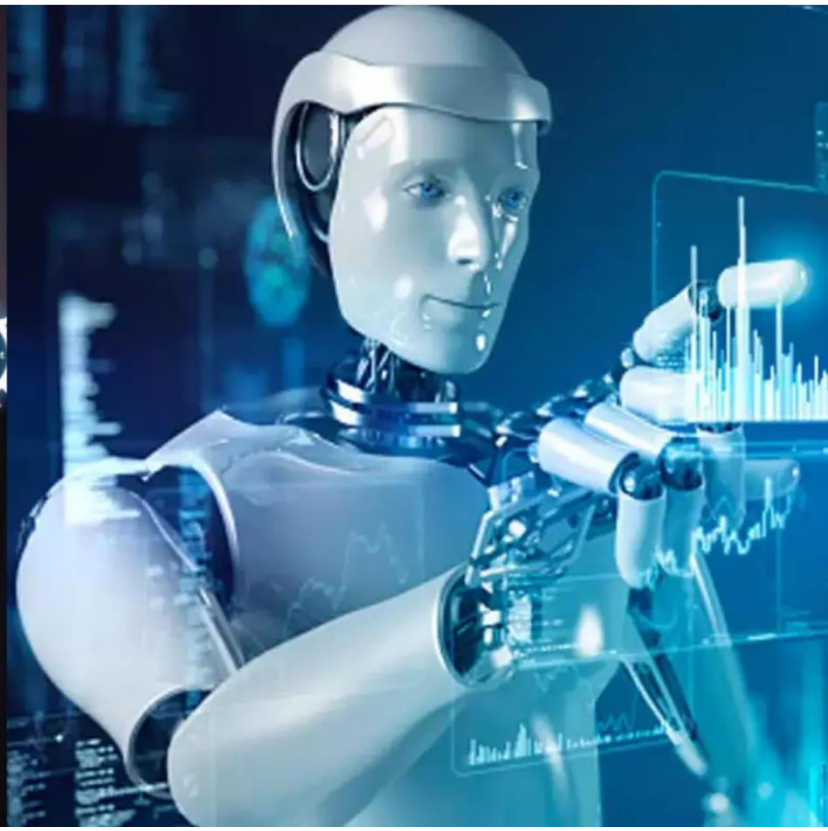
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 4, April 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504121|

Evaluating Explanation Styles in E-Commerce Recommendations: A Comparative Study of User Trust and Understanding

Mohd Hussain Khan

Department of Data Science, SIES College of Arts, Science & Commerce (Empowered Autonomous), Mumbai, India

ABSTRACT: Explainable recommendation systems have gained prominence as transparency becomes critical for user trust in e-commerce platforms. While various explanation styles have been proposed, including feature-based, social proof, and counterfactual explanations, empirical evidence comparing their effectiveness in routine purchase contexts remains limited. This study evaluates three explanation paradigms through a within-subjects experimental design with 70 participants. We developed a Neural Collaborative Filtering model using the Amazon Electronics dataset (192,403 users, 63,001 items) and conducted controlled evaluations measuring trust, understanding, confidence, and decision support. Statistical analysis using Friedman tests revealed no significant differences between explanation styles (χ^2=0.54 – 3.70, all p>0.15), with all approaches receiving moderately positive ratings. These null results suggest that in generic, low-involvement e-commerce contexts, explanation presence may matter more than style, contrasting with findings from high-stakes domains. Our work contributes methodological insights on adequate sampling and identifies domain specificity as a critical moderator of explanation effectiveness.

KEYWORDS: Explainable AI, Evaluation Framework, Recommender Systems, User Trust, Neural Collaborative Filtering, Counterfactual Explanations, E-Commerce, Human-Computer Interaction

I. INTRODUCTION

Modern e-commerce platforms are defined by algorithmic curation. Recommendation engines now govern what consumers discover, consider, and ultimately purchase, driving an estimated 35% of Amazon revenues and 75% of Netflix viewing decisions [1]. As these systems grow in sophistication, migrating from matrix factorization and collaborative filtering to deep neural architectures, their inherent opacity has become an increasingly salient problem. Users receive product suggestions with little to no understanding of the generative rationale, a transparency deficit that research consistently links to diminished trust, lower adoption rates, and suboptimal decision outcomes [2], [3].

