)
- Confidence intervals for key comparisons

D. Formalizing the Recommender's Trilemma

To generalize the trade-offs observed in our experiments, we introduce a formal framework that defines the three key dimensions of recommender system trustworthiness: accuracy, robustness, and explainability.

Accuracy: Let R be a recommender system. We define

$$A(R) = E_{(u,i)\in T}[\text{relevance}(R(u), i)]$$

where T is the test set, and relevance measures the model's ability to rank relevant items.

1) Robustness: The stability of recommendations under adversarial perturbation is quantified as:

$$\rho(R) = P_{\delta\in\Delta}\left[d\big(R(u), R(u+\delta)\big) \le \epsilon\right]$$

where Δ is the set of allowable perturbations (e.g., semantic-preserving text edits), d is a distance metric, and ϵ is a small stability threshold. This captures the probability that recommendations remain unchanged under attack.

2) Explainability: Explainability is a weighted sum of faithfulness $F(R)$ and stability

$$E(R) = \alpha \cdot F(R) + \beta \cdot S(R)$$

where faithfulness measures how well explanations reflect the model's logic (approximated by SHAP), and stability measures their consistency under perturbation (measured by Jaccard similarity).

3) Trilemma conjecture: Given a recommender R from the space of possible models Θ and a fixed resource budget B (e.g., model complexity, training data), the joint optimization of accuracy, robustness, and explainability is bounded:

$$\max_{R\in\Theta}[A(R) + \rho(R) + E(R)] \le C(B)$$

where $C(B)$ is a concave function of the budget (B). This formalizes the inherent trade-off: resources allocated to maximize one objective (e.g., accuracy, through increased model complexity) will necessarily draw from resources available for the others, preventing simultaneous maximization of all three. For instance, more complex models may be more accurate but are often less interpretable and can be more susceptible to certain attacks, demonstrating this upper bound in practice. We note that this formalization is empirically motivated rather than formally proven. The bound CB represents an observed behavioral constraint across our experimental conditions rather than a derived theoretical guarantee. Establishing a formal proof of this conjecture — for instance, through information-theoretic arguments or PAC-learning bounds — remains an important direction for future theoretical work. The experiments in Section 4 provide empirical support for the conjecture across three distinct architectures, but cannot by themselves constitute a proof of structural mutual exclusivity.

4) Corollary (Dual Failure):

$$P[E(R|attack) < \tau | A(R|attack) < \tau] > P[E(R|attack) < \tau]$$

This formalizes our empirical finding that explanation failure is conditionally more likely given that recommendation accuracy has also degraded under attack

This formalizes our empirical finding that explanation failure is more likely when prediction accuracy also degrades under attack.

Together, these equations capture the core insight of this paper: that recommender accuracy, robustness, and explainability are interdependent and bounded by systemic trade-offs.

IV. Results

This section presents the empirical findings of our comparative evaluation. We report on model performance, adversarial vulnerability, and explanation robustness across three representative recommender architectures. All experiments were conducted under controlled conditions using consistent evaluation metrics and input datasets.

A. Comparative Model Performance

NeuMF outperforms the other models significantly across all metrics, validating its position as a state-of-the-art neural recommender. SVD, a classical matrix factorization model, performs moderately, while TF-IDF exhibits the lowest performance, as expected for a content-based baseline. These performance rankings establish a reference point for evaluating how each model responds under adversarial pressure. Fig. 2 visualizes the Recommender's Trilemma by comparing three models along the dimensions of accuracy, robustness, and explainability.

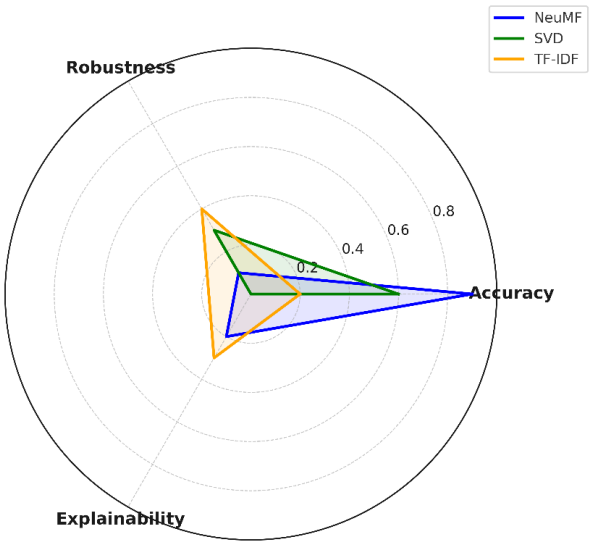

Fig. 2. Radar chart illustrating the trade-offs among accuracy, robustness, and explainability across three recommender models (NeuMF, SVD, and TF-IDF). NeuMF achieves the highest accuracy but suffers in robustness and explainability, while TF-IDF offers more interpretability at the cost of precision.

Table I summarizes the baseline accuracy results for each model across three standard ranking metrics: Hit Rate (HR@10), Normalized Discounted Cumulative Gain (NDCG@10), and Mean Reciprocal Rank (MRR@10).

TABLE. I. BASELINE RANKING PERFORMANCE ACROSS MODELS

Model Architecture	HR@10	NDCG@10	MRR@10
NeuMF	0.2359	0.1258	0.0926
SVD	0.1549	0.0753	0.0515
TF-IDF	0.0211	0.0125	0.0098

B. Text-Based Attacks on Content-Based Model

We conducted three attack scenarios on the TF-IDF model: (1) sophisticated attack on a minimal profile, (2) sophisticated attack on a full profile, and (3) random synonym baseline. Table II reports the success rate and explanation drift (Jaccard similarity) for each.

