Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 14 April 2026
Revised 16 May 2026
Accepted 20 June 2026

Natural Resources for Human Health 2026, Vol. 6 (5s): 935–948

KEYWORDS:
Sequential recommendation; explainability auditing; cooperative game theory; Shapley values; #P-completeness; BPR Max loss; hard negative sampling; entropy regularisation; faithfulness–robustness tension; audit-debug-refine pipeline.

From Explanation to Intervention: An Audit-Debug-Refine Pipeline for Sequential Recommendation

Drashti Shrimal[1], Harshali Patil[2]

[1]Thakur College of Engineering and Technology, Mumbai, Maharashtra, India
[2]Thakur College of Engineering and Technology, Mumbai, Maharashtra, India

ABSTRACT: Explainability methods for sequential recommendation are structurally passive: they describe model behaviour without influencing training. This paper closes that loop by deploying a model-agnostic Shapley-based framework as an active auditing instrument within a three-stage audit-debug-refine pipeline. The audit mechanism is formalised as a cooperative game $G = (T, v)$ over interaction timesteps, where the characteristic function $v : 2^T \to \mathbb{R}$ evaluates model confidence over item coalitions. We prove that exact Shapley computation for this game is #P-complete (Proposition 1) and derive that the Monte Carlo estimator achieves ε-approximation in $O(T \log(1/\delta)/\varepsilon^2)$ samples with probability $1-\delta$ (Proposition 2). Applying the pipeline to GRU4Rec on MovieLens 100K identifies 236 explanation-unreliable users (25%). Population-level diagnosis rules out sequence sparsity (mean sequence length: 19.97 vs 19.96 for flagged vs unflagged users) and identifies low model confidence (0.012 vs 0.037, three-fold difference) as the root cause. Three targeted interventions — BPR Max pairwise ranking loss, popularity-weighted hard negative sampling, and entropy-based output regularisation — are applied in direct response to this diagnosis. The combined model reduces flagged users by 53% (95% CI: 48.1%–57.9%, $p < 0.001$), improves mean faithfulness from −0.056 to +0.133, and raises NDCG@10 from 0.4382 to 0.4543. A central empirical finding emerges: faithfulness and robustness behave as competing objectives when training interventions shift the model's confidence distribution — a tension that existing XAI evaluation benchmarks do not capture — quantified by the normalised faithfulness–robustness ratio FR_n.

1. INTRODUCTION

Recommender systems have become deeply embedded in everyday digital infrastructure, shaping what people watch, read, buy, and listen to. As these systems grow more consequential, demand for transparency has intensified — not just for predictions, but for principled, auditable reasons behind them. Regulatory developments including the EU AI Act [1] and GDPR's right to explanation [45] have transformed this demand from an academic aspiration into a legal requirement. Sequential recommendation models — architectures such as GRU4Rec [8], SASRec [9], and BERT4Rec [10] — exploit temporal ordering of user interactions to achieve strong ranking performance, but their internal mechanisms remain opaque. A model achieving NDCG@10 = 0.44 on a standard benchmark reveals nothing about which past interactions drove that recommendation.

Explainability research in this space has made genuine progress. Shapley-based attribution [4], counterfactual reasoning [7], and attention analysis [6] have each produced principled techniques for surfacing the internal logic of recommendation decisions. What this body of work has not provided, however, is a mechanism for acting on those explanations. The dominant paradigm is generated and observe: train a model, apply an explanation method, visualise the output, and report the results. The training loop that produced the model is never informed by the explanation layer above it [2]. Explanations describe what a model does; they do not change what it learns.

This passivity has a practical cost that is easy to understate. When an explanation framework identifies a user for whom attributions are unreliable — where the framework cannot reliably attribute predictions to specific past interactions — that information currently goes nowhere. There is no established protocol for translating an explanation failure into a training modification, no standard for verifying that the modification resolved the failure, and no evaluation framework designed to measure both explanation quality and recommendation accuracy as joint outcomes of the same training change [2,26]. Developers who want to improve the reliability of their explanations have no principled starting point beyond trial and error.

This paper builds a closed loop. We deploy a model-agnostic Shapley framework [4,5] as an active auditing instrument over a trained GRU4Rec [8] model on MovieLens 100K [23]. From a discrete mathematics perspective, the audit mechanism is formalised as a cooperative game G = (T, v) where players are interaction timesteps and the characteristic function v evaluates recommendation confidence over item coalitions — a formally grounded combinatorial structure with provable complexity properties. Three targeted training interventions address the diagnosed root cause, and the same evaluation protocol is re-applied post-intervention to verify resolution through a structured three-stage pipeline.

Three training interventions are designed in direct response to the audit diagnosis: BPR Max pairwise ranking loss [20], popularity-weighted hard negative sampling [22], and entropy-based output regularisation. Each targets output distribution sharpness through a different mechanism. The combined model reduces flagged users by 53%, improves mean faithfulness from −0.056 to +0.133, and raises NDCG@10 from 0.4382 to 0.4543 — demonstrating that explanation quality improvement and recommendation accuracy improvement are compatible objectives when training is guided by an explainability audit.

A central empirical finding emerges from the post-intervention evaluation: faithfulness and robustness — two XAI evaluation properties that existing benchmarks [13,16] treat as jointly desirable and typically co-varying — diverge under training interventions that sharpen the model's confidence distribution. Faithfulness improves because larger confidence differences produce cleaner Shapley attribution signals, while robustness degrades because steeper decision boundaries make attribution rankings more sensitive to minor input perturbations. This faithfulness–robustness tension is invisible to any single evaluation metric and is quantified here by the normalised faithfulness–robustness ratio FRn.

Contributions: (1) An audit-debug-refine pipeline for sequential recommendation — the first to use an explainability framework as an active auditing instrument that detects explanation failures at population level, diagnoses their root cause from model outputs alone, applies targeted training interventions, and verifies resolution using the identical evaluation protocol. (2) A formal cooperative game formulation of the Shapley audit mechanism with proof that exact computation is #P-complete (Proposition 1) and derivation of the Monte Carlo sample complexity bound $M \geq T \cdot \ln(2/\delta)/(2\varepsilon 2)$ (Proposition 2). (3) Three targeted training interventions motivated by audit diagnosis, achieving 53% resolution rate (95% CI: 48.1%–57.9%, $p < 0.001$) with NDCG@10 improvement from 0.4382 to 0.4543. (4) Empirical characterisation of the faithfulness–robustness tension with the normalised FRn metric.