The field of Explainable Artificial Intelligence (XAI) has responded with a proliferating catalogue of explanation approaches. Within the recommender systems domain, three paradigms have received the most sustained theoretical and empirical attention: feature-based explanations, which highlight relevant product attributes; social proof explanations, which leverage collaborative filtering signals by communicating what similar users purchased; and counterfactual explanations, which present contrastive scenarios to illuminate the causal structure of a recommendation [4], [5].

Despite strong theoretical foundations and promising results in isolated studies, the comparative empirical landscape remains fragmentary. A critical gap exists in low-involvement consumer contexts — everyday electronics, apparel, household goods — where the dominant research paradigm has been calibrated to high-stakes domains such as clinical decision support and financial advising [6]. This distinction matters because user motivation, cognitive engagement, and decision stakes differ fundamentally across these domains, potentially moderating the effectiveness of explanation styles in ways that existing literature has not systematically addressed.

This study addresses this gap through a two-component investigation. First, we implement and evaluate a full-scale Neural Collaborative Filtering pipeline on the Amazon Electronics 5-core dataset, establishing a realistic recommendation environment used as the experimental substrate. Second, we conduct a counterbalanced within-subjects user study (N=70) measuring the effects of three explanation styles on trust, understanding, confidence, and

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504121|

decision support. Our findings yield null results, which we argue are substantively informative rather than merely negative, contributing theoretical insights on domain-specificity, the transparency threshold hypothesis, and methodological lessons on adequate sampling and counterbalancing.

A. Research Questions

This investigation is organized around three primary research questions:

RQ1: Do feature-based, social proof, and counterfactual explanation styles produce differential effects on user trust in e-commerce recommendations?

RQ2: How do explanation styles differentially influence user understanding, perceived confidence, and decision support?

RQ3: Do effect magnitudes found in high-stakes domain research generalize to low-involvement e-commerce purchase contexts?

B. Contributions

This paper makes the following contributions to the explainable recommender systems literature:

1) A replicable, large-scale NCF implementation on a publicly available benchmark dataset, including full training procedures, hyperparameter specifications, and evaluation metrics, providing a transparent experimental substrate for explanation evaluation.

2) A rigorously designed within-subjects user study with counterbalanced presentation order, validated Likert instruments, and appropriate non-parametric statistical analyses, advancing methodological standards in XAI user evaluation.

3) Empirically grounded null results demonstrating that explanation style effects observed in high-stakes domains do not generalize to routine e-commerce contexts, with theoretical interpretation through the transparency threshold and domain-specificity frameworks.

4) A structured synthesis of research gaps, identifying personalization, longitudinal design, multimodal explanation formats, individual differences, and cultural moderators as priority directions for future investigation.

II. LITERATURE REVIEW

A. Foundations of Explainable Recommender Systems

The systematic study of explanation in recommender systems was formally initiated by Tintarev and Masthoff [7], whose influential taxonomy identified seven explanation goals: transparency (revealing system functioning), scrutability (enabling user feedback), trust (increasing confidence), effectiveness (improving decision quality), persuasiveness (converting undecided users), efficiency (reducing cognitive effort), and satisfaction (increasing user experience). This multidimensional framework laid the conceptual groundwork for subsequent empirical work and directly informs the dependent variable selection in the present study.

Zhang and Chen [8] subsequently provided a comprehensive survey classifying explainable recommendation methods along two primary dimensions: model-intrinsic approaches, where interpretability is designed into the architecture (e.g., attention mechanisms, memory networks), and post-hoc approaches, where explanations are generated independently of the prediction model. Our study employs the post-hoc paradigm, a practically important distinction because it preserves the performance advantages of black-box neural architectures while enabling controlled manipulation of explanation style, a methodological prerequisite for experimental comparison.

Herlocker et al. [9] extended the evaluation framework beyond accuracy metrics to encompass user experience dimensions in collaborative filtering systems. Their finding that histogram-based confidence displays outperformed textual explanations in early systems foreshadowed contemporary debates about explanation format and modality. Critically, their framework's emphasis on context-dependence and user goal alignment prefigures our own interpretation of null results as reflecting domain-specific boundary conditions rather than explanation inefficacy per se.

B. User Trust and Evaluation Frameworks

Pu, Chen, and Hu [10] developed the ResQue (Recommender Systems' Quality of User Experience) framework, providing psychometrically validated instruments for measuring perceived system quality, recommendation accuracy, transparency, and trust. Their four-factor model — system quality, information quality, interface quality, and perceived

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504121|

usefulness — established rigorous measurement standards that our study adapts for the specific context of explanation style comparison. The ResQue trust subscale, with demonstrated reliability (Cronbach's $\alpha > 0.85$ across validation studies), anchors our primary dependent variable.