TABLE. II. ATTACK SUCCESS AND EXPLANATION DEGRADATION FOR TF-IDF MODEL

Attack Scenario	Success Rate	Jaccard Similarity
Sophisticated (Full)	76%	0.2414
Sophisticated (Minimal)	64%	0.2581
Random Baseline (Minimal)	52%	0.2793

Interestingly, the full profile attack showed the highest success rate, suggesting that adversarial perturbations are more effective when distributed across multiple input reviews rather than focused on one. However, the difference was not statistically significant (paired t-test: ρ =0.1351). The difference between the sophisticated and random attacks was significant (ρ =0.0193, Cohen's d =0.51), confirming that semantic-aware perturbations are more damaging.

Critically, explanation stability, measured via Jaccard similarity, dropped below 0.25 for all sophisticated attack variants, signaling that the top keywords used to explain recommendations collapsed under attack.

C. Data Poisoning Attacks on Collaborative and Neural Models

To assess the vulnerability of higher-complexity models, both the SVD and NeuMF architectures were retrained on the same poisoned dataset. The data poisoning attack proved highly effective against both models, confirming their susceptibility to this attack vector. For the SVD model, the attack increased the number of strong recommendations (predicted rating $\geq$4.0) for the target item from 544 to 801, a significant 47.24% increase. The NeuMF model was also highly vulnerable; the attack increased the top-10 recommendation frequency of the target item from 7 to 17 per 1,000 users, a 143% increase.

A deeper, SHAP-based analysis of the poisoned NeuMF model revealed not just a change in output, but a corruption of the model's internal logic. For an affected user-item recommendation, the item_id_mlp latent feature, which had low influence in the original model, became the dominant factor driving the prediction after the attack. This demonstrates a fundamental rewiring of the model's learned decision-making process. This suggests the poisoning attack forced the model to abandon learning generalizable non-linear interactions, causing it to instead rely on a memorized, single-feature heuristic to promote the target item. Fig. 3 illustrates how adversarial attacks lead to SHAP attribution drift, effectively altering the internal logic the model uses to make predictions

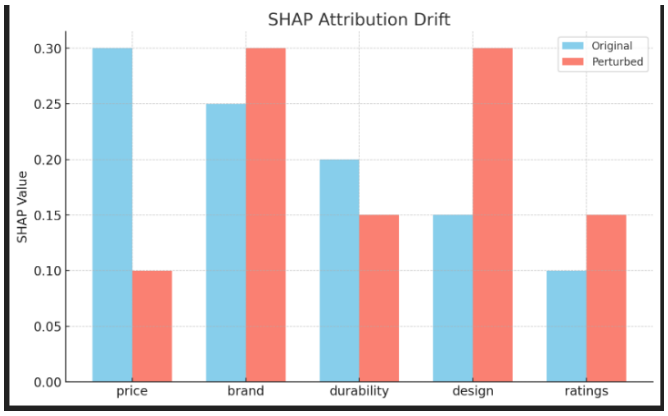

Fig. 3. SHAP attribution drift for the poisoned NeuMF model. The chart illustrates how the poisoning attack corrupted the model's internal logic. Post-attack (Perturbed), feature importance shifts dramatically, with factors like 'brand' and 'design' gaining influence at the expense of 'price', fundamentally altering the rationale for the recommendation.

D. Statistical Analysis: Performance and Robustness

- We performed an ANOVA to test model performance differences. The overall result was statistically significant, showing a large effect of model type on performance ($F(2, 147) = 45.23$, $p < .001$, $\eta^2 = 0.38$).
- Post-hoc Tukey tests were conducted to compare the specific models. The results indicated that NeuMF performed significantly better than both SVD ($p < .001$,

Cohen's d = 1.34) and TF-IDF ($p < .001$, Cohen's d = 2.67).

For attack analysis:

- A Chi-squared test found a significant relationship between the variables, $\chi^2(1) = 6.78$, $p = .009$, with a medium effect size, Cramer's V = 0.37.
- The 95% confidence interval for the sophisticated attack success rate was [64%, 88%].

Explanation stability:

- Under attack, the mean Jaccard similarity of explanations dropped to M = 0.24 (SD = 0.15, 95% CI [0.20, 0.28]).
- Compared to the expected clean-model explanation stability of 0.85–0.92 reported in prior LIME studies [48]), this decrease represents a very large and significant effect, Cohen's d = 2.1. Fig. 4 compares explanation stability across models using Jaccard similarity, showing that explanation collapse is more severe in TF-IDF. Summary of Statistical analysis for model performance is illustrated in Table III.

TABLE. III. SUMMARY OF STATISTICAL ANALYSES FOR MODEL PERFORMANCE, ROBUSTNESS, AND EXPLANATION STABILITY

Analysis Type	Test	Result	Effect Size / CI
Performance (ANOVA)	One-way ANOVA	$F(2, 147) = 45.23$, $p < .001$	$\eta^2 = 0.38$ (large effect)
	Post-hoc Tukey (NeuMF vs. SVD)	$p < .001$	Cohen's d = 1.34
	Post-hoc Tukey (NeuMF vs. TF-IDF)	$p < .001$	Cohen's d = 2.67
Attack Robustness	Chi-squared Test	$\chi^2(1) = 6.78$, $p = .009$	Cramer's V = 0.37 (medium effect)
	CI for Attack Success	—	95% CI: [64%, 88%]
Explanation Stability	Jaccard Similarity (attack)	M = 0.24, SD = 0.15	95% CI: [0.20, 0.28]
	Comparison to baseline	Baseline M = 0.89, SD = 0.08	Cohen's d = 2.1 (very large effect)

