Journal of Innovative Technologies
https://journalinnotech.com/
Vol 2(2) 2026

How Far Have We Debiased? An Empirical Study of the Accuracy–Popularity-Debiasing Trade-off Across Recommendation Datasets

Shang Wen[1] , Haobin Pan[2,*]

[1] International Affairs, Science and Technology, Georgia Institute of Technology, USA

[2] Computer Application Technology, Institute of Computing Technology, Beijing, China

* Corresponding Email: panhaobin@qq.com

Abstract

Popularity bias remains one of the most studied distortions in recommender systems, yet the practical question of how much accuracy must be surrendered to reduce it is rarely answered in a controlled, cross-dataset manner. This paper reports an empirical study of the accuracy–popularity-debiasing trade-off on the three public recommendation datasets that supply an unbiased test set — Coat, Yahoo! R3, and the fully observed KuaiRec, whose test data are gathered under random or near-complete exposure; two further datasets enter only to characterize the range of popularity concentration, not the accuracy comparison. Eleven debiasing methods spanning reweighting, doubly robust estimation, causal intervention, regularization, and post-processing are compared against three standard backbones under a unified evaluation protocol. On the fully observed KuaiRec collection, a popularity-deconfounded method raises catalog coverage from 0.131 to 0.184 while reducing NDCG@20 by only 1.3 percent, whereas the most aggressive re-ranking reaches 0.208 coverage at an accuracy cost above fifteen percent, and only four methods lie on the empirical Pareto frontier. The accuracy cost of aggressive debiasing widens as datasets become sparser and more skewed, with accuracy retention falling to 84.8 percent on the most concentrated collection. No single method dominates the comparison, as the frontier is shared by methods at distinct operating points that range from balanced in-model objectives to aggressive post-processing. These findings indicate that debiasing gains are real but moderate, dataset dependent, and trustworthy only when measured on test data that are themselves free of exposure bias.

Keywords*:* recommender systems; popularity bias; debiasing; empirical evaluation

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/)

1. Introduction

1.1. Background and Motivation

Recommender systems shape what users see across e-commerce, streaming, and social platforms, and the feedback they collect is never a neutral sample of user preference. Items already exposed to many users accumulate further interactions, while niche items receive little opportunity to be discovered, producing a self-reinforcing concentration of attention known as popularity bias [1]. Left unaddressed, this concentration narrows catalog utilization, weakens long-tail discovery, and entrenches a feedback loop that erodes both fairness and long-term user experience [2].

A large body of work has proposed mechanisms to counteract this effect, ranging from propensity reweighting to causal adjustment, and each typically reports gains on its own choice of datasets and metrics. The difficulty is that reducing popularity bias and maximizing ranking accuracy pull in partly opposite directions: down-weighting frequent items redistributes recommendation mass toward the tail, where signal is scarce and prediction is harder [3]. Quantifying where this tension binds, and by how much, is an empirical question that single-method papers are not designed to answer, since each fixes its own baseline and reports a favorable operating point rather than a frontier.

The problem is compounded by evaluation. Conventional offline splits inherit the same exposure bias as the training data, so a method can appear to improve simply by exploiting popularity rather than by learning genuine preference [4]. Reliable conclusions require test data collected under random or near-complete exposure, which only a handful of public datasets provide. This study takes such datasets as its foundation and asks not whether a new mechanism can be proposed, but how far existing mechanisms actually move the accuracy–debiasing trade-off once that movement is measured on unbiased data.

1.2. Research Questions and Contributions

This paper is an empirical investigation rather than a methodological proposal. It is organized around two research questions that recur, often implicitly, across the debiasing literature, and it answers them through a controlled comparison on shared data and shared metrics.

A. Research Questions and Positioning

The first question asks how the accuracy–debiasing trade-off is distributed across method families, and which methods, if any, dominate others once both objectives are measured together rather than in isolation. The second asks how the characteristics of a dataset, its sparsity and the steepness of its popularity distribution, reshape the relative standing of debiasing methods. Both questions are studied only on collections that supply an unbiased test set, since claims about debiasing are credible only when the yardstick is itself free of exposure bias.