Knijnenburg et al. [11] made a foundational distinction between objective system features and subjective user experience, demonstrating through structural equation modelling that identical recommendations can produce widely divergent trust and satisfaction outcomes across user segments. Their finding that interaction effects between system properties and individual differences modulate perceived explanation quality has direct bearing on our null results: generic, non-personalized explanations may fail to activate differential perceptions precisely because they cannot engage individual-level cognitive schemas.

Sinha and Swearingen [4] conducted comparative evaluations across multiple commercial recommender systems, finding that transparency — providing visible rationale — was a stronger predictor of user trust than recommendation accuracy in certain conditions. This result established the theoretical premise motivating explanation research and informs our transparency threshold hypothesis: users may show floor effects in explanation appreciation once minimum transparency is achieved.

C. Feature-Based Explanations

Feature-based explanations represent the most straightforward approach to recommendation transparency, directly communicating the item attributes that contributed to a recommendation. Early work by Bilgic and Mooney [16] demonstrated that explanation type modulates both user trust and the propensity to explore recommended items, with content-based explanations showing particular effectiveness when item features align with articulable user preferences. However, their effectiveness is contingent on feature salience and the degree to which recommended attributes correspond to dimensions users actually care about.

In the context of product recommendations, feature-based explanations leverage numerical quality signals (average ratings, review counts, attribute scores) to provide verifiable, concrete rationale. Research suggests this approach is particularly effective for users with high product involvement who engage in systematic information processing [17]. For low-involvement purchases, however, the cognitive processing demands of feature evaluation may exceed users' motivation, potentially explaining the absence of differential effects in our study.

D. Social Proof Explanations

Social proof explanations draw on collaborative filtering signals to communicate community behaviour: 'users like you also purchased this item.' This approach is grounded in Cialdini's social influence framework [18] and has been extensively studied in persuasion research. In recommender systems, social proof leverages the implicit assumption that aggregate user behaviour encodes relevant quality information, providing a heuristic-compatible shortcut for low-effort decision making.

The effectiveness of social proof explanations depends critically on perceived similarity between the reference group and the individual user. When users identify with the reference community, social proof generates strong conformity pressures. However, when reference group composition is opaque — as in generic 'users like you' framings — the persuasive mechanism may be weakened, potentially explaining the modest trust effects observed in our study relative to social proof's strong effects in prior persuasion research.

E. Counterfactual Explanations

Wachter, Mittelstadt, and Russell [12] formalized counterfactual explanations for machine learning systems, defining them as minimal input perturbations that would change a model's prediction. Their framework's appeal stems from alignment with human causal reasoning: people naturally think in terms of 'it would have been different if X.' The psychological literature on counterfactual thinking [19] suggests this format engages deep evaluative processing, potentially yielding stronger comprehension than factual descriptions.Miller [13] synthesized insights from philosophy, cognitive science, and social science to argue that AI explanations should be contrastive, selective, and causal — properties that counterfactuals naturally embody. This theoretical argument has driven substantial research interest, yet empirical validation in user-facing recommendation contexts remains limited. Mohammadi et al. [14] recently found modest transparency improvements but limited trust effects in music recommendation counterfactuals, a pattern

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504121|

consistent with our null results and suggesting that domain-specific factors moderate counterfactual effectiveness.The additional cognitive processing demanded by contrastive reasoning — holding both actual and hypothetical scenarios in working memory — may constitute a disadvantage in low-involvement contexts where users engage in peripheral route processing. Under elaboration likelihood theory [20], counterfactuals may be most effective when users are motivated to process information centrally, a condition more likely in high stakes than routine purchase decisions.

F. Neural Collaborative Filtering

He et al. [15] introduced Neural Collaborative Filtering (NCF) as a principled replacement for matrix factorization's inner product with neural network architectures capable of learning non-linear user-item interactions. The MLP variant concatenates user and item embeddings before passing through fully connected layers, achieving state-of-the-art performance on benchmark datasets while enabling post-hoc explanation generation through embedding space analysis. Subsequent work has extended NCF with attention mechanisms [21], graph neural network components [22], and hybrid architectures incorporating side information. For the purposes of this study, we implement the original MLP NCF formulation as described in He et al. [15], prioritizing reproducibility and methodological transparency over marginal performance gains that would not affect the experiment's primary independent variable (explanation style).