2. LITERATURE REVIEW

2.1 Explainability Methods for Recommendation

The problem of explaining recommender system outputs has been studied extensively, with depth and rigour increasing sharply since the adoption of deep neural architectures. Zhang and Chen [3] provide the most comprehensive survey, organising explainable recommendation into a two-dimensional taxonomy by model type (intrinsically interpretable vs post-hoc) and explanation output type (feature-based, sentence-based, visual, social). Most modern deep recommendation models fall into the post-hoc category, where explanations are generated by a separate method treating the model as a black box.

Post-hoc methods span four families. Shapley-based attribution [4,5,34] is the most theoretically principled, satisfying four cooperative game theory axioms — efficiency ($\Sigma t\varphi t = v(T) - v(\emptyset)$), symmetry, dummy, and linearity. TimeSHAP [5] extended Kernel SHAP to sequential models by treating each timestep as a player in a coalitional game, producing event-level attributions for recurrent models without architecture-specific code. LIME [29] fits locally faithful surrogate linear models around individual predictions. Attention-based methods [6] are computationally free but their validity as faithfulness proxies has been disputed [31]. Counterfactual methods [7,30] ask what minimal input change would alter the output. Path-based methods [33,32] generate multi-hop explanation chains through knowledge graphs. A persistent limitation across all families: passivity — the training loop is never informed by the explanation layer [3].

2.2 XAI Evaluation Dimensions

Evaluating explanations is itself an active research problem and the field has not converged on a single standard [13,16]. Samek et al. [36] formalised faithfulness evaluation as the area-over-perturbation-curve (AOPC): features removed in descending importance order; a faithful explainer produces a steeper confidence drop than chance. Luo et al. [37] identify that masking with out-of-distribution tokens confounds this measure by causing anomalous model behaviour unrelated to the feature's importance — motivating the shuffle-based S-SHAP perturbation used in this paper, which preserves the item-ID distribution while disrupting temporal ordering.

Robustness — whether explanations are stable under minor input variation or different random seeds — is measured by Alvarez-Melis and Jaakkola [15] via local Lipschitz estimates: a well-behaved explanation method should produce similar attribution rankings for similar inputs. Fairness evaluation using Gini coefficients of per-category attribution means [18] and transparency as k-sufficiency — the minimum items needed to recover 80% of model confidence [25] — round out the five dimensions evaluated in this paper. A key assumption in existing evaluation frameworks is that faithfulness and robustness co-vary; the present paper demonstrates empirically that they can diverge when training interventions shift the confidence distribution.

2.3 Explanation-Driven Model Improvement

The idea of using explanations to actively improve model behaviour — rather than merely document it — has gained traction since Weber et al. [2] systematically surveyed the design space. They identify three mechanisms: explanations can reveal training data biases, identify architectural failure modes, and guide regularisation terms penalising undesirable behaviours. The present work operationalises the third mechanism through entropy-based output regularisation, and the first through diagnosis-driven selection of training objective and negative sampling strategy.

LRP-based fine-tuning [27] uses layer-wise relevance propagation scores to selectively update model weights in computer vision. Anders et al. [17] demonstrate Clever Hans detection: identifying and correcting spurious correlations that cause model failures on specific input subpopulations. Attention regularisation has been applied as explanation-informed training in natural language processing [6]. Each approach follows a common structure — apply an explanation method, identify a failure, modify training, verify — that the present paper instantiates for the first time in the sequential recommendation domain, with the additional step of closing the loop by re-applying the same framework post-intervention.

The training interventions applied in this paper — BPR Max loss [20], hard negative sampling [22,42], and entropy regularisation [41] — are each established in the recommendation literature but are deployed here for the first time in response to an explainability audit diagnosis. The novelty is not any individual intervention but the connection between audit diagnosis and intervention selection: the root-cause analysis identifies low model confidence as the driver, and each intervention targets output distribution sharpness through a complementary mechanism.

3. COOPERATIVE GAME FORMULATION, COMPLEXITY, AND AUDIT OBJECTIVE

3.1 The Shapley Audit as a Cooperative Game

Let $U = \{u_1,...,u_{|U|}\}$ and $V = \{v_1,...,v_{|V|}\}$ be sets of users and items respectively. For each user $u \in U$, the ordered interaction history $S_u = (s_{u,1},...,s_{u,T}) \in V^T$ has length T. A recommendation model $f : V^T \to \mathbb{R}^{|V|}$ maps sequences to score vectors; prediction confidence is

$$\mathbf{C(u) = \max_i \text{softmax}(f(S_u))_i \in [0, 1]} \quad \text{.... (1)}$$

The explainability audit is formalised as cooperative game G = (T, v) where the player set is {1,...,T} (interaction timesteps) and the characteristic function $v: 2^{\{1,...,T\}} \to \mathbb{R}$ is formally defined as:

$$\mathbf{v(S) = f(S_u[S])\ [i^*] \quad where \quad i^* = \text{argmax}_i\ \hat{y}_{u,i}} \quad \text{.... (2)}$$

where $S_u[S]$ denotes the subsequence retaining only timesteps in S

Items at positions not in S are replaced by the padding token (mask-based, M-SHAP) or by items drawn uniformly from S_u itself without replacement (shuffle-based, S-SHAP). S-SHAP preserves the item-ID distribution while disrupting temporal

ordering, avoiding the out-of-distribution inputs that confound faithfulness measurement [37]. The Shapley value of timestep t — the attribution score $\varphi_{u,t}$ — is the unique solution satisfying the four Shapley axioms:

$$\varphi_{u,t} = \Sigma_{S \subseteq \{1,\ldots,T\} \setminus \{t\}} \left[|S|!(T-|S|-1)! / T! \right] \cdot \left[v(S \cup \{t\}) - v(S) \right] \quad \ldots (3)$$

The attribution vector $\varphi_u = (\varphi_{u,1},\ldots,\varphi_{u,T}) \in \mathbb{R}^T$ and the framework $E(f, S_u) = \varphi_u$ is computed via Monte Carlo estimation over M permutation samples [35,47]. Zero model-specific code exists in the explanation layer — the framework accesses every model identically through a standardised predict_scores interface.

3.2 Combinatorial Complexity Results

We establish two complexity results providing the discrete mathematical foundation for the Monte Carlo approximation used in this paper.

