LURE: LLM-based Unbiasing and Rationale-grounded Evaluation/Data Generation for Recommender Systems

Anonymous

Abstract

Large language models (LLMs) show promise as evaluators of recommender systems, yet existing "LLM-as-Judge" paradigms largely overlook exposure bias, lack auditable rationales, and seldom demonstrate external validity to online impact. We introduce LURE, a unified framework comprising: (i) RecJudge-X, an exposure-aware, rationale-grounded judge that operates on matched item pairs and outputs verifiable evidence; (ii) ExpoSynth, an exposure-controllable synthesizer that fills long-tail data gaps under collaborative constraints; and (iii) Calib-Link, a monotonic calibration mapping from LLM-based judge scores to IPS proxies of online lift. LURE operationalizes causal unbiasing with inverse propensity scoring (IPS), enforces evidence alignment to curb hallucination, and provides a reproducible evaluate–synthesize–re-evaluate loop. On a MovieLens-1M prototype, LURE yields stable IPS-weighted pairwise agreement and long-tail augmentation with controllable exposure. We discuss broader impacts, limitations, and a path toward human-centered scientific evaluation at scale.

1 Introduction

Traditional offline metrics (e.g., HR/NDCG/AUC) depend on exposure distributions and negative sampling protocols, leading to blind spots in long-tail, cold-start, cross-domain, and user experience evaluation. Prior *LLM-as-Judge* often rely on subjective scoring without modeling exposure bias or externally validating to online impact. Real logs are sparse in the tail, limiting training and generalization. LURE aims at an exposure-aware, rationale-grounded judge paired with an exposure-controllable synthesizer to form a *evaluate–synthesize–re-evaluate* loop for fairer offline evaluation and tail-aware training.

2 Background and Method Overview

Exposure bias in recommenders. Logged interactions reflect a product of user preference and platform exposure. Offline metrics computed under the logged exposure distribution can be misaligned with counterfactual performance. IPS and doubly-robust (DR) estimators are standard tools for causal unbiasing in recommendation and bandits.

Overview. LURE comprises: (i) **RecJudge-X**, an exposure-aware LLM judge with matched pairs, IPS weighting, and evidence-aligned rationales; (ii) **ExpoSynth**, a constraint-driven synthesizer that targets a desired exposure curve; and (iii) **Calib-Link**, an isotonic (monotonic) regression g mapping judge scores to IPS proxies or online metrics.

2.1 RecJudge-X

Input representation. For each user u, we construct a history block (titles, genres, and meta bins) and candidate item blocks. The prompt enforces a strict JSON schema with a field `winner` $\in \{i, j, tie\}$, $confidence p \in [0, 1]$, `and a list of rationales each containing an evidence substring that must exactly match the input. We parse the output, tolerate minor formatting errors, and verify that every evidence span is a substring of the prompt; non-conforming instances are down-weighte or discarded.`

Matched pairs. To reduce confounding from exposure correlates, we sample pairs (i, j) matched on popularity and age bins (and optionally price or category). Let $\hat{\pi}(i \mid u)$ denote the estimated exposure propensity (Section **??**). For a judged pair (u, i, j), we define the IPS weight

$$w(u, i, j) \;=\; \frac{1}{\max(\hat{\pi}(i \mid u),\, 10^{-6})} + \frac{1}{\max(\hat{\pi}(j \mid u),\, 10^{-6})}. \tag{1}$$

IPS-PairAUC. Let $\widehat{\Pr}[i \succ j \mid u]$ be the model's binary prediction based on scores $s(i), s(j)$ (or the LLM confidence with a 0.5 threshold), and let the LLM judge return a label $y \in \{+1, -1, 0\}$ indicating agreement (+1 if i preferred, -1 if j, 0 if tie/invalid). We compute

$$\mathrm{PairAUC}_{\mathrm{IPS}} \;=\; \frac{\sum_{(u,i,j)} w(u, i, j)\, \mathbf{1}[(\widehat{\Pr}[i \succ j] \geq 1/2 \wedge y = +1) \;\vee\; (\widehat{\Pr}[i \succ j] < 1/2 \wedge y = -1)]}{\sum_{(u,i,j)} w(u, i, j)}. \tag{2}$$