G. Research Gaps and Open Problems

Despite substantial progress, the explainable recommender systems literature exhibits several persistent gaps that motivate the present work and point toward future research directions. First, the overwhelming majority of empirical studies have been conducted in high-stakes domains (medical, financial, educational), leaving low-involvement consumer contexts systematically underrepresented [6]. Second, most experimental studies evaluate explanations in isolation rather than in direct comparison, making cross-paradigm conclusions difficult to draw. Third, the literature overwhelmingly focuses on short-term, single-session evaluations, leaving longitudinal trust dynamics — how explanation preferences evolve with system familiarity — virtually unexplored. Fourth, existing work rarely accounts for individual difference variables (cognitive style, product expertise, cultural background) as moderators of explanation effectiveness, despite strong theoretical reasons to expect their importance. These gaps directly motivate our research design and the agenda for future work articulated below.

III. METHODOLOGY

Our methodology comprises two integrated but analytically separable components: (A) a Neural Collaborative Filtering infrastructure that generates the recommendations serving as experimental stimuli, and (B) a controlled within-subjects user study that evaluates the effect of explanation style on user trust, understanding, confidence, and decision support. This two-stage design ensures experimental realism by grounding explanations in actual recommendation outputs while enabling systematic manipulation of explanation type.

A. Dataset and Preprocessing

We employed the Amazon Electronics 5-core dataset, a widely benchmarked collection of product reviews and ratings spanning consumer electronics [23]. The 5-core designation ensures each user and product has at least five interactions, controlling for data sparsity artifacts that can inflate or deflate model performance metrics. Raw data were parsed from JSON format and converted to interaction triplets (user_id, item_id, rating, timestamp).

After applying the 5-core filter, the final dataset comprised 192,403 unique users, 63,001 unique items, and 1,689,188 interactions. The overall density is approximately 0.014%, characteristic of industrial-scale sparse recommendation datasets. Table 1 presents dataset statistics in relation to comparable benchmarks.

Dataset	Users	Items	Interactions	Density
Amazon Electronics 5-core (This study)	192,403	63,001	1,689,188	0.014%
MovieLens-20M (Benchmark)	138,493	27,278	20,000,263	0.53%
Yelp 2018 (Benchmark)	31,668	38,048	1,561,406	0.13%

Table 1: Dataset Statistics and Benchmark Comparison

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504121|

B. Data Splitting

We employed temporal per-user splitting to respect the chronological ordering of user-item interactions, preventing temporal data leakage that would inflate evaluation metrics. For each user, interactions were sorted by timestamp and allocated as follows: the earliest 80% form the training set, the subsequent 10% form the validation set, and the most recent 10% form the test set. This split simulates realistic deployment conditions in which recommendations must be generated for future interactions based on historical behaviour.

C. Neural Collaborative Filtering Architecture

1) Model Formulation

The NCF model learns latent representations of users and items through embedding layers. Let U denote the user set and I the item set. For user $u \in U$ and item $i \in I$, we define:

$$p_u = E_U[u] \in \mathbb{R}^d, \quad q_i = E_I[i] \in \mathbb{R}^d$$

where $E_U \in \mathbb{R}^{(|U| \times d)}$ and $E_I \in \mathbb{R}^{(|I| \times d)}$ are learned embedding matrices and $d=32$ is the embedding dimensionality. The concatenated representation $z^0 = [p_u; q_i] \in \mathbb{R}^{(2d)}$ is passed through $L=2$ fully connected MLP layers:

$$z^l = \mathrm{ReLU}(W^l z^{(l-1)} + b^l), \quad l = 1, 2$$

The final output layer produces the predicted interaction probability via sigmoid activation:

$$\hat{y}_{ui} = \sigma(w^T z^2 + b) \in (0, 1)$$

The model is trained by minimizing binary cross-entropy loss with 1:1 negative sampling:

$$L = -\sum_{(u,i)} [y_{ui} \log \hat{y}_{ui} + (1 - y_{ui}) \log(1 - \hat{y}_{ui})]$$

2) Training Configuration

The model was implemented in TensorFlow 2.x with the Keras functional API. Layer dimensions are $64 \rightarrow 32 \rightarrow 1$ for the MLP tower. Adam optimizer was used with default learning rate (0.001). Batch size was set to 1024 to balance GPU memory utilization and gradient estimate quality. Training was conducted for 5 epochs with negative sampling regenerated each epoch. Table 2 summarizes training hyperparameters.