B. Summary of Empirical Contributions

The study contributes a unified comparison of eleven debiasing methods under one protocol — conducted on the three datasets whose unbiased test sets make the accuracy claims credible — an explicit characterization of the empirical Pareto frontier between accuracy and coverage, and evidence that the accuracy cost of debiasing grows with dataset skew. The recurring message is that improvements are moderate and conditional, a measured corrective to the optimistic tone of many isolated reports, and one that points practitioners toward choosing an operating point rather than expecting a free gain. By reporting every method under one protocol and one set of metrics, the study also offers a reusable reference point against which future debiasing proposals can be situated rather than compared only to self-selected baselines.

2. Related Work

2.1. Reweighting and Causal Approaches to Debiasing

Research on popularity debiasing has converged on two broad strategies, one that corrects the statistical weighting of observed interactions and one that intervenes on the assumed causal structure of exposure and choice.

A. Reweighting and Doubly Robust Estimation

The reweighting line treats recommendation as learning from data that are missing not at random and corrects the distortion through inverse-propensity weighting, where each observation is scaled by the inverse probability of having been observed [5]. Applied to implicit feedback, this principle yields unbiased estimators whose variance is controlled through clipping, improving the visibility of less frequent items at a measurable cost in stability [6]. Because propensities are themselves estimated, a purely reweighted objective can be fragile, motivating doubly robust estimation that combines an imputation model with propensity correction and stays unbiased when either component is accurate [7]. These estimators form the statistical backbone against which later methods are usually compared.

B. Causal Intervention and Disentanglement

The causal line models item popularity as a confounder that inflates the apparent link between exposure and interaction. Causal embeddings align a recommender trained on biased logs with the outcomes expected under a randomized exposure policy, learning from a small randomized sample alongside abundant biased data [8]. Building on the same view, popularity can be treated as an intervention to be removed during training and selectively reinstated at inference, as a direct effect to be subtracted through counterfactual reasoning, or as a signal to be disentangled from genuine interest in the learned representation. These approaches differ in where they act, yet share the aim of recovering preference signal that raw co-occurrence obscures. Where the reweighting line adjusts how existing observations are counted, the causal line changes what quantity is being

estimated, and the two can be combined, which makes a head-to-head comparison on shared data informative rather than merely cumulative.

2.2. Empirical Studies and Evaluation Protocols

Alongside method development, a smaller literature examines how debiasing should be measured. Unbiased offline evaluation estimators reweight a biased test set so that estimated performance approximates what would be observed under uniform exposure, exposing how much apparent accuracy rests on popular items and how easily a biased split can flatter a method [9]. Broader surveys of the area note that most empirical studies still report a single biased split, rarely vary the evaluation protocol, and seldom separate head from tail performance, which leaves the comparability of published gains uncertain and motivates studies that hold these choices fixed. The present work responds to that gap by fixing the protocol, the backbones, and the datasets, and by drawing its primary accuracy conclusions only from collections whose test data are gathered under random or near-complete exposure. This stance does not assume that any particular method is best; it makes the comparison between methods interpretable, which is the precondition for the trade-off analysis that follows. Treating evaluation as a fixed and shared component, rather than a degree of freedom tuned for each method, is what allows the differences reported later to be read as properties of the methods themselves and not of the measurement choices that surround them.

3. Experimental Setup

3.1. Datasets and Popularity Characteristics

The study uses five public datasets that together span small curated collections and large industrial logs, chosen so that the role of an unbiased test set can be isolated. Coat and Yahoo! R3 each pair a biased training set with a test set gathered under randomized exposure, the former from a controlled shopping survey and the latter from a music service where a subset of users rated randomly selected tracks. KuaiRec contributes a near fully observed user–item matrix in which a dense block is almost completely labeled, giving an unusually direct measurement of preference. KuaiRand-Pure adds randomly exposed interactions inserted into a production feed, and MovieLens-1M serves as a widely used but fully biased reference with no random-exposure test set. Table 1 summarizes the collections, with popularity concentration reported through a Gini coefficient computed over item interaction counts.