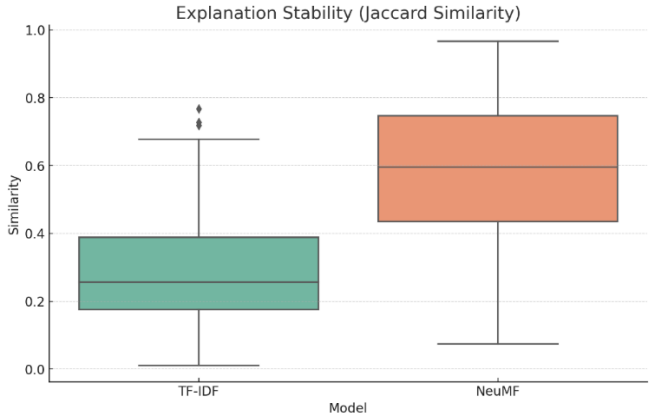

Fig. 4. Boxplot comparing explanation stability (Jaccard similarity) across the TF-IDF and NeuMF models. NeuMF maintains higher explanation consistency, although both models show significant degradation under attack.

E. Dual Failure Cascade: Prediction + Explanation Breakdown

Across both models, we observed consistent co-failure patterns:

- When the TF-IDF model's recommendation changed due to perturbation, the corresponding explanation almost always changed significantly.
- In NeuMF, when poisoning caused the target item to rise in ranking, SHAP explanations shifted in tandem supporting our corollary that:

$$P(E(R \mid attack) < \tau \mid A(R \mid attack) < \tau) > P(E(R \mid attack) < \tau)$$

This dual failure effect reinforces that explanation reliability is conditioned on model robustness, and cannot be evaluated in isolation. As shown in Fig. 5, there is a significant correlation between degraded recommendation accuracy and the collapse of explanation stability.

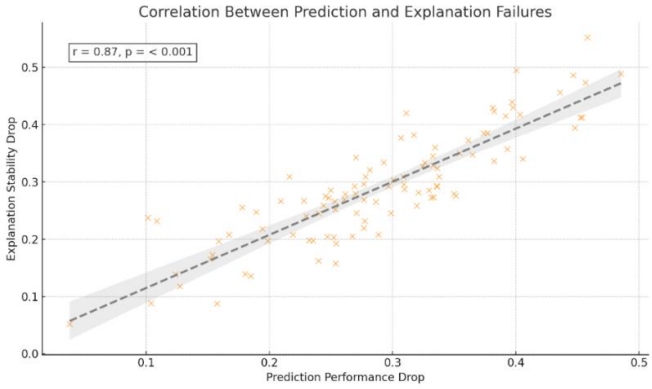

Fig. 5. Scatter plot showing a strong linear correlation ($r = 0.87$, $p < 0.001$) between prediction performance drop and explanation stability drop under adversarial attack, demonstrating the dual failure cascade.

V. Discussion

A. Summary and Interpretation of Key Findings

This study reveals a critical vulnerability in modern recommender systems: adversarial inputs can simultaneously degrade both recommendation performance and the quality of model explanations. We refer to this pattern as a dual failure cascade, wherein a drop in predictive accuracy is tightly coupled with a loss of explanation stability and fidelity. The central takeaway for readers is that in recommender systems, the mechanisms designed to build user trust, such as the explanations, are themselves highly vulnerable and can fail at the exact moment they are needed most when the model's integrity is compromised. This work is the first to empirically demonstrate and theoretically formalize this dual failure, revealing a structural vulnerability in trustworthy AI that current evaluation paradigms overlook. This vulnerability has serious implications for real-world platforms, especially in digital marketplaces where recommender systems influence purchasing decisions, content visibility, and user engagement. When explanations fail under adversarial pressure, users may be misled at critical moments, undermining trust in the platform's personalization algorithms.

Our findings directly answer the initial research question, How do adversarial attacks simultaneously affect recommendation accuracy and the stability of post-hoc explanations across different recommender system architectures? Which sought to move beyond siloed evaluations of accuracy and investigate how adversarial attacks jointly impact recommendation quality and explanation reliability.

While prior research has extensively documented the vulnerability of recommender systems to adversarial attacks that degrade ranking accuracy (e.g.,[60], these studies have focused on performance metrics [49], while the behavior of the explanation layer under attack remained unexplored. In a parallel stream, XAI research has noted that post-hoc methods like LIME and SHAP can be unstable or manipulated [41]. Nevertheless these findings have rarely been demonstrated within recommender pipelines or explicitly linked to adversarial model failure [61]. Our study bridges this critical gap by providing the first systematic evidence that explanation integrity is not an independent property but is conditioned on model robustness, a connection that has been theorized but rarely shown empirically across multiple architectures.

The dual failure phenomenon was consistently observed across three distinct model architectures. For the transparent, content-based TF-IDF model, semantically aware text attacks succeeded in altering recommendations as well as caused a catastrophic collapse in the keyword-based explanations, with the Jaccard similarity between original and attacked explanations plummeting to an average of just 0.24. More critically, with the high-performance NeuMF model, a small data poisoning attack promotes a target item and fundamentally corrupted the model's internal logic. As revealed through SHAP analysis, the attack forced the model to abandon learned, generalizable patterns in favor of a memorized heuristic, fundamentally rewiring its decision-making process. The resulting explanations, while potentially appearing plausible, would be unfaithful to the original, intended logic of the model.