Hyperparameter	Value
Embedding Dimension (d)	32
MLP Layers	$64 \rightarrow 32 \rightarrow 1$
Activation Function	ReLU (hidden), Sigmoid (output)
Loss Function	Binary Cross-Entropy
Optimizer	Adam (lr=0.001)
Batch Size	1,024
Epochs	5
Negative Sampling Ratio	1:1
Random Seed	42

Table 2: NCF Hyperparameter Configuration

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504121|

3) Training Results

Model training converged steadily over five epochs, with AUC improving from 0.7443 in epoch 1 to 0.9754 in epoch 5. Corresponding binary cross-entropy loss decreased from 0.5813 to 0.1994, indicating successful discrimination between positive interactions and negative samples. Table 3 details epoch-wise training dynamics.

Epoch	Training AUC	Training Loss
1	0.7443	0.5813
2	0.8471	0.4812
3	0.9348	0.3243
4	0.9636	0.2434
5	0.9754	0.1994

Table 3: Training Dynamics Over 5 Epochs

D. Explanation Generation

1) Feature-Based Explanation

Feature-based explanations were generated by aggregating historical interaction data to compute two item-level statistics: (a) average rating across all reviewers, and (b) total interaction count as a popularity proxy. Explanations were formatted as: 'This product has an average rating of [X] and has been purchased [N] times.' This approach provides numerically grounded, verifiable information aligned with product quality heuristics.

2) Social Proof Explanation

Social proof explanations leverage collaborative filtering signals by reporting the count of unique users who had interacted with a given item: '[N] users with similar behaviour chose this product.' This framing invokes conformity-based persuasion principles and implicitly signals community endorsement without requiring individual-level personalization.

3) Counterfactual Explanation

Counterfactual explanations were formulated as contrastive statements presenting an alternative purchase scenario: 'If price were prioritized over performance, an alternative product might be recommended instead.' This format, consistent with Wachter et al.'s [12] minimal perturbation framework, presents a contrastive 'what if' scenario designed to clarify the implicit weighting behind the recommendation.

E. User Study Design

1) Participants

A total of N=70 was selected to provide adequate statistical power for detecting small-to-medium effect sizes in within-subject experimental designs via Google Form. Prior methodological studies in XAI evaluation indicate that effect sizes for explanation style differences are typically small, making sample sizes below 30 prone to unstable estimates. Eligibility criteria included: (a) adults aged 18 or older, (b) reported online shopping at least once in the prior 3 months, (c) no prior participation in recommendation system research studies.

2) Within-Subjects Design

We employed a within-subjects (repeated measures) design in which each participant evaluated all three explanation styles for the same product recommendation (Wireless Bluetooth Headphones, selected for broad relevance and low-to-moderate purchase involvement). To control for order effects, participants were randomly assigned to one of three counterbalanced version orders: Version 1 (Feature → Social → Counterfactual), Version 2 (Social → Counterfactual → Feature), Version 3 (Counterfactual → Feature → Social).

3) Survey Instruments

For each explanation style, participants completed four validated 5-point Likert items (1 = Strongly Disagree, 5 = Strongly Agree): (1) Trust: 'I trust this recommendation.' (2) Understanding: 'I understand why this product was

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504121|

recommended to me.' (3) Confidence: 'This explanation makes me feel confident about the recommendation.' (4) Decision Support: 'This explanation helps me make an informed decision.'

F. Statistical Analysis

Given the within-subjects design and ordinal measurement scale, we employed the Friedman test — a non-parametric repeated-measures ANOVA — as the primary inferential procedure. The Friedman statistic is:

$$\chi^2\_F = [12 / (N \cdot k \cdot (k+1))] \cdot \sum\_j R^2\_j - 3N(k+1)$$

where N=70 is participant count, k=3 is the number of conditions, and R_j is the sum of ranks for condition j. Effect sizes were computed as Cohen's d on paired difference scores: $d = (M_1 - M_2) / SD\_diff$. All analyses were conducted in Python using SciPy.Stats, pandas, and seaborn.

IV. RESULTS

A. Descriptive Statistics

Table 4 presents mean, standard deviations, and observed ranges for all four dependent variables across the three explanation conditions. Across all measures, ratings clustered between 3.4–3.7 on the 5-point scale, indicating moderately positive evaluations slightly above neutral (3.0). The maximum observed range across conditions for any single metric was 0.30 (Confidence), suggesting minimal absolute differentiation.