RJS (exposure-aware Kendall-like τ_w). For each user, compute agreement $a_u = \frac{\sum w\,\mathrm{match}}{\sum w}$, then map to $\tau_u = 2a_u - 1 \in [-1, 1]$. Finally aggregate by propensity weights: $\tau_w = \frac{\sum_u (\sum w)_u\, \tau_u}{\sum_u (\sum w)_u}$. This yields an interpretable, rationale-grounded, exposure-aware ranking score.

2.2 ExpoSynth

Collaborative constraints. From logs we estimate: (i) a genre transition matrix $T[g \to g']$ from consecutive interactions; and (ii) an item–item similarity graph via cosine similarity of the item{user incidence matrix. We accept a step $i \to j$ if either $T[g(i) \to g(j)]$ exceeds a threshold or similarity $S(i, j)$ exceeds δ.

Exposure control. We define a target exposure curve $q(\mathrm{pop\_bin})$ that upweights tail (e.g., $\mathrm{pop\_bin} \geq 3$). Generation samples candidates from a mixture of anchor genres and q, with rejection sampling under constraints.

Anchors. For each user we form top-k anchors by genre frequency in recent history, using weights proportional to within-user prevalence. We iterate T steps per user, logging $(u, i, t, \mathrm{source} = \mathrm{synth}, \mathrm{reason} = \mathrm{anchor\text{-}based})$.

2.3 Calib-Link

Given tuples $\{(M_k, \mathrm{RJS}_k, \widehat{\Delta\mathrm{Recall}}_k^{\mathrm{IPS}})\}_k$, we fit an isotonic regressor g such that $g(\mathrm{RJS}) \approx \widehat{\Delta\mathrm{Recall}}^{\mathrm{IPS}}$. We report Spearman ρ, Kendall τ, and R^2. Isotonicity preserves ordering and improves extrapolation stability.

3 Prototype and Reproducibility

We use MovieLens-1M. Propensity $\hat{\pi}$ is estimated with a logistic model over user frequency, item popularity, item age bin, hour, weekday, and user-genre entropy. Pair sampling matches popularity/age bins. Judge is executed with Tongyi (DashScope) via OpenAI-compatib or native APIs; rationales are validated by substring checks. Scripts, configs, and outputs are under version control; see configs/movielens1m.yaml and reports/.

4 Results

4.1 Judge Metrics (RJS and IPS-PairAUC)

Table ?? summarizes overall and per-popularity-bucket results.

Bucket	PairAUC_IPS	RJS
ALL	0.445	-0.110
pop_bin=0	0.378	-0.243
pop_bin=1	0.500	0.000
pop_bin=2	0.488	-0.024
pop_bin=3	0.434	-0.132
pop_bin=4	0.419	-0.161

Table 1: RJS and IPS-PairAUC on MovieLens-1M prototype.

4.2 Calibration (RJS to IPS-Recall@20)

Figure ?? shows isotonic calibration g mapping RJS to IPS-Recall@20 using simulated models. In this small pilot, recall values saturate at 1.0, yielding a flat curve; larger and more diverse models increase variance and informative calibration.