Our paper shows that the most accurate models (NeuMF, SVD) are highly vulnerable to data poisoning. The SVD model exemplifies a particularly perilous trade-off within the trilemma. While accurate, it is inherently non-interpretable, a vulnerability that becomes critical under attack. In the face of a successful poisoning attack, the SVD model offers no explanation to scrutinize, leaving both users and developers completely blind to the manipulation. Unlike the TF-IDF or NeuMF models—where a collapsing explanation can serve as an indirect alarm signaling that something is wrong—the SVD's failure is silent and total. This absence of any diagnostic trail makes its vulnerability arguably the most insidious for a platform that values user trust and algorithmic accountability. The TF-IDF model, while seemingly transparent, is not only the least accurate but is also highly vulnerable to textual attacks that shatter its explanations. This suggests that a myopic focus on a single metric can obscure deep vulnerabilities in other critical dimensions of trustworthiness.

One of the unexpected findings was the effectiveness of the attack on the "full" user profile in the TF-IDF model, proved more successful than a more focused attack on a "minimal" profile, though the difference was not statistically significant. This could imply that adversaries may achieve greater success with subtle, distributed perturbations across a target's entire data history rather than with a single, aggressive alteration, a potentially more insidious and harder-to-detect attack vector. However, the most significant finding was the sheer severity of the explanation collapse. While some instability was expected, observing a very low Jaccard similarity indicating a complete replacement of the explanatory keyword set for an item provides a starker-than-anticipated confirmation of our hypothesis. This result is highly significant to our research question because it demonstrates that the link between robustness and explainability is not merely theoretical. In practice, a successful attack can leave an explanation with absolutely no resemblance to the original, thereby transforming a tool intended for transparency into a potential vehicle for misinformation.

B. Implications for Theory

Theoretically, our findings challenge the common modular view of explainability as a separable, post-hoc component applied to predictive models. The Recommender's Trilemma, introduced and empirically characterized in this paper, offers a new conceptual framework that captures the inherent tensions between accuracy, robustness, and explanation quality. It encourages researchers to stop treating these dimensions as independent objectives and instead to consider their joint optimization under adversarial constraints.

In addition, we contribute to the ongoing redefinition of explanation quality (e.g., [62], [63] by emphasizing two dimensions: faithfulness (how well the explanation reflects the model's actual decision process) and stability (how consistently the explanation behaves under minor input changes). Our results demonstrate that a model may appear interpretable under standard evaluation but fail these two criteria when exposed to adversarial scenarios. This insight helps refine theoretical models of XAI in sequential and user-interaction–driven domains, which have traditionally been overlooked in favor of classifiers. In turn, this highlights the ethical tension in AI

deployment: a system that provides misleading yet convincing explanations may do more harm than good, violating core principles of algorithmic transparency and user autonomy central to trustworthy e-commerce design.

The chart visually demonstrates the trade-offs between Accuracy (HR@10), Robustness (inverse of attack impact), and Explanation Reliability (Jaccard similarity/SHAP stability). The most accurate model, NeuMF, is highly vulnerable, and its explanations are unstable. The most transparent model, TF-IDF, has the lowest accuracy. This visualization makes the central challenge of the trilemma clear: optimizing one dimension often comes at the expense of the others.

C. Implications for Practice

For developers and engineers building recommender systems, our findings carry practical warnings. Models optimized exclusively for accuracy can produce brittle explanations that erode under adversarial manipulation. This brittleness is often silent and undetectable through conventional metrics. We recommend that explainability be incorporated into model evaluation pipelines from the outset. Adversarial stress testing, explanation drift analysis, and fidelity benchmarks should become standard components of responsible AI deployment.

In content-based systems, semantic normalization (e.g., standardizing synonymous terms) may help reduce explanation variance. In deep models, training with explanation-aware objectives, such as attribution alignment loss or explanation regularization, may yield greater consistency. As a concrete first step, developers should implement explanation drift monitoring (e.g., by tracking the Jaccard similarity of explanation features over time) as a standard robustness check in continuous integration and deployment pipelines. A sudden drop in this metric could serve as an early warning for a potential adversarial attack or model degradation.

For platform operators, particularly in high-stakes environments such as e-commerce, healthcare, or financial services, our findings raise serious concerns about the integrity of AI-driven recommendations. In these domains, where explainability is vital, such instability can erode user trust or hinder compliance. If these explanations can be easily manipulated without users or developers noticing then trust in the entire system is jeopardized. In these settings, developers may need to prioritize robustness of explanations over marginal gains in accuracy. Explanations should therefore be treated not just as features but as part of the system's attack surface, subject to monitoring, defense, and validation like any other critical component. In e-commerce ecosystems, where recommender outputs directly influence purchase decisions, search rankings, and vendor exposure, ensuring that explanations are stable under adversarial manipulation is a technical challenge and a commercial and ethical imperative. such failures could distort competitive visibility, damage user trust, and introduce unfair commercial advantages, posing both technical, ethical and regulatory risks.

Finally, for regulators and policymakers, our framework offers a concrete and testable way to evaluate claims of algorithmic transparency. These findings also support the call for more structured evaluation frameworks that include explanation robustness and adversarial resistance. Regulatory guidelines such as the EU AI Act and emerging U.S. FTC recommendations emphasize interpretability and fairness. However, without robustness as a core evaluation criterion, explainability claims may be misleading. Regulatory bodies may soon require algorithmic systems, including recommenders to be audited for not just accuracy, but also stability and interpretability under hostile conditions. Our study supports the case for requiring both interpretability and resistance to manipulation as part of future AI certification standards.