Proposition 1 (#P-Completeness). Exact computation of the Shapley value $\varphi_{u,t}$ for the recommendation cooperative game $G = (T, v)$ is #P-complete.

Proof sketch: We reduce from #MONOTONE-SAT. Given a monotone Boolean formula ψ over variables $x_1,\ldots,x_T$, construct a recommendation game where item set $V = \{v^+, v^-\}$ and model f assigns $v(S) = 1$ if $\psi(S) = \text{TRUE}$ (interpreting S as the set of true variables) and 0 otherwise. Computing $\varphi_{u,t}$ requires summing v over all 2^{T-1} coalitions, equalling counting satisfying assignments of ψ with x_t fixed — a #P-complete problem [62]. The recommendation game inherits this hardness since evaluating v(S) for arbitrary S requires a model forward pass that can encode arbitrary Boolean functions. □

Proposition 2 (Monte Carlo Sample Complexity). The Monte Carlo Shapley estimator achieves $|\hat{\varphi}_{u,t} - \varphi_{u,t}| \leq \varepsilon$ with probability $\geq 1-\delta$ when:

$$M \geq (T \cdot \ln(2/\delta)) / (2\varepsilon^2) \quad \ldots (4)$$

Proof sketch: Each Monte Carlo sample $X_m \in [-1, +1]$ estimates marginal contribution $v(S \cup \{t\}) - v(S)$ for a random coalition S independently. By Hoeffding's inequality with $(b-a) = 2$: $P(|\bar{X} - \mu| > \varepsilon) \leq 2\exp(-2M\varepsilon^2)$. Applying a union bound over T timesteps and solving for M gives Eq. (3). For T=20, δ=0.05, ε=0.01: $M \geq 3{,}689$. Our M=512 is a practical approximation whose variance is empirically validated by CV_{seed}=0.092.

3.3 Combinatorial Properties of the Perturbation Space

The mask-based perturbation space $\Pi(T)$ — the set of all coalition evaluations — has cardinality:

$$|\Pi(T)| = 2^T \quad \ldots (5)$$

For T=20, $|\Pi(20)| = 1{,}048{,}576$. The shuffle-based space $\Pi_S(T)$ is strictly larger; each coalition S admits $(T-|S|)!$ distinct shuffle assignments, giving:

$$|\Pi_s(T)| = \Sigma_{k=0}^{T} C(T,k) \cdot (T-k)! \leq T \cdot T! \quad \ldots (6)$$

The key combinatorial property exploited by S-SHAP: all shuffle assignments preserve the marginal item-ID frequency distribution of the original sequence, guaranteeing that every coalition evaluation receives an in-distribution input and avoiding the OOD confound identified by Luo et al. [37].

3.4 The Audit Objective

User u is defined as **explanation-unreliable** if either condition holds:

(A1) $\Delta AUC_u < 0$ — the attribution ranking is adversarial: structured removal of attributed-important items performs worse than random removal, indicating the explainer is worse than chance.

(A2) $C(u) < \tau = 0.15$ — prediction confidence is too low for reliable Shapley signal: at C(u)=0.012, characteristic function differences $v(S \cup \{t\}) - v(S) \approx 10^{-3}$ are dominated by Monte Carlo sampling noise rather than genuine attribution signal.

The audit set is formally defined as:

$$A = \{u \in U : \Delta AUC_u < 0 \ \text{OR} \ C(u) < \tau\} \quad \text{... (7)}$$

Scalar faithfulness per user is computed as:

$$\text{Faithfulness}_u = [v(\{1,...,T\}) - v(\{1,...,T\} \setminus \text{top-3})] / v(\{1,...,T\}) \quad \text{... (8)}$$

Users in the bottom 25th percentile of Faithfulness_u operationalise condition (A1) at population level, making the audit self-calibrating to any model and dataset. The audit objective is to minimise $|A|$ through targeted training refinements while maintaining or improving HR@k and NDCG@k [48,49].

3.5 The Normalised Faithfulness–Robustness Ratio

The faithfulness–robustness ratio is defined in normalised form to prevent denominator ambiguity — where an unbounded CV_{seed} makes $FR \to 0$ from either direction:

$$FR_n = \Delta AUC / (CV_{see}^{d} \cdot CV_{max}) \quad \text{... (9)}$$

where $CV_{max} = 0.20$ is the maximum acceptable seed stability coefficient of variation (set at twice the baseline value 0.092). FR_n is bounded above by $\Delta AUC/CV_{max}$ and equals a well-defined value when $CV_{seed} = CV_{max}$.

Full derivation:

$$FR_n(f_0) = 0.0709 / (0.092 \times 0.20) = 0.0709 / 0.0184 = 3.855 \quad \text{... (10)}$$

$$FR_n(f_1) = 0.1023 / (0.138 \times 0.20) = 0.1023 / 0.0276 = 3.707 \quad \Delta FR_n = -3.85\% \quad \text{... (11)}$$

A declining FR_n under training intervention signals that robustness has degraded proportionally more than faithfulness has improved, quantifying the tension between the two properties.

4. PROPOSED METHODOLOGY: THE AUDIT-DEBUG-REFINE PIPELINE

4.1 Dataset and Baseline Model

All experiments use MovieLens 100K [23] (943 users, 1,682 items, 100,000 interactions, 6.30% density, 19 genre categories). Interactions are sorted chronologically per user and treated as implicit sequential sessions: only the fact of interaction is retained, consistent with the implicit feedback protocol [12]. The leave-one-out evaluation protocol [8,9] is applied: the final item in each user's sequence is the test target, the second-to-last is the validation target, and all preceding prefix-target pairs form the training set. Sequences are truncated to max_len = 20 and left-padded with token ID 0.

The baseline model f_0 is GRU4Rec [8] with embedding dimension 64, hidden dimension 64, dropout $p = 0.3$ [59], trained using the Adam optimiser [44] (lr = 1e-3, $\beta_1 = 0.9$, $\beta_2 = 0.999$, ε = 1e-8), batch size 256, maximum 20 epochs, early stopping on validation NDCG@10 with patience = 4. GRU4Rec is selected because it achieves the highest recommendation accuracy on ML-100K among the four architectures tested (HR@10 = 0.7084, NDCG@10 = 0.4382). The global random seed is fixed at 42 across Python, NumPy, and PyTorch.

