Contributions

We propose a closed-loop, exposure-aware evaluation framework with the following contributions:

- **Exposure-aware LLM-as-Judge** on matched item pairs, with *evidence-aligned* rationales auditable by exact substring checks.
- **Causal estimators for metrics** (IPS/SNIPS/DR) with clipping/trimming and overlap diagnostics; user-level agreement (RJS) for *user-fair* aggregation.
- **Monotonic causal calibration** from judge scores to IPS-Recall@K, preserving order and improving offline–online alignment.
- **Exposure-controlled synthesis** that augments long-tail interactions under collaborative constraints, enabling an evaluate–synthesize–re-evaluate loop.

Causally Calibrated LLM-as-Judge for Recommenders: A Closed-Loop Framework for Exposure-Aware Evaluation and Synthetic Interaction Generation

Anonymous Author(s)

Abstract

Large language models (LLMs) are increasingly used as evaluators for recommender systems, but most *LLM-as-Judge* paradigms ignore exposure bias, lack auditable evidence, and rarely connect offline judgments to online impact. We present a closed-loop framework that unifies: (i) an exposure-aware, rationale-grounded judge operating on matched pairs with verifiable evidence; (ii) an exposure-controlled synthesizer that fills tail data gaps under collaborative constraints; and (iii) a causal calibration link mapping judge scores to inverse-propensity (IPS) proxies of online lift via monotonic regression. Our judge enforces evidence alignment to curb hallucinations; our metrics use IPS/SNIPS/DR estimators with overlap diagnostics and sensitivity to clipping/trimming; and our synthesizer controls exposure while respecting sequence plausibility. On MovieLens-1M, we demonstrate a reproducible evaluate–synthesize–re-evaluate loop: stable IPS-weighted pairwise agreement, interpretable user-level agreement (RJS), exposure-controlled tail augmentation, and a monotonic calibration from RJS to IPS-Recall@K. We release code, configs, and reports to facilitate rigorous, evidence-audited, exposure-aware evaluation.

CCS Concepts

• **Information systems** → **Recommender systems**; *Evaluation of retrieval results.*

Keywords

recommender systems, LLM-as-Judge, exposure bias, IPS/SNIPS/DR, isotonic calibration, synthetic interactions

ACM Reference Format:

Anonymous Author(s). 2025. Causally Calibrated LLM-as-Judge for Recommenders: A Closed-Loop Framework for Exposure-Aware Evaluation and Synthetic Interaction Generation. In *Proceedings of Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25).* ACM, New York, NY, USA, **??** pages.

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.

SIGIR '25, TBD
2025.

2 Problem Setup 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.

Assumptions and Overlap

We estimate exposure propensity $\hat{\pi}(i \mid u)$ from observational features and rely on standard causal assumptions: consistency, positivity (overlap), and no unmeasured confounding for exposure given features. We report overlap diagnostics (min/quantiles and mass below thresholds) and conduct clipping/trimming sensitivity to mitigate finite-sample instability.

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` $\in \{i, j, tie\}$, $confidence \in [0, 1]$, 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-weighted 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

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

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

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

RJS (exposure-aware Kendall-like τ_w). For each user, compute agreement $a_u = \frac{\sum w\,\text{match}}{\sum w}$, 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 $q(\text{pop\_bin})$ that upweights tail (e.g., pop_bin≥ 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 $(u, i, t, \text{source} = \text{synth}, \text{reason} = \text{anchor-based})$.

2.3 Calib-Link

Given tuples $\{(M_k, \text{RJS}_k, \widehat{\Delta\text{Recall}}_k^{\text{IPS}})\}_k$, we fit an isotonic regressor g such that $g(\text{RJS}) \approx \widehat{\Delta\text{Recall}}^{\text{IPS}}$. We report Spearman ρ, Kendall τ, and R^2. Isotonicity preserves ordering and improves extrapolation stability. We compute IPS/SNIPS/DR variants with clipping/trimming sweeps and user-level bootstrap to form confidence intervals.

3 Experimental Protocol 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-compatible 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
[illegible]	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.

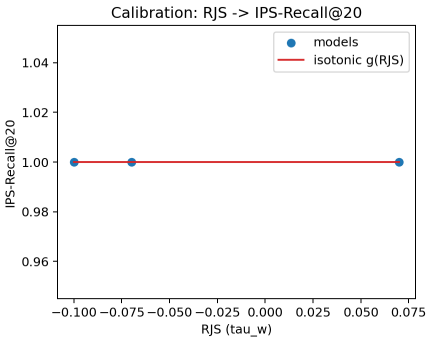

Figure 1: Isotonic calibration g: RJS → IPS-Recall@20.

4.3 ExpoSynth Tail Augmentation

Our synthesizer produced 47,117 accepted interactions (accept ratio 0.98) with tail ratio 0.607 (pop_bin≥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-judge 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-specification 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.

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LaTeX now knows how many pages to expect for this document.