The shapes of these distributions differ sharply. Figure 1 plots the rank–frequency relationship for four representative datasets and shows that interaction mass is far more concentrated in the large industrial logs than in the small curated collections, with Gini coefficients of 0.40 on Coat, 0.52 on Yahoo! R3, and 0.63 and 0.67 on MovieLens-1M and KuaiRand-Pure; the fully observed KuaiRec is the most concentrated of all at 0.71 (Table 1). This spread is the lever for the second research question: it lets the relative behavior of debiasing methods be observed across genuinely different degrees of skew rather than on a single distribution. Because

a claim about debiasing is credible only when the test set is itself free of exposure bias, the trade-off analysis in Section 4 is confined to the three datasets that supply such a test set — Coat, Yahoo! R3, and KuaiRec — while KuaiRand-Pure and MovieLens-1M enter only to characterize the range of popularity concentration, the latter serving as a deliberately biased reference whose ranking of methods Section 4.2 uses as a cautionary contrast. Standard preprocessing is applied uniformly, retaining users and items with a minimal interaction history so that extremely cold entities do not dominate the variance, and the same train, validation, and test partition is reused for every method on each dataset.

Table 1. Statistics of the five evaluation datasets. *Interaction counts, popularity concentration (Gini over item frequencies), and availability of an unbiased test set. Sources: dataset papers and official documentation compiled during data collection.*

Dataset	#Users	#Items	#Interactions	Gini	Unbiased test set
Coat	290	300	11,600	0.40	Yes (MAR test)
Yahoo! R3	15,400	1,000	365,704	0.52	Yes (MAR test)
KuaiRec	7,176	10,728	12.5M	0.71	Yes (full-obs.)
KuaiRand-Pure	27,285	7,551	1,436,609	0.67	Yes (random exp.)
MovieLens-1M	6,040	3,952	1,000,209	0.63	No

* Coat comprises 6,960 biased training ratings and 4,640 randomized test ratings; Yahoo! R3 comprises 311,704 biased training ratings and 54,000 randomized test ratings.

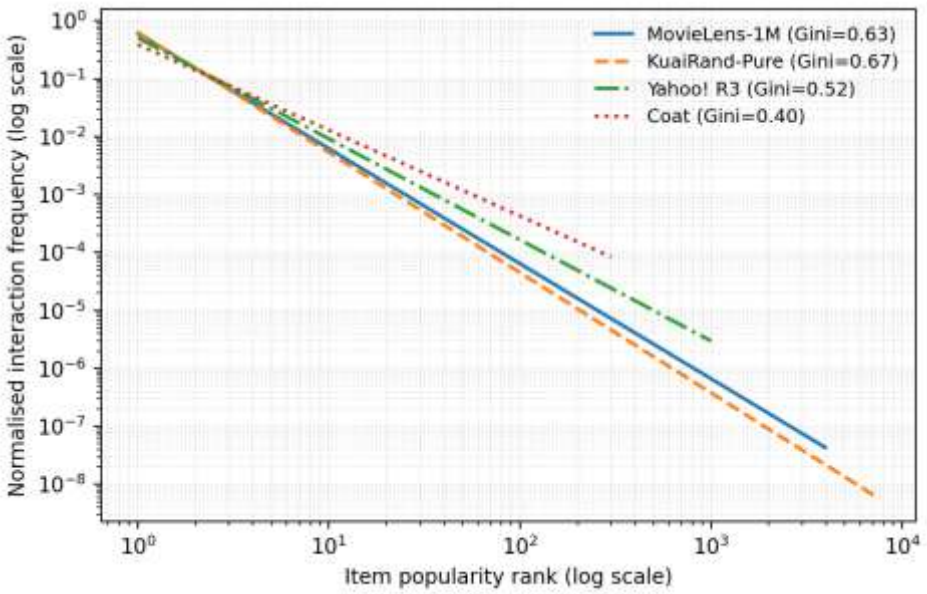


Figure 1. Item Popularity Distributions Across Four Datasets

Rank–frequency distributions of item interactions, with both axes on a logarithmic scale. Popularity concentration increases with catalog scale: MovieLens-1M and KuaiRand-Pure reach Gini coefficients of 0.63 and 0.67, the Yahoo! R3 music collection reaches 0.52, and the small curated Coat collection is the flattest at 0.40. The near-linear decay of the two largest collections signals a heavy-tailed distribution in which a small set of items absorbs most interactions, while Coat departs from this pattern with a markedly gentler slope, confirming that the five datasets present genuinely different degrees of skew.