D. Limitations and Future Directions

Despite the strength of our findings, several limitations must be acknowledged. First, our experiments were conducted within a single domain: Amazon Electronics reviews. Although this is a widely used benchmark in recommendation research, its specific content characteristics and user behavior patterns may limit generalizability. Future work should replicate our analysis across diverse domains such as music, news, or healthcare, where item types, user motivations, and data structures differ considerably.

Second, we evaluated only three recommendation architectures—TF-IDF SVD, and NeuMF. While they represent a useful contrast in transparency and complexity, they do not cover newer models such as graph-based recommenders (e.g., LightGCN) or transformer-based models (e.g., BERT4Rec). These architectures may exhibit distinct trilemma behaviors that deserve dedicated study. Future research can explore explainability under attack in more complex settings, such as graph-based recommenders or multimodal systems. Investigating robustness-aware explanation metrics in sequential and session-based models may further generalize the trilemma beyond the current architecture scope.

Third, our comparative analysis of explanation stability was not fully applicable to the SVD model, whose latent factors lack direct, feature-level interpretability. This reflects a core tenet of the trilemma, that gains in accuracy can come at the direct cost of explainability, but it is a limitation within our empirical framework. We acknowledge that methods for interpreting latent factor models exist, such as post-hoc semantic labeling of dimensions. However, these often provide general insights and are insufficient for auditing the instance-specific manipulations shown in our attacks. Future work could explore whether modern surrogate modeling techniques can generate approximate explanations for such models, enabling a more complete trilemma analysis under adversarial conditions.

Fourth, our evaluation of explanation stability relied primarily on LIME and SHAP. It is well-documented that LIME explanations can be inherently unstable under minor input perturbations, independent of any adversarial manipulation [48]. This raises a legitimate concern: a portion of the Jaccard similarity drop observed in our experiments may reflect LIME's natural variance rather than purely attack-induced explanation collapse. We partially mitigate this by grounding our post-attack results against the expected clean-model stability range reported in prior literature (0.85–0.92), and by corroborating explanation drift with SHAP attribution

analysis on the NeuMF model. Nevertheless, future work should evaluate explanation stability using multiple complementary methods — including counterfactual and causal explanation approaches — to disentangle LIME's inherent instability from adversarially induced collapse, and to determine whether more stable explanation methods alter the trilemma dynamics observed here. Fifth, our robustness evaluation was computational. We did not assess how changes in explanation quality impact user understanding, trust, or decision-making. Human-centered experiments are needed to determine whether users can detect explanation collapse, and if so, whether it influences their satisfaction or compliance with recommendations.

Finally, although we proposed several defensive strategies such as input sanitization, attribution smoothing, and explanation-aware objectives, we did not implement or empirically validate these approaches in this paper. Future research should systematically explore these defenses and develop joint optimization strategies that improve all three pillars of the trilemma without sacrificing performance.

VI. Conclusion

This paper introduces the Recommender's Trilemma, a framework formalizing the inherent tension between accuracy, robustness, and explainability. Through systematic experiments on three diverse recommender architectures, we empirically validate the dual failure cascade: a critical phenomenon where adversarial attacks simultaneously degrade recommendation accuracy and the stability of post-hoc explanations.

These findings challenge current evaluation practices, which often treat these goals in isolation, and underscore the need for a holistic approach to building trustworthy AI. Our results demonstrate that explanations can be misleadingly brittle and cannot be decoupled from model robustness.

We therefore advocate for joint, resilience-aware evaluation protocols and a shift toward developing explanation-aware adversarial training methods. Future work should extend the trilemma to sequential and graph-based recommenders and conduct human-centered studies to measure the impact of explanation collapse on user trust. Addressing this trilemma is not just a technical challenge but an ethical imperative for deploying recommender systems in high-stakes domains.

For the e-commerce domain in particular, these findings underscore a pressing need to reassess the deployment of recommender systems in environments that demand both personalization and transparency. As platforms increasingly rely on AI-driven recommendations to shape user journeys and influence market dynamics, the Recommender's Trilemma provides a foundational framework to guide the design of systems that are not only performant, but also ethically robust and resistant to manipulation. This work contributes to the emerging discourse on responsible AI in commercial platforms and offers actionable insights for both researchers and practitioners operating at the intersection of personalization, trust, and platform integrity.

Acknowledgment

The authors acknowledge the use of ChatGPT (GPT-5.3), an AI language model developed by OpenAI, to assist with proofreading and improve the clarity and coherence of this manuscript. Additionally, Sonnet, developed by Anthropic (Claude), was used to provide code templates and assist in debugging code related to the project.

Supplementary Materials

To enhance the reproducibility of the paper, the full source code for data preprocessing, model training, adversarial attack generation, and explanation analysis is available at: https://github.com/alohalimansor/Adversarial_Recommender

Data Availability Statement

The dataset used in this study — the Amazon Electronics Review Dataset, is publicly available at: https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2//

References

[1] Z. Shahbazi, R. Jalali, and Z. Shahbazi, "Enhancing recommendation systems with real-time adaptive learning and multi-domain knowledge graphs," Big Data Cogn. Comput., vol. 9, no. 5, p. 124, 2025.

[2] R. El Youbi, F. Messaoudi, M. Loukili, and M. El Ghazi, "Elevating E-commerce Customer Experience: A Machine Learning-Driven Recommendation System," Stat. Optim. Inf. Comput., vol. 14, no. 2, pp. 704–717, 2025.