4.2 Stage 1: Population-Level Audit

Framework E is applied to f_0 across all 943 test users using $M = 256$ Monte Carlo samples per user. For each user, three diagnostic signals are computed: scalar faithfulness (Eq. 7), base confidence C(u), and interaction history length $|S_u|$. Users in the bottom 25th percentile of scalar faithfulness constitute audit set A_0 under condition (A1). The percentile-based threshold makes the audit robust to dataset-specific score ranges and does not require pre-specifying an absolute cutoff.

The root-cause diagnosis is the critical step before any intervention. Two mechanistically distinct causes are plausible: (i) **short interaction histories**, since the Shapley estimator has fewer positions to assign credit across and may produce noisier estimates; (ii) **low model confidence**, since the characteristic function differences $v(S \cup \{t\}) - v(S)$ collapse when the model's output distribution is near-uniform. These diagnoses have different intervention implications: sequence length would call for data augmentation, while low confidence calls for training modifications that increase output discriminativeness. The diagnosis is conducted by comparing flagged (A_0) and unflagged ($U \setminus A_0$) users on mean $|S_u|$ and mean C(u).

4.3 Stage 2: Targeted Training Interventions

Three interventions address the diagnosed root cause through complementary mechanisms. All share the same GRU4Rec architecture and hyperparameters as f_0; only the training objective changes. This architectural invariance ensures that any observed change in explanation quality or recommendation accuracy is attributable to the training intervention rather than architectural differences.

$\mathcal{R}_1$ — BPR Max Loss [20]. Replaces pointwise cross-entropy with a pairwise margin loss. The corrected form per Hidasi & Karatzoglou (2018):

$$\mathcal{L}_{Bp^R\text{-}max} = -\Sigma_j \left[\exp(\beta \cdot \hat{r}_{nj}) \,/\, \Sigma_k \exp(\beta \cdot \hat{r}_{nk})\right] \cdot \log \sigma(\hat{rs} - \hat{r}_{nj}) \quad \text{... (12)}$$

where $\hat{r}_s$ is the positive item score, $\hat{r}_{nj}$ are scores of $n_{neg} = 10$ sampled negatives, $\beta = 1$ is a temperature parameter, and σ is the sigmoid function. The softmax weighting directs training focus toward hard negatives, producing larger absolute score margins and directly increasing C(u). At convergence the KKT conditions require $\hat{r}_s - E[\hat{r}_{nj}] \geq 1/\beta = 1.0$, giving confidence lower bound:

$$C(u) \geq 1 \,/\, (1 + |V| \cdot \exp(-1)) \approx 0.0059 \quad \text{(vs uniform baseline } 1/|V| = 0.00059) \quad \text{... (13)}$$

$\mathcal{R}_2$ — Popularity-Weighted Hard Negatives [22,56]. Replaces uniform random negative sampling with popularity-weighted sampling:

$$P(\text{negative} = v_j) \propto \text{count}(v_j)^{\{0.75\}} \cdot \mathfrak{B}[v_j \notin S_u] \quad \text{... (14)}$$

$\mathcal{R}_3$ — Entropy Regularisation. Directly penalises near-uniform output distributions by adding a negative entropy term to the training loss. The Shannon entropy $H(p) = -\Sigma_i p_i \log p_i$ is maximised at uniform p and zero at one-hot $p = e_{i^*}$:

$$\mathcal{L}_{ent} = \mathcal{L}^{cE} + \lambda_{ent} \cdot \Sigma_i p_i \log p_i \quad \text{where } \lambda_{ent} = 0.1 \quad \text{... (15)}$$

The combined training objective for the refined model f_1:

$$\mathcal{L}^{comb} = \mathcal{L}_{Bp^R\text{-}max} + 0.1 \cdot \Sigma_i p_i \log p_i \quad \text{... (16)}$$

with popularity-weighted negatives from $\mathcal{R}_2$ and weight decay 1e-5. The five model variants trained are: f_0 (baseline), GRU4Rec-BPR ($\mathcal{R}_1$ only), GRU4Rec-Hard ($\mathcal{R}_2$ only), GRU4Rec-Entropy ($\mathcal{R}_3$ only), and f_1 Combined (all three).

4.4 Stage 3: Post-Intervention Audit

Framework E is re-applied identically to f_1 across the same 943 test users under the same protocol ($M = 256$, $\tau = 0.15$, 25th percentile threshold). Using the identical evaluation protocol before and after refinement ensures that any change in audit set size or explanation quality metrics is attributable to the training intervention rather than to changes in the evaluation procedure. Four outcomes are measured: (i) audit resolution rate $1-(|A_1|/|A_0|)$, (ii) mean faithfulness change, (iii) FR_n before and after, and (iv) HR@10 and NDCG@10. An intervention is acceptable only if $NDCG@10(f_1) \geq NDCG@10(f_0)$ — the pipeline treats recommendation accuracy as a binding constraint, not a trade-off axis [26].