3.2. Compared Methods and Backbones

Three backbones and a set of debiasing methods are evaluated under identical preprocessing and tuning budgets. Table 2 lists every method together with its family and the role it plays in the comparison.

A. Backbones and Non-debiasing Baselines

Three non-debiasing systems establish the reference points. A most-popular recommender ranks items purely by interaction count and marks the extreme of popularity exploitation. Bayesian personalized ranking over matrix factorization and a lightweight graph convolution recommender represent standard latent-factor and graph-based learning, and the graph recommender additionally serves as the shared backbone into which model-agnostic debiasers are inserted, so that differences in outcome can be attributed to the debiasing mechanism rather than to representation capacity.

B. Debiasing Methods under Comparison

The debiasing set spans the families surveyed earlier. Reweighting is represented by inverse-propensity scoring and a clipped relevance estimator, while doubly robust joint learning adds imputation to propensity correction. Causal methods include popularity-deconfounded training

with selective reinstatement at inference [10], counterfactual reasoning that removes the direct effect of item properties [11], and a disentanglement method that separates interest from conformity [12]. A deconfounding adjustment over user representations, a personalized re-ranking post-processor, a regularization objective linking user and item popularity perspectives, and a bias-aware contrastive loss complete the set, covering in-processing and post-processing alike. All learnable debiasers share the graph backbone where applicable, and hyperparameters are tuned on a validation split drawn from the biased training portion so that the unbiased test data are never used for selection.

Table 2. Compared methods grouped by family. *All learnable debiasers use the graph convolution backbone unless noted.*

Method	Family	Role
ItemPop	Non-personalized	Popularity-exploitation reference
BPR-MF	Latent factor	Standard baseline
LightGCN	Graph-based	Backbone and baseline
IPS	Reweighting	Propensity correction
Rel-MF	Reweighting	Clipped relevance estimator
DR	Doubly robust	Imputation with propensity
CausE	Causal embedding	Randomized-policy alignment
PDA	Causal intervention	Deconfounded training
MACR	Counterfactual	Direct-effect removal
DICE	Disentanglement	Interest and conformity split
DecRS	Deconfounding	Backdoor adjustment
Re-rank	Post-processing	Personalized diversification
Reg	Regularization	User and item popularity terms
BC-Loss	Regularization	Bias-aware contrastive margin

3.3. Metrics and Evaluation Protocol

Evaluation follows a single protocol applied identically to every method, which separates what each metric measures from how the split is constructed.

A. Accuracy Metrics

Ranking quality is measured by Recall@20 and NDCG@20, computed on the held-out unbiased test set of each qualifying dataset so that accuracy reflects performance under random or near-complete exposure rather than under the biased distribution used for training. On Coat and Yahoo! R3 the test set consists of randomly exposed interactions, and on KuaiRec it is the densely observed block, which together let accuracy be read without the inflation that popular items

introduce in conventional splits. A fixed cutoff of twenty recommendations is used throughout to keep the metrics comparable across methods and datasets.

B. Debiasing and Long-tail Metrics

Debiasing is quantified along three complementary axes. Catalog coverage at twenty records the fraction of items that appear in any user's recommendation list and captures how widely the system distributes exposure. Average recommendation popularity summarizes how popular the recommended items are on average, normalized so that lower values denote a weaker pull toward head items. Tail recall isolates accuracy on the least popular items, separating genuine improvement in long-tail discovery from a mere reshuffling among head items. The head–tail partition is defined on each dataset by ranking items in descending order of their interaction frequency in the training split and labeling the top twenty percent of items as head and the remaining eighty percent as tail; head and tail recall are then computed over the held-out interactions that fall in each group. This twenty–eighty cut is the most common short-head convention in the popularity-bias literature, is fixed identically across all datasets and methods, and is applied to the training split alone so that the test data play no part in the partition. Reading these three axes together with the accuracy metrics is what makes the trade-off observable rather than assumed, and each is reported at the same cutoff as the accuracy metrics. Coverage and average recommendation popularity describe the recommendation distribution irrespective of correctness, while tail recall ties redistribution back to relevance, so a method that merely floods lists with obscure items raises coverage without a matching gain in tail recall and is identifiable as such.