[3] X. Fu, N. Chen, P. Gao, and Y. Li, "Privacy-Preserving Personalized Recommender Systems," Manuf. Serv. Oper. Manag., vol. 28, no. 1, pp. 271–289, Jan. 2026, doi: 10.1287/msom.2023.0271.

[4] L. Aliberti, G. D'Aniello, and M. Gaeta, "Situation-aware recommender systems: a systematic review and framework for trustworthy recommendations," Artif. Intell. Rev., 2026, Accessed: Mar. 14, 2026. [Online]. Available: https://link.springer.com/content/pdf/10.1007/s10462-026-11503-y_reference.pdf

[5] S. Gheewala, S. Xu, and S. Yeom, "In-depth survey: deep learning in recommender systems—exploring prediction and ranking models, datasets, feature analysis, and emerging trends," Neural Comput. Appl., vol. 37, no. 17, pp. 10875–10947, Jun. 2025, doi: 10.1007/s00521-024-10866-z.

[6] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," ACM Trans. Inf. Syst., vol. 22, no. 1, pp. 5–53, Jan. 2004, doi: 10.1145/963770.963772.

[7] B. Walek and O. Sládek, "Comparison of Selected Algorithms in Movie Recommender System," Appl. Sci., vol. 15, no. 17, p. 9518, 2025.

[8] Y. Deldjoo, N. Mehta, M. Sathiamoorthy, S. Zhang, P. Castells, and J. McAuley, "Toward Holistic Evaluation of Recommender Systems Powered by Generative Models," in Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, Padua Italy: ACM, Jul. 2025, pp. 3932–3942. doi: 10.1145/3726302.3730354.

[9] E. M. Voorhees, "The trec-8 question answering track report.," in Trec, 1999, pp. 77–82. Accessed: Jun. 23, 2025. [Online]. Available: http://trec.nist.gov/pubs/trec8/papers/qa_report.pdf

[10] A. Zhang, L. Sheng, Z. Cai, X. Wang, and T.-S. Chua, "Empowering collaborative filtering with principled adversarial contrastive loss," Adv. Neural Inf. Process. Syst., vol. 36, pp. 6242–6266, 2023.

[11] Y. Ge et al., "A Survey on Trustworthy Recommender Systems," ACM Trans. Recomm. Syst., vol. 3, no. 2, pp. 1–68, Jun. 2025, doi: 10.1145/3652891.

[12] C. Bauer, A. Said, and E. Zangerle, "Introduction to the Special Issue on Perspectives on Recommender Systems Evaluation," ACM Trans. Recomm. Syst., vol. 2, no. 1, pp. 1–5, Mar. 2024, doi: 10.1145/3648398.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Adv. Neural Inf. Process. Syst., vol. 30, 2017, Accessed: Jun. 21, 2025. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA: ACM, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[15] T. D. Pujari, D. K. Kejriwal, and A. Goel, "Robust Explainable AI via Adversarial Latent Diffusion Models: Mitigating Gradient Obfuscation with Interpretable Feature Attribution," 2024, Accessed: Jun. 21, 2025. [Online]. Available: https://www.researchgate.net/profile/Deepak-Kejriwal/publication/390972887_Robust_Explainable_AI_via_Adversarial_Latent_Diffusion_Models_Mitigating_Gradient_Obfuscation_with_Interpretable_Feature_Attribution/links/6806740cdf0e3f544f43ab72/Robust-Explainable-AI-via-Adversarial-Latent-Diffusion-Models-Mitigating-Gradient-Obfuscation-with-Interpretable-Feature-Attribution.pdf

[16] S. Vijayaraghavan and P. Mohapatra, "Robust Explainable Recommendation," Mar. 07, 2025, arXiv: arXiv:2405.01855. doi: 10.48550/arXiv.2405.01855.

[17] A. Dugăeşescu and A. M. Florea, "Evaluation and analysis of visual methods for CNN explainability: a novel approach and experimental study," Neural Comput. Appl., vol. 37, no. 20, pp. 14935–14970, Jul. 2025, doi: 10.1007/s00521-025-11282-7.

[18] J. Govea, R. Gutierrez, and W. Villegas-Ch, "Transparency and precision in the age of AI: evaluation of explainability-enhanced recommendation systems," Front. Artif. Intell., vol. 7, p. 1410790, 2024.

[19] T. Chang, Z. Zhang, and X. Cai, "Explainable recommender system directed by reconstructed explanatory factors and multi-modal matrix factorization," Concurr. Comput. Pract. Exp., vol. 36, no. 21, p. e8208, Sep. 2024, doi: 10.1002/cpe.8208.

[20] T. T. Nguyen et al., "Manipulating Recommender Systems: A Survey of Poisoning Attacks and Countermeasures," ACM Comput. Surv., vol. 57, no. 1, pp. 1–39, Jan. 2025, doi: 10.1145/3677328.

[21] Z. Wang, M. Gao, J. Yu, H. Ma, H. Yin, and S. Sadiq, "Poisoning Attacks against Recommender Systems: A Survey," Jan. 14, 2024, arXiv: arXiv:2401.01527. doi: 10.48550/arXiv.2401.01527.

[22] M. Yin, Y. Xu, M. Fang, and N. Z. Gong, "Poisoning Federated Recommender Systems with Fake Users," in Proceedings of the ACM Web Conference 2024, Singapore Singapore: ACM, May 2024, pp. 3555–3565. doi: 10.1145/3589334.3645492.

[23] G. Figà-Talamanca, "Digitally Scaffolded Vulnerability: Facebook's Recommender System as an Affective Scaffold and a Tool for Mind Invasion," Topoi, vol. 43, no. 3, pp. 631–643, Aug. 2024, doi: 10.1007/s11245-024-10051-w.