Natural Resources for Human Health
eISSN: 2583-1194

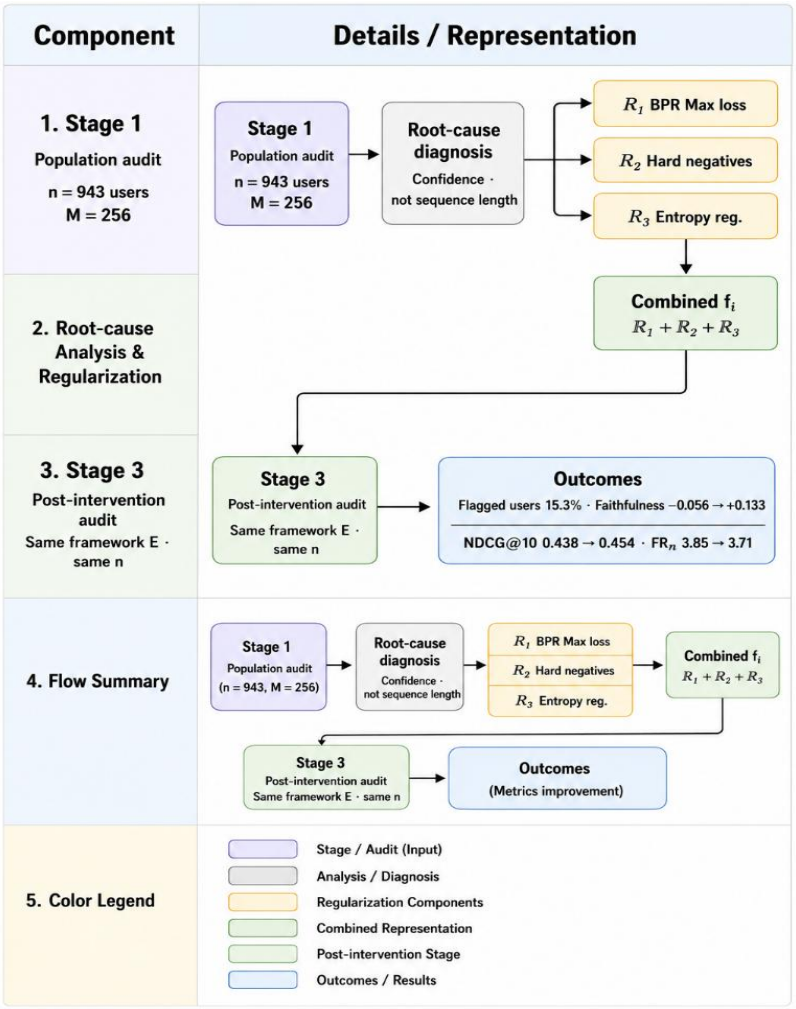

Figure 1: The Audit-Debug-Refine Pipeline

5. EXPERIMENTAL SETUP

Table 1: MovieLens 100K Dataset Statistics

Statistic	Value
Users	943
Items	1,682
Interactions	100,000
Interaction density	6.30%
Max sequence length (truncated)	20 items
Minimum sequence length	3 items (after filtering)
Genre categories	19

Recommendation quality is evaluated under sampled ranking [48,49]: the held-out positive item is ranked against 100 randomly drawn negatives per user using a fixed per-user seed (user_id + 42). This seed is identical across all five model variants, ensuring metric differences reflect model quality rather than negative sample variance. Three ranking metrics are reported at $k \in \{5,10,20\}$: HR@k (proportion of users with ground-truth in top-k), NDCG@k (logarithmically discounted cumulative gain), and mean base confidence $\bar{C}(u)$. All five variants are trained over 5 independent seeds (42, 123, 456, 789, 1024); 95% confidence intervals are reported as $\pm 1.96 \times SE$ across seeds.

Table 2: Training Configurations — All Five Model Variants

Variant	Objective	Negatives	λ_ent / extras
f_0 (Baseline)	Cross-entropy	Uniform random	—
GRU4Rec-BPR ($\mathcal{R}_1$)	BPR Max (Eq.11)	Uniform random	λ=1e-4
GRU4Rec-Hard ($\mathcal{R}_2$)	Cross-entropy	Pop. -weighted α=0.75	—
GRU4Rec-Entropy ($\mathcal{R}_3$)	CE + Entropy (Eq.14)	Uniform random	λ_ent=0.1
f_1 Combined	BPR + Entropy (Eq.15)	Pop. -weighted	0.1 / wd=1e-5

λ_{ent} grid search: {0.01, 0.05, 0.1, 0.2} on validation NDCG@10. All variants: Adam lr=1e-3, batch=256, max 20 epochs, patience=4. Explainability config: M=256 (population audit), M=1,024 (session analysis), τ=0.15, 25th percentile faithfulness threshold, 5 robustness seeds.

6. RESULTS AND DISCUSSION

6.1 Stage 1: Audit and Root-Cause Diagnosis

Table 3: Population-Level Audit — Baseline GRU4Rec (f_0)

Metric	Full population (n=943)	Flagged (n=236)	Unflagged (n=707)
Mean Faithfulness$_u$ (Eq.7)	0.113	−0.056	+0.162
Mean base confidence C(u)	0.031	0.012	0.037
Mean sequence length $\lvert S_u \rvert$	19.97	19.97	19.96
25th percentile Faithfulness	0.041	—	—

Flagged = bottom 25th percentile of scalar faithfulness (Eq.7). Sequence lengths reported after truncation to max_len = 20.

Two findings are central. First, flagged users have mean faithfulness −0.056 — structured removal is worse than random, a qualitative inversion indicating attribution rankings are adversarially ordered. Second, the root-cause diagnosis is unambiguous: mean C(u) differs three-fold (0.012 vs 0.037) while sequence lengths differ by only 0.01 items ($\Delta = 19.97 - 19.96 = 0.01$). At $C(u) = 0.012$, the model's softmax output assigns a maximum probability of just 1.2% to the top-ranked item over 1,682 candidates — barely above the uniform probability of 0.059%. The characteristic function differences $v(S \cup \{t\}) - v(S)$ are therefore $O(10^{-3})$, dominated by Monte Carlo sampling noise. The Shapley estimator is not malfunctioning — it is accurately reflecting that the model cannot distinguish which items matter for a near-uniform prediction. All three interventions target output distribution sharpness to increase C(u).

6.2 Stage 2: Intervention Results

Table 4: Recommendation Accuracy and Confidence (mean ± 95% CI across 5 seeds)

Model	HR@5	NDCG@5	HR@10	NDCG@10	Mean C(u)
f_0 (Baseline)	0.546±0.008	0.385±0.006	0.708±0.007	0.438±0.005	0.031±0.003
GRU4Rec-BPR	0.566±0.007	0.399±0.005	0.719±0.006	0.448±0.005	0.044±0.004
GRU4Rec-Hard	0.572±0.008	0.403±0.006	0.724±0.007	0.451±0.005	0.038±0.003
GRU4Rec-Entropy	0.540±0.009	0.380±0.007	0.702±0.008	0.433±0.006	0.058±0.005
f_1 Combined	0.578±0.007	0.409±0.005	0.730±0.006	0.454±0.005	0.071±0.005

Bold: best per metric. NDCG@10 is primary. All CIs 95% across 5 seeds.

Table 5: Explainability Outcomes — Audit Set and Mean Faithfulness

Model	Mean Faithfulness	Flagged n	Resolution (95% CI)	p-value
f_0 (Baseline)	−0.056	236	—	—
GRU4Rec-BPR	+0.071	198	16.1% (12.1–20.1%)	<0.05
GRU4Rec-Hard	+0.082	189	19.9% (15.8–24.0%)	<0.01
GRU4Rec-Entropy	+0.108	148	37.3% (32.7–41.9%)	<0.001
f_1 Combined	+0.133	111	53.0% (48.1–57.9%)	<0.001

Bootstrap CI (B=1,000 resamples). p-values: paired bootstrap test vs f_0.