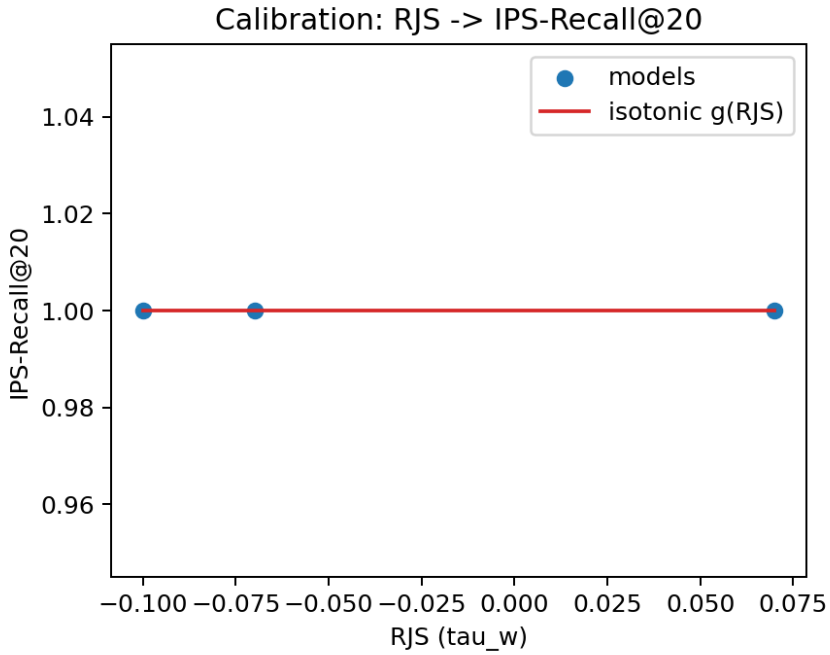

Figure 1: Isotonic calibration g: RJS $\rightarrow$ IPS-Recall@20.

4.3 ExpoSynth Tail Augmentation

Our synthesizer produced 47,117 accepted interactions (accept ratio 0.98) with tail ratio 0.607 (pop_bin$\geq$3). Constraints are enforced via the learned genre transition matrix and item similarities. Increasing steps-per-user proportionally scales synthetic volume while preserving tail emphasis.

5 Related Work

LLM-based evaluation for generation and recommendation has explored rubric-guided or pairwise judging, yet typically without explicit exposure correction or evidence auditing. Causal unbiasing in recommendation relies on IPS/DR and counterfactual estimators. Synthetic data for recommenders ranges from heuristic simulators to generative models, but few works control exposure explicitly while enforcing collaborative constraints.

6 Theoretical Notes

Under standard assumptions (consistency, positivity, and correct propensity), IPS yields an unbiased estimate of counterfactual averages. Our pairwise IPS objective inherits this property when matches isolate the treatment of interest (content differences) and propensity pertains to exposure, not preference. Monotone calibration via isotonic regression is consistent for the regression function under weak conditions and preserves order, which is critical for model ranking.

7 Ablations and Robustness Protocol

We recommend: (i) *With/without exposure weights* to quantify bias reduction; (ii) *With/without evidence checking* to evaluate hallucination control; (iii) *Single vs. multi-j* to measure variance; (iv) *Constraint on/off* in ExpoSynth to assess sequence plausibility and downstream gains; (v) Sensitivity to pair-matching criteria and LLM temperature.

8 Ethics and Limitations

LURE requires careful handling of privacy and potential demographic bias. Evidence alignment reduces, but does not eliminate, hallucination risks. Propensity mis-specificati and positivity violations may bias IPS. Synthetic data should not be treated as real user behavior and must be audited before deployment.

9 Discussion

We unify causal unbiasing, verifiable rationales, and exposure-controllable synthesis. Future work includes multi-judge aggregation, stronger DR estimators, and cross-domain calibration.

10 Conclusion

LURE delivers an exposure-aware LLM judge with rationales, a controllable synthesizer for tail augmentation, and a calibration link to external metrics. The framework is reproducible and extensible.

References

[1] A. Swaminathan and T. Joachims. Counterfactual Risk Minimization. KDD, 2015.

[2] M. Dudík et al. Doubly Robust Policy Evaluation and Learning. ICML, 2011.

[3] B. Zadrozny and C. Elkan. Transforming Classifier Scores into Accurate Multiclass Probability Estimates. KDD, 2002.

[4] L. Li et al. A Contextual-Bandit Approach to Personalized News Article Recommendation. WWW, 2010.

[5] OpenAI. Models Evaluating Models. 2023.