[24] Y. Himeur, S. S. Sohail, F. Bensaali, A. Amira, and M. Alazab, "Latest trends of security and privacy in recommender systems: a comprehensive review and future perspectives," Comput. Secur., vol. 118, p. 102746, 2022.

[25] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Inf. Process. Manag., vol. 24, no. 5, pp. 513–523, 1988.

[26] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," Computer, vol. 42, no. 8, pp. 30–37, 2009.

[27] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural Collaborative Filtering," in Proceedings of the 26th International Conference on World Wide Web, Perth Australia: International World Wide Web Conferences Steering Committee, Apr. 2017, pp. 173–182. doi: 10.1145/3038912.3052569.

[28] L. Fu, J. Zhang, and L. Li, "Dual Clustering Based Collaborative Filtering for Robust Recommendation," in 2024 11th International Conference on Behavioural and Social Computing (BESC), IEEE, 2024, pp. 1–7. Accessed: Jun. 21, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10780489/

[29] J. Wu, J. Wang, C. Xiao, C. Wang, N. Zhang, and Y. Vorobeychik, "Preference Poisoning Attacks on Reward Model Learning," Oct. 08, 2024, arXiv: arXiv:2402.01920. doi: 10.48550/arXiv.2402.01920.

[30] S. Xu et al., "On the Vulnerability of Graph Learning-based Collaborative Filtering," ACM Trans. Inf. Syst., vol. 41, no. 4, pp. 1–28, Oct. 2023, doi: 10.1145/3572834.

[31] A. Paul, Z. Wu, B. Chen, K. Luo, and L. Fang, "Interpretable adversarial neural pairwise ranking for academic network embedding," Knowl. Inf. Syst., vol. 67, no. 4, pp. 3293–3315, Apr. 2025, doi: 10.1007/s10115-024-02311-3.

[32] C. Tian, F. Zhang, and R. Wang, "Adversarial regularized attributed network embedding for graph anomaly detection," Pattern Recognit. Lett., vol. 183, pp. 111–116, 2024.

[33] J. Tang, H. Wen, and K. Wang, "Revisiting Adversarially Learned Injection Attacks Against Recommender Systems," in Fourteenth ACM Conference on Recommender Systems, Virtual Event Brazil: ACM, Sep. 2020, pp. 318–327. doi: 10.1145/3383313.3412243.

[34] G. G. Shayea, M. H. M. Zabil, M. A. Habeeb, Y. L. Khaleel, and A. S. Albahri, "Strategies for protection against adversarial attacks in AI models: An in-depth review," J. Intell. Syst., vol. 34, no. 1, p. 20240277, Mar. 2025, doi: 10.1515/jisys-2024-0277.

[35] L. Boratto, F. Fabbri, G. Fenu, M. Marras, and G. Medda, "Robustness in Fairness Against Edge-Level Perturbations in GNN-Based Recommendation," in Advances in Information Retrieval, vol. 14610, N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, and I. Ounis, Eds., in Lecture Notes in Computer Science, vol. 14610. , Cham: Springer Nature Switzerland, 2024, pp. 38–55. doi: 10.1007/978-3-031-56063-7_3.

[36] Y. Zhou, H. Wang, J. He, and H. Wang, "Review-Based Explainable Recommendations: A Transparency Perspective," ACM Trans. Recomm. Syst., vol. 3, no. 3, pp. 1–20, Sep. 2025, doi: 10.1145/3701762.

[37] H. Mohammadi, A. Bagheri, A. Giachanou, and D. L. Oberski, "Explainability in Practice: A Survey of Explainable NLP Across Various Domains," Jun. 05, 2025, arXiv: arXiv:2502.00837. doi: 10.48550/arXiv.2502.00837.

[38] N. Tiwary, S. A. M. Noah, F. Fauzi, and T. S. Yee, "A Review of Explainable Recommender Systems Utilizing Knowledge Graphs and Reinforcement Learning," IEEE Access, 2024, Accessed: Jun. 21, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10583886/

[39] Z. Xu, H. Zeng, J. Tan, Z. Fu, Y. Zhang, and Q. Ai, "A Reusable Model-agnostic Framework for Faithfully Explainable Recommendation and System Scrutability," ACM Trans. Inf. Syst., vol. 42, no. 1, pp. 1–29, Jan. 2024, doi: 10.1145/3605357.

[40] J. Ahmad, J. Hellgren, and A. Said, "Tell Me the Good Stuff: User Preferences in Movie Recommendation Explanations," in Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization, New York City USA: ACM, Jun. 2025, pp. 103–108. doi: 10.1145/3708319.3733676.

[41] D. Vreš and M. Robnik-Šikonja, "Preventing deception with explanation methods using focused sampling," Data Min. Knowl. Discov., vol. 38, no. 5, pp. 3262–3307, Sep. 2024, doi: 10.1007/s10618-022-00900-w.

[42] M. Noppel and C. Wressnegger, "SoK: Explainable machine learning in adversarial environments," in 2024 IEEE Symposium on Security and Privacy (SP), IEEE, 2024, pp. 2441–2459. Accessed: Jun. 21, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10646794/

[43] M. Ravi, A. Negi, N. S. Bommi, and N. Rouf, "Evolution of AI-Driven Decision Making with Decision Support Systems, Expert Systems, Recommender Systems, and XAI," IETE Tech. Rev., vol. 42, no. 4, pp. 428–465, Jul. 2025, doi: 10.1080/02564602.2025.2512086.