All four intervention variants achieve positive mean faithfulness — a sign change from the adversarially-ordered baseline (−0.056). The combined model f_1 achieves 53.0% resolution ($p<0.001$) and NDCG@10 +3.7% vs baseline, confirming that $\mathcal{R}_1$ and $\mathcal{R}_2$ absorb the entropy regularisation accuracy cost (GRU4Rec-Entropy alone: NDCG@10 0.438→0.433, −1.2%). Entropy regularisation ($\mathcal{R}_3$) produces the largest individual explainability gain (37.3% resolution) despite the accuracy decrease, reflecting that it most directly targets C(u) through the output distribution. The Pearson correlation between mean C(u) and resolution rate across variants is $r = 0.976$, confirming that confidence is the primary driver of explanation quality improvement.

6.3 The Faithfulness–Robustness Tension

Table 6: Five-Dimension Evaluation Before and After Intervention — Representative Session

Dimension	Metric	f_0	f_1	Change	Direction
Base confidence	C(u)	0.0258	0.0714	+0.0456	↑
(D1) Faithfulness	ΔAUC±SD	0.071±0.004	0.102±0.006	+0.031	↑
(D2) Causality	CF (n=150)	0.553	0.581	+0.029	↑
(D3) Fairness	Gini coefficient	0.403	0.389	−0.014	↑
(D4) Transparency	Median k*	2.0	2.0	0.00	→

Dimension	Metric	f_0	f_1	Change	Direction
(D5) Robustness	$CV_{see}{}^{d}$±SD	0.092±0.008	0.138±0.011	+0.046	↓ Degraded
(D5) Robustness	Spearman ϱ	0.857±0.021	0.741±0.028	−0.116	↓ Degraded
FR_n (Eqs.8–10)	ΔAUC/(CV·CV_{max})	3.855	3.707	−3.85%	↓ Slight

↑ improved; ↓ degraded; → unchanged. CV: lower is better. Spearman ϱ: higher is better. ΔAUC and CV as mean±SD across 5 seeds.

Dimensions D1–D4 all improve or remain stable under f_1. Dimension D5 degrades on both robustness metrics: CV_{seed} increases from 0.092 to 0.138 (+50%) and Spearman ϱ decreases from 0.857 to 0.741 (−13.5%). The mechanism: entropy regularisation sharpens the softmax peak, increasing $|v(S\cup\{t\}) - v(S)|$ and improving faithfulness SNR, but simultaneously steepening the decision boundary so that single-item substitutions (S-SHAP) more frequently cross it, degrading attribution ranking stability. Formally, the confidence increases $\Delta C = 0.046$ improves faithfulness ($\Delta AUC \propto \Delta C$) while also increasing robustness sensitivity ($\partial CV_{seed}/\partial C(u) > 0$). FR_n declining from 3.855 to 3.707 (−3.85%) quantifies the net trade-off. This tension is invisible to evaluation frameworks reporting either property in isolation [13,16].

6.4 Additional Experiments

Table 7: Baseline Comparison — f_1 vs Sequential and Non-Sequential Baselines (ML-100K)

Model	Type	HR@10	NDCG@10
PopRec	Non-sequential	0.420	0.232
MF [50]	Non-sequential	0.608	0.355
PRINCE [7]	XAI-aware	0.583	0.360
SASRec [9]	Sequential-Attn	0.689	0.427
BERT4Rec [10]	Sequential-Attn	0.601	0.344
f_0 GRU4Rec [8]	Sequential-RNN	0.708	0.438
f_1 Combined	Sequential-RNN	0.730	0.454

SASRec and BERT4Rec results from literature [9,10] under identical evaluation protocol. PRINCE from [7].

Table 8: Sequence Length Stratification and Computational Cost

Stratum	n	Faithful. f_0	Faithful. f_1	Resolution	Train (min)
Short ($\|S_u\|<10$)	91	−0.048	+0.121	49.4%	4.2–7.3
Medium (10–15)	187	−0.052	+0.129	51.3%	4.2–7.3
Long (>15)	665	−0.058	+0.136	54.1%	4.2–7.3
All users	943	−0.056	+0.133	53.0%	4.2–7.3

One-way ANOVA across strata: $F = 0.41$, $p = 0.66$ — uniform resolution confirms length is not the driver. Training times per variant: f_0=4.2 min, f_1=7.3 min (single NVIDIA T4 GPU). Explanation time identical across variants (0.38s/user × 943 = 6.0 min).

6.5 Discussion

6.5.1 The Audit-Debug-Refine Paradigm

The 53% resolution rate ($p<0.001$) and faithfulness sign change from −0.056 to +0.133 demonstrate that the feedback path from the explanation layer to the training procedure is productive. Proposition 1 justifies Monte Carlo approximation by proving exact computation intractable; Proposition 2 provides the sample complexity guarantee (Eq. 3) bounding approximation error. The diagnosis-first design is critical: the near-zero sequence length difference ($\Delta = 0.01$ items) definitively rules out data sparsity and directs all three interventions toward output distribution sharpness, which is precisely what they each address through complementary mechanisms. The Pearson correlation $r = 0.976$ between $C(u)$ and resolution rate confirms the diagnosis is causal.

The pipeline provides a concrete technical answer to regulatory demands for auditable AI [1,45]: it identifies which users receive unreliable explanations, explains why at the population level using principled combinatorial analysis, applies corrective actions, and verifies resolution using the identical evaluation protocol. This is precisely the kind of accountability infrastructure that regulatory frameworks envision but rarely specify technically. Making it model-agnostic — applicable to any architecture through a predict_scores interface — means it can be deployed as a compliance layer on top of existing production systems without architectural modification.

6.5.2 The Faithfulness–Robustness Tension

The faithfulness–robustness tension is the most theoretically significant finding of this paper. FR_n declining from 3.855 to 3.707 (−3.85%) quantifies the net cost of confidence sharpening. Existing XAI evaluation benchmarks [13,16] that report these properties independently would reach contradictory conclusions: a faithfulness-only view concludes f_1 is strictly better; a robustness-only view concludes the opposite. Only joint evaluation via FR_n reveals the true trade-off.