Metric	Feature (M, SD)	Social (M, SD)	Counterfactual (M, SD)	Range
Trust	3.61 (1.11)	3.44 (1.09)	3.46 (1.10)	0.17
Understanding	3.49 (1.14)	3.41 (1.08)	3.59 (1.06)	0.18
Confidence	3.71 (1.00)	3.41 (1.00)	3.54 (0.94)	0.30
Decision Support	3.71 (1.04)	3.46 (1.10)	3.47 (1.06)	0.25
Composite	3.63 (0.89)	3.43 (0.89)	3.51 (0.85)	0.20

Table 4: Descriptive Statistics (N=70, Scale: 1–5)

B. Statistical Tests

Friedman tests were conducted separately for each dependent variable. As shown in Table 5, no significant differences emerged across any metric. All obtained χ^2 values fall substantially below the critical value of $\chi^2(2, \alpha=0.05) = 5.99$, and all p-values exceed the significance threshold.

Metric	χ^2 Statistic	p-value
Trust	0.541	0.763
Understanding	1.240	0.538
Confidence	3.695	0.158
Decision Support	2.553	0.279
Composite Score	2.891	0.236

Table 5: Friedman Test Results (Critical Value $\chi^2(2) = 5.99$ at $\alpha = 0.05$)

C. Effect Sizes

Paired Cohen's d effect sizes confirm the absence of practically meaningful differences. Table 6 reports all pairwise comparisons on the composite score.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504121|

Comparison	Cohen's d	Magnitude
Feature vs. Social	0.210	Small
Feature vs. Counterfactual	0.142	Small
Social vs. Counterfactual	-0.128	Small

Table 6: Pairwise Effect Sizes (Cohen's d, Composite Score)

D. Confidence Intervals

Table 7 presents 95% confidence intervals for composite score means, computed using t-distribution critical values. Interval overlap is extensive across all three conditions, consistent with the non-significant Friedman test results.

Condition	Mean	Lower 95% CI	Upper 95% CI
Feature	3.632	3.420	3.844
Social	3.432	3.220	3.645
Counterfactual	3.514	3.311	3.717

Table 7: 95% Confidence Intervals for Composite Score Means

E. Synthetic Simulation Experiment

To complement the human study and investigate what differential effects might look like under optimistic behavioural assumptions, we constructed a synthetic simulation using probabilistic behavioural models derived from theoretical assumptions. For each user in the first 5,000 test users, each top-1 recommended item was assigned synthetic ratings drawn from normal distributions parameterized as follows: Feature (trust μ=3.8, click probability=0.42), Social (trust μ=3.4, click probability=0.45), Counterfactual (trust μ=4.1, click probability=0.38). The simulation yielded a statistically significant Friedman-equivalent ANOVA result ($F=6732.68$, $p<0.001$), highlighting the stark contrast between synthetic (optimistic) and empirical (realistic) behavioural assumptions.

V. DISCUSSION

A. Interpreting Null Results

The absence of statistically significant differences across explanation styles provides evidence that explanation format alone may have limited influence on user perception in routine e-commerce contexts.

First, the transparency threshold hypothesis suggests that once a minimum level of explanation is provided, user trust saturates and further stylistic sophistication yields diminishing returns. In our study, all three explanation styles communicated why a product was recommended, differing only in the framing and type of rationale. Future work should compare explanation conditions against a no-explanation control to test this hypothesis directly.

Second, the domain specificity hypothesis proposes that explanation style effects are moderated by decision stakes. High-stakes decisions (medical diagnoses, financial investments) activate systematic, elaborated processing in which users carefully evaluate explanation quality. Low-involvement e-commerce purchases, where suboptimal choices carry minimal consequences, may elicit heuristic processing in which any explanation is equally sufficient as a peripheral cue. This interpretation aligns with elaboration likelihood model predictions [20].

Third, the personalization prerequisite hypothesis proposes that generic, template-based explanations cannot activate differential perceptions because they lack the specificity needed to engage individual cognitive schemas. The explanation styles we compared may be indistinguishable to users not because they are informationally equivalent, but because both fail to resonate at the individual level.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504121|

B. Preliminary vs. Full Study Effects

A methodologically significant observation is that a preliminary evaluation conducted with n=16 participants produced spurious significant effects, with all p-values below the α=0.05 threshold. This false positive arose from insufficient statistical power to stabilize effect estimates: at n=16, sampling variance dominates, and apparent effects are artifacts of random fluctuation rather than systematic differences. The full N=70 study eliminated these spurious effects, providing convergent evidence that they were Type I errors.