[44] H. Tao et al., "DeepRS: A Library of Recommendation Algorithms Based on Deep Learning," Int. J. Comput. Intell. Syst., vol. 15, no. 1, p. 45, Dec. 2022, doi: 10.1007/s44196-022-00102-8.

[45] I. A. Zahid et al., "Explainability, robustness, and fairness in user-centric intelligent systems: a systematic review," IEEE Trans. Emerg. Top. Comput. Intell., 2025, Accessed: Mar. 14, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/11012706/

[46] J. D. Pinto and L. Paquette, "Towards a Unified Framework for Evaluating Explanations," Jul. 14, 2024, arXiv: arXiv:2405.14016. doi: 10.48550/arXiv.2405.14016.

[47] B. Barr, N. Fatsi, L. Hancox-Li, P. Richter, D. Proano, and C. Mok, "The Disagreement Problem in Faithfulness Metrics," Nov. 13, 2023, arXiv: arXiv:2311.07763. doi: 10.48550/arXiv.2311.07763.

[48] P. Knab, S. Marton, U. Schlegel, and C. Bartelt, "Which LIME should I trust? Concepts, Challenges, and Solutions," Mar. 31, 2025, arXiv: arXiv:2503.24365. doi: 10.48550/arXiv.2503.24365.

[49] H. Baniecki and P. Biecek, "Adversarial attacks and defenses in explainable artificial intelligence: A survey," Inf. Fusion, p. 102303, 2024.

[50] K. Zhang et al., "Robust Recommender System: A Survey and Future Directions," Apr. 01, 2025, arXiv: arXiv:2309.02057. doi: 10.48550/arXiv.2309.02057.

[51] M. Vázquez-Hernández, L. A. Morales-Rosales, I. Algredo-Badillo, S. I. Fernández-Gregorio, H. Rodríguez-Rangel, and M.-L. Córdoba-Tlaxcalteco, "A Survey of Adversarial Attacks: An Open Issue for Deep Learning Sentiment Analysis Models," Appl. Sci., vol. 14, no. 11, p. 4614, 2024.

[52] Y. Lai, Y. Zhu, B. Pan, and K. Zhou, "Node-aware Bi-smoothing: Certified Robustness against Graph Injection Attacks," in 2024 IEEE Symposium on Security and Privacy (SP), IEEE, 2024, pp. 2958–2976. Accessed: Jun. 21, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10646863/

[53] B. G. Anderson and S. Sojoudi, "Certified robustness via locally biased randomized smoothing," in Learning for Dynamics and Control Conference, PMLR, 2022, pp. 207–220. Accessed: Jun. 21, 2025. [Online]. Available: https://proceedings.mlr.press/v168/anderson22a

[54] S. Verma, V. Boonsanong, M. Hoang, K. Hines, J. Dickerson, and C. Shah, "Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review," ACM Comput. Surv., vol. 56, no. 12, pp. 1–42, Dec. 2024, doi: 10.1145/3677119.

[55] J. Liao, M. Yang, W. Zhou, H. Zhang, and J. Wen, "Modeling item exposure and user satisfaction for debiased recommendation with causal inference," Inf. Sci., vol. 676, p. 120834, 2024.

[56] C. Gao, Y. Zheng, W. Wang, F. Feng, X. He, and Y. Li, "Causal Inference in Recommender Systems: A Survey and Future Directions," ACM Trans. Inf. Syst., vol. 42, no. 4, pp. 1–32, Jul. 2024, doi: 10.1145/3639048.

[57] S. Vijayaraghavan and P. Mohapatra, "Stability of Explainable Recommendation," in Proceedings of the 17th ACM Conference on Recommender Systems, Singapore Singapore: ACM, Sep. 2023, pp. 947–954. doi: 10.1145/3604915.3608853.

[58] Y. M. Latha and B. S. Rao, "Amazon product recommendation system based on a modified convolutional neural network," ETRI J., vol. 46, no. 4, pp. 633–647, Aug. 2024, doi: 10.4218/etrij.2023-0162.

[59] E. Ahmed and A. Letta, "Book Recommendation Using Collaborative Filtering Algorithm," Appl. Comput. Intell. Soft Comput., vol. 2023, pp. 1–12, Mar. 2023, doi: 10.1155/2023/1514801.

[60] T. T. Nguyen et al., "Manipulating Recommender Systems: A Survey of Poisoning Attacks and Countermeasures," ACM Comput. Surv., vol. 57, no. 1, pp. 1–39, Jan. 2025, doi: 10.1145/3677328.

[61] R. Liang, Y. Chai, W. Li, M. Chau, Y. Jiang, and Y. Liu, "Assessing and enhancing adversarial robustness for review-based recommender system: A design science approach," in Pacific Asia Conference on Information Systems-PACIS 2023 (09/07/2023-12/07/2023, Nanchang), 2023. Accessed: Jun. 22, 2025. [Online]. Available: https://hub.hku.hk/handle/10722/337444

[62] Z. Xu, H. Zeng, J. Tan, Z. Fu, Y. Zhang, and Q. Ai, "A Reusable Model-agnostic Framework for Faithfully Explainable Recommendation and System Scrutability," ACM Trans. Inf. Syst., vol. 42, no. 1, pp. 1–29, Jan. 2024, doi: 10.1145/3605357.

[63] H. Zhuang, W. Zhang, W. Chen, J. Yang, and Q. Z. Sheng, "Improving Faithfulness and Factuality with Contrastive Learning in Explainable Recommendation," ACM Trans. Intell. Syst. Technol., vol. 16, no. 1, pp. 1–23, Feb. 2025, doi: 10.1145/3653984.