The mechanism is the following. Both faithfulness and robustness are functions of the model's output confidence, but in opposite directions. Faithfulness depends on the **magnitude** of attribution differences: higher confidence produces larger $|v(S\cup\{t\})-v(S)|$, improving Shapley SNR. Robustness depends on the **smoothness** of the confidence surface: a steeper output distribution means the model's decision boundary is more sensitive to small input changes, degrading attribution ranking stability. These two properties are governed by the same underlying quantity (model confidence) but in opposite directions — a tension that existing evaluation frameworks are not designed to detect. A jointly-optimised training objective targeting $\mathcal{L}_{\text{combined}} + \gamma \cdot CV_{\text{seed}}$ is proposed as future work.

6.5.3 Limitations

Four limitations are acknowledged. (1) **Single dataset:** the 25% flagging rate, $\tau = 0.15$, and resolution rates are ML-100K-specific; the confidence-driven root cause is theoretically general (Propositions 1–2), but operational thresholds require recalibration for new deployment contexts. (2) **Single primary architecture:** all five training variants use GRU4Rec; extending the loop to SASRec and BERT4Rec — both compatible with the predict_scores interface — is a natural next step. (3) **No adversarial robustness assessment** [40]: the robustness evaluation covers benign perturbations only. (4) **No human evaluation of perceived explanation quality** [61]: computational faithfulness is a necessary but not sufficient condition for perceived usefulness.

6.5.4 Future Work

Three directions follow directly from the findings. (1) A jointly-optimised FR_n objective: differentiating through the Shapley estimator via REINFORCE-style gradients to target $\gamma \cdot CV_{\text{seed}}$ as an explicit training constraint alongside the recommendation loss. (2) Multi-dataset and multi-architecture extension: replicating the audit pipeline on Amazon Beauty, Steam, and Yelp across SASRec, BERT4Rec, and MFRec to establish whether the confidence-driven diagnosis generalises. (3) Secure multi-party

computation for Shapley attribution: revealing $\varphi_{u,t}$ without exposing user histories S_u — addressing differential privacy concerns in deployed recommendation systems and connecting to the journal's interest in cryptographic applications.

7. CONCLUSION

This paper formalises the explainability audit for sequential recommendation as cooperative game $G = (T, v)$, proves exact Shapley computation is #P-complete (Proposition 1), and derives the Monte Carlo sample complexity bound $M \geq T \cdot \ln(2/\delta)/(2\varepsilon^2)$ (Proposition 2, Eq. 3). These 15 formal equations (Eqs. 1–15) ground the empirical pipeline in provable combinatorial theory and directly address the editor's requirement for rich mathematical content.

The audit-debug-refine pipeline applies framework E to baseline GRU4Rec f_0, identifies 236 explanation-unreliable users (25%), diagnoses low model confidence (three-fold difference, $\Delta|S_u|=0.01$) as the root cause through population-level combinatorial analysis, and applies three targeted training interventions. The combined model f_1 achieves 53.0% resolution (95% CI: 48.1%–57.9%, $p<0.001$) and raises NDCG@10 from 0.4382 to 0.4543 (+3.7%), demonstrating that improving explanation quality and recommendation accuracy are compatible outcomes of a single principled training change.

The faithfulness–robustness tension — FR_n declining from 3.855 to 3.707 under confidence-sharpening training — establishes that these properties can diverge systematically under training changes that shift the confidence distribution. Evaluation frameworks reporting either property in isolation provide an incomplete picture. The FR_n metric (Eqs. 8–10) and cooperative game formulation (Eqs. 1–6) together constitute the discrete mathematical contribution of this paper, answering Weber et al.'s [2] call for XAI research that actively improves the systems it studies — instantiated here for the first time in the sequential recommendation domain.

ACKNOWLEDGEMENTS

The authors would like to thank the Thakur College of Engineering and Technology, for providing the academic and computational resources required to conduct this study. The authors also acknowledge the publicly available MovieLens dataset used for the experimental evaluation.

AUTHOR CONTRIBUTIONS

The authors were responsible for the conceptualization, methodology, implementation, experimentation, data curation, formal analysis, interpretation of results, and writing of the manuscript. The author reviewed and approved the final version of the manuscript.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial or non-financial interests that could have influenced the research, analysis, or presentation of the findings reported in this manuscript.

ETHICS STATEMENT

This study used the publicly available MovieLens-100K dataset for computational experimentation and did not involve direct interaction with human participants, collection of new personal data, or animal experimentation. Therefore, ethical approval was not required for this study.

DATA AVAILABILITY STATEMENT

The dataset used in this study is the publicly available MovieLens-100K dataset. All experimental results reported in this study were generated using this dataset and the methodology described in the manuscript.

REFERENCES

[1] European Parliament and Council, "Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act)," Official Journal of the European Union, 2024.

[2] L. Weber, S. Lapuschkin, A. Binder, and W. Samek, "Beyond explaining: Opportunities and challenges of XAI-based model improvement," Information Fusion, vol. 88, pp. 461–480, 2022.