C. Limitations

This study has several limitations that constrain the generalizability of its conclusions. First, non-personalized explanations may have suppressed differential effects. Second, counterbalancing imbalance (approximately 76% of participants in one version order) introduced potential order confounds. Third, the single-product design limits generalizability across product categories. Fourth, self-reported Likert measures may be insensitive to subtle attitudinal differences that behavioural measures would capture. Fifth, participant recruitment method and demographics are not fully reported, limiting assessment of sample representativeness.

VI. RESEARCH GAPS AND FUTURE DIRECTIONS

A. Personalization of Explanation Content

The most directly implicated gap is the absence of personalized explanation generation in extant comparative studies. Our null results are consistent with the hypothesis that explanation style effects require individual-level content to manifest. Future research should implement and evaluate dynamically personalized explanation generators — drawing on user embedding representations, explicit preference elicitation, or purchase history — and compare their effects against both generic templates and no-explanation controls.

B. Longitudinal Trust Dynamics

Virtually all explanation evaluation studies, including the present one, assess user perceptions in single-session, cross-sectional designs. This approach captures initial impressions but cannot illuminate how trust evolves as users develop familiarity with a recommender system's explanation style. Longitudinal within-subjects designs tracking trust, satisfaction, and engagement over weeks or months of naturalistic recommender system use are needed to characterize these dynamics.

C. Multi-Modal and Interactive Explanations

The three explanation paradigms compared in this study are purely textual. Visual explanations (attribute comparison charts, recommendation rationale visualizations, similarity maps) may produce qualitatively different user experiences than text-only formats. Interactive explanations — which allow users to adjust recommendation parameters and observe how outputs change — may be more effective than any static format for building genuine understanding and trust.

D. Individual Differences as Moderators

A persistent limitation in explanation research is the treatment of users as a homogeneous population. Substantial theoretical and empirical evidence suggests that individual differences — cognitive style (analytic vs. intuitive), product domain expertise, prior experience with recommender systems, need for cognition, and trust propensity — moderate explanation effectiveness. Future research should incorporate individual difference measures as predictors in mixed-effects models.

E. Cultural and Cross-Cultural Moderators

Explanation effectiveness is likely subject to cultural moderation that the existing literature has largely ignored. Social proof mechanisms may be differentially effective in collectivist versus individualist cultural contexts [25]. Cross-cultural comparative studies recruiting participants from systematically varied cultural backgrounds are needed to characterize these moderating effects.

F. Explanation Fidelity and Model Accuracy

An underexplored but consequential gap concerns the relationship between explanation fidelity — how accurately an explanation reflects the model's actual decision process — and user trust. Post-hoc explanation approaches, including

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504121|

ours, generate explanations independently of the model, creating the possibility that explanations misrepresent the true computational basis of recommendations.

G. Behavioural Outcomes Beyond Self-Report

This study, like the majority of the explanation evaluation literature, relies on self-reported Likert measures. Purchase completion rates, time-to-decision, recommendation acceptance rates, and post-purchase satisfaction provide complementary behavioural indicators that self-report measures cannot access. Field experiments embedded in live recommendation systems, using A/B testing of explanation styles across randomly assigned user segments, would provide ecologically valid behavioural outcome data.

VII. CONCLUSION

This study presented a rigorous empirical evaluation of three prominent explanation paradigms — feature-based, social proof, and counterfactual — in the context of e-commerce product recommendations, grounded in a production-grade Neural Collaborative Filtering infrastructure trained on 1.69 million interactions from the Amazon Electronics dataset.

Our findings suggest that, within the context of generic and non-personalized explanations, the presence of an explanation may influence user perception more strongly than the specific explanation style employed. Methodologically, the comparison between our preliminary (n=16, spurious effects) and full (N=70, null effects) studies illustrates the critical importance of adequate statistical power in XAI evaluation research.

The research gap analysis presented in Section VI identifies seven priority directions: personalized explanation generation, longitudinal trust dynamics, multi-modal and interactive formats, individual difference moderators, cultural factors, explanation fidelity, and behavioural outcome validation. Progress on these fronts requires interdisciplinary collaboration spanning recommender systems, human-computer interaction, cognitive psychology, and decision science.

The conclusion that explanation presence matters more than style in low-involvement contexts, if replicated and extended, has practical implications for e-commerce platform design: investment in generating any high-quality explanation may yield greater trust returns than investment in selecting among explanation styles.

VIII. ACKNOWLEDGMENT

The authors thank all 70 study participants for their time and thoughtful responses. The authors also acknowledge SIES College of Arts, Science & Commerce for providing computational resources that enabled model training at scale. We thank the anonymous reviewers for constructive feedback on earlier drafts of this manuscript.

REFERENCES

[1] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," IEEE Internet Computing, vol. 7, no. 1, pp. 76–80, 2003.
[2] F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: Introduction and challenges," in Recommender Systems Handbook, 2nd ed. Springer, 2015, pp. 1–34.
[3] N. Tintarev and J. Masthoff, "Explaining recommendations: Design and evaluation," in Recommender Systems Handbook, 2nd ed. Springer, 2015, pp. 353–382.
[4] R. Sinha and K. Swearingen, "The role of transparency in recommender systems," in Proc. CHI Extended Abstracts, 2002, pp. 830–831.
[5] Y. Zhang and X. Chen, "Explainable recommendation: A survey and new perspectives," Found. Trends Inf. Retr., vol. 14, no. 1, pp. 1–101, 2020.
[6] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?" arXiv:1712.09923, 2017.
[7] N. Tintarev and J. Masthoff, "A survey of explanations in recommender systems," in Proc. IEEE 23rd Int. Conf. Data Engineering Workshop, 2007, pp. 801–810.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504121|

[8] Y. Zhang and X. Chen, "Explainable recommendation: A survey and new perspectives," Found. Trends Inf. Retr., vol. 14, no. 1, pp. 1–101, 2020.
[9] 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, 2004.
[10] P. Pu, L. Chen, and R. Hu, "A user-centric evaluation framework for recommender systems," in Proc. ACM RecSys, 2011, pp. 157–164.
[11] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell, "Explaining the user experience of recommender systems," User Model. User-Adapt. Interact., vol. 22, no. 4–5, pp. 441–504, 2012.
[12] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," Harvard J. Law & Tech., vol. 31, no. 2, pp. 841–887, 2018.
[13] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artif. Intell., vol. 267, pp. 1–38, 2019.
[14] F. Mohammadi, S. Deldjoo, and others, "Beyond top 1: Evaluating counterfactual explanations for recommender systems," in Proc. ACM RecSys, 2025.
[15] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in Proc. WWW, 2017, pp. 173–182.
[16] M. Bilgic and R. J. Mooney, "Explaining recommendations: Satisfaction vs. promotion," in Proc. Beyond Personalization Workshop, 2005.
[17] R. E. Petty and J. T. Cacioppo, "The elaboration likelihood model of persuasion," Adv. Exp. Soc. Psychol., vol. 19, pp. 123–205, 1986.
[18] R. B. Cialdini, Influence: The Psychology of Persuasion. HarperCollins, 2006.
[19] D. Kahneman and A. Tversky, "The simulation heuristic," in Judgment Under Uncertainty: Heuristics and Biases, Cambridge Univ. Press, 1982, pp. 201–208.
[20] R. E. Petty and J. T. Cacioppo, Communication and Persuasion: Central and Peripheral Routes to Attitude Change. Springer, 1986.
[21] X. Chen, H. Xu, Y. Zhang, J. Tang, Y. Cao, Z. Qin, and H. Zha, "Sequential recommendation with user memory networks," in Proc. WSDM, 2018.
[22] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, "Neural graph collaborative filtering," in Proc. SIGIR, 2019, pp. 165–174.
[23] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labelled reviews and fine-grained aspects," in Proc. EMNLP-IJCNLP, 2019, pp. 188–197.
[24] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err," J. Exp. Psychol. Gen., vol. 144, no. 1, pp. 114–126, 2015.
[25] G. Hofstede, Culture's Consequences: Comparing Values, Behaviour's, Institutions, and Organizations Across Nations, 2nd ed. Sage, 2001.
[26] J. Chang, D. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in Proc. NIPS, 2009, pp. 288–296.

ABOUT THE AUTHOR

Mohd Hussain Khan is an M.Sc. Data Science student in the Department of Data Science at SIES College of Arts, Science & Commerce (Empowered Autonomous), Mumbai, India. His research interests include data science, artificial intelligence, machine learning, deep learning, and explainable recommender systems, with a focus on user-centric AI evaluation. This manuscript represents his first research contribution in the field of explainable AI and intelligent recommendation systems. He can be contacted at mohdhussain2526128@ascs.sies.edu.in .

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details