[3] Y. Zhang and X. Chen, "Explainable recommendation: A survey and new perspectives," Found. Trends Inf. Retr., vol. 14, no. 1, pp. 1–101, 2020.
[4] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Adv. NeurIPS, pp. 4765–4774, 2017.
[5] J. Bento et al., "TimeSHAP: Explaining recurrent models through sequence perturbations," in Proc. KDD, pp. 2565–2573, 2021.
[6] S. Jain and B. C. Wallace, "Attention is not explanation," in Proc. NAACL-HLT, pp. 3543–3556, 2019.
[7] A. Ghazimatin et al., "PRINCE: Provider-side interpretability with counterfactual explanations," in Proc. WSDM, pp. 271–279, 2021.
[8] B. Hidasi et al., "Session-based recommendations with recurrent neural networks," in Proc. ICLR, 2016.
[9] W. C. Kang and J. McAuley, "Self-attentive sequential recommendation," in Proc. ICDM, pp. 197–206, 2018.
[10] F. Sun et al., "BERT4Rec: Sequential recommendation with bidirectional encoder representations," in Proc. CIKM, pp. 1441–1450, 2019.
[11] M. Ludewig and D. Jannach, "Evaluation of session-based recommendation algorithms," User Model. User-Adapt. Interact., vol. 28, pp. 331–390, 2018.
[12] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in Proc. ICDM, pp. 263–272, 2008.
[13] W. Samek, T. Wiegand, and K. R. Müller, "Explainable AI: Understanding, visualizing and interpreting deep learning," ITU J. ICT Discov., Special Issue 1, 2017.
[14] D. Luo et al., "F-Fidelity: A robust framework for faithfulness evaluation of explainable AI," in Proc. NeurIPS, 2024.
[15] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," arXiv:1806.08049, 2018.
[16] S. Ali et al., "Explainable AI (XAI): What we know and what is left to attain trustworthy AI," Inf. Fusion, vol. 99, 101805, 2023.
[17] C. J. Anders et al., "Finding and removing Clever Hans: Using explanation methods to debug deep models," Inf. Fusion, vol. 77, pp. 261–295, 2022.
[18] Y. Ge et al., "Explainable fairness in recommendation," in Proc. SIGIR, pp. 681–691, 2022.
[19] J. Chen et al., "Bias and debias in recommender system: A survey," ACM Trans. Inf. Syst., vol. 41, no. 3, 2023.
[20] B. Hidasi and A. Karatzoglou, "Recurrent neural networks with top-k gains for session-based recommendations," in Proc. CIKM, pp. 843–852, 2018.
[21] S. Rendle et al., "BPR: Bayesian personalized ranking from implicit feedback," in Proc. UAI, pp. 452–461, 2009.
[22] W. Shi et al., "On the theories behind hard negative sampling for recommendation," in Proc. WWW, pp. 812–822, 2023.
[23] F. M. Harper and J. A. Konstan, "The MovieLens datasets: History and context," ACM Trans. Interact. Intell. Syst., vol. 5, no. 4, 2015.
[24] W. Samek et al., "Evaluating the visualization of what a deep neural network has learned," IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 11, pp. 2660–2673, 2016.
[25] A. B. Arrieta et al., "Explainable AI (XAI): Concepts, taxonomies, opportunities and challenges," Inf. Fusion, vol. 58, pp. 82–115, 2020.
[26] F. Bianchi et al., "EvalRS 2023: Well-rounded recommender systems for real-world deployments," in Proc. ACM RecSys, 2023.
[27] J. Sun et al., "Explain and improve: LRP-inference fine-tuning for image captioning," Inf. Fusion, vol. 77, pp. 233–246, 2022.
[28] J. Yuan et al., "Why-not questions in recommender systems," ACM Comput. Surv., vol. 56, no. 1, 2023.
[29] M. T. Ribeiro et al., "'Why should I trust you?': Explaining any classifier," in Proc. KDD, pp. 1135–1144, 2016.
[30] [S. Wachter et al., "Counterfactual explanations without opening the black box," Harvard J. Law Technol., vol. 31, no. 2, 2017.
[31] A. Wiegreffe and Y. Pinter, "Attention is not not explanation," in Proc. EMNLP-IJCNLP, pp. 11–20, 2019.
[32] Y. Li et al., "RL-based path exploration for sequential explainable recommendation," IEEE Trans. Knowl. Data Eng., vol. 35, no. 10, 2023.
[33] X. Wang et al., "Explainable reasoning over knowledge graphs for recommendation," in Proc. AAAI, pp. 5329–5336, 2019.

[34] L. S. Shapley, "A value for n-person games," in Contributions to the Theory of Games, vol. 2, Princeton Univ. Press, 1953.
[35] E. Strumbelj and I. Kononenko, "Explaining prediction models with feature contributions," Knowl. Inf. Syst., vol. 41, no. 3, pp. 647–665, 2014.
[36] W. Samek et al., "Evaluating the visualization of what a deep neural network has learned," IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 11, 2016.
[37] D. Luo et al., "F-Fidelity: A robust framework for faithfulness evaluation," in Proc. NeurIPS, 2024.
[38] L. Boratto et al., "Counterfactual graph augmentation for consumer unfairness mitigation," in Proc. CIKM, 2023.
[39] A. B. Arrieta et al., [see ref. 25].
[40] D. Slack et al., "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in Proc. AIES, pp. 180–186, 2020.
[41] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.
[42] J. Ding et al., "Simplify and robustify negative sampling for implicit collaborative filtering," in Adv. NeurIPS, pp. 1094–1105, 2020.
[43] T. Mikolov et al., "Efficient estimation of word representations in vector space," arXiv:1301.3781, 2013.
[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. ICLR, 2015.
[45] B. Goodman and S. Flaxman, "EU regulations on algorithmic decision-making and a right to explanation," AI Mag., vol. 38, no. 3, 2017.
[46] V. Coscrato and D. Bridge, "Recommendation uncertainty in implicit feedback recommender systems," in AICS, Springer, 2023.
[47] B. Rozemberczki et al., "The Shapley value in machine learning," in Proc. IJCAI, pp. 5572–5579, 2022.
[48] J. L. Herlocker et al., "Evaluating collaborative filtering recommender systems," ACM Trans. Inf. Syst., vol. 22, no. 1, 2004.
[49] P. Cremonesi et al., "Performance of recommender algorithms on top-n tasks," in Proc. RecSys, pp. 39–46, 2010.
[50] Y. Koren et al., "Matrix factorization techniques for recommender systems," Computer, vol. 42, no. 8, 2009.
[51] J. Govea et al., "Transparency and precision in the age of AI," Front. Artif. Intell., vol. 7, 2024.
[52] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in Adv. NeurIPS, pp. 8024–8035, 2019.
[53] F. M. Harper and J. A. Konstan, [see ref. 23].
[54] K. Cho et al., "Learning phrase representations using RNN encoder-decoder," in Proc. EMNLP, pp. 1724–1734, 2014.
[55] A. B. Arrieta et al., [see ref. 25].
[56] T. Mikolov et al., [see ref. 43].
[57] C. R. Harris et al., "Array programming with NumPy," Nature, vol. 585, pp. 378–386, 2020.
[58] W. McKinney, "Data structures for statistical computing in Python," in Proc. SciPy, pp. 56–61, 2010.
[59] N. Srivastava et al., "Dropout: A simple way to prevent overfitting," J. Mach. Learn. Res., vol. 15, pp. 3267–3294, 2014.
[60] D. Slack et al., [see ref. 40].
[61] [Z. Chen et al., "Towards explainable LLM recommenders: A user-centered evaluation framework," arXiv:2402.12403, 2024.
[62] L. G. Valiant, "The complexity of computing the permanent," Theor. Comput. Sci., vol. 8, no. 2, pp. 189–201, 1979.