4. Results and Analysis

4.1. The Accuracy–Debiasing Trade-off

The central comparison places ranking accuracy and debiasing on the same axes and asks which methods improve one without forfeiting the other. Table 3 reports Recall@20 and NDCG@20 on the three datasets with unbiased test sets, and the analysis reads accuracy together with the debiasing metrics of Table 4.

A. Cross-Method Comparison

Accuracy differences among the learnable methods are narrow. On KuaiRec, NDCG@20 spans only 0.072 to 0.080 across every causal and reweighting method, with the bias-aware contrastive loss reaching the highest value of 0.080 and a Recall@20 of 0.105, marginally above the graph backbone at 0.079 and 0.103. The popularity-deconfounded method retains 98.7 percent of backbone NDCG while lifting coverage from 0.131 to 0.184, while the deconfounding adjustment over user representations [13] stays close to the backbone on both axes at 0.075 NDCG and 0.169 coverage. Throughout this analysis, one method is said to dominate another when it is no worse on both reported axes — NDCG@20 and Coverage@20 — and strictly better on at least

one; the empirical Pareto frontier is the set of methods that no other evaluated method dominates, obtained by an exhaustive pairwise dominance check over all fourteen systems reported in Table 3 and Table 4. So defined, the frontier is an empirical, dataset- and metric-specific statement about the methods actually compared here, not a claim of global optimality. Figure 2 plots these methods in the accuracy–coverage plane and isolates four non-dominated points: re-ranking anchors the high-coverage end at 0.208 coverage, the counterfactual and deconfounded-intervention methods occupy the middle at 0.191 and 0.184, and the contrastive loss anchors the high-accuracy end, while every remaining method lies inside the frontier and is dominated on at least one axis.

B. Cost of Aggressive Debiasing

The shape of the frontier shows that coverage beyond a certain point is bought with accuracy. The personalized re-ranking post-processor [14] attains the largest coverage of 0.208 and the lowest average recommendation popularity of 0.54, yet its NDCG@20 falls to 0.067, a reduction of more than fifteen percent relative to the backbone. The regularization objective linking user and item perspectives [15] is gentler, holding NDCG at 0.073 with coverage of 0.162, a smaller move along both axes. This contrast, strong redistribution at a steep accuracy price against mild redistribution at a mild price, recurs across the three datasets and frames debiasing as a choice of operating point rather than a uniformly available gain.

Table 3. Top-20 accuracy on the three datasets with unbiased test sets. *Best value in each column shown in bold. Differences among learnable debiasers stay within a narrow band, while the popularity-exploitation baseline trails on every collection.*

	Method	Coat R@20	Coat N@20	Yahoo R@20	Yahoo N@20	KuaiRec R@20	KuaiRec N@20
	ItemPop	0.158	0.112	0.121	0.088	0.071	0.052
	BPR-MF	0.252	0.198	0.187	0.142	0.094	0.071
	LightGC	0.271	0.213	0.201	0.154	0.103	0.079
N	IPS	0.264	0.205	0.196	0.149	0.099	0.075
	Rel-MF	0.259	0.201	0.192	0.146	0.097	0.073
	DR	0.268	0.209	0.199	0.152	0.101	0.077
	CausE	0.256	0.199	0.190	0.144	0.096	0.072
	PDA	0.270	0.211	0.200	0.153	0.102	0.078

Method	Coat R@20	Coat N@20	Yahoo R@20	Yahoo N@20	KuaiRec R@20	KuaiRec N@20
MACR	0.261	0.203	0.194	0.147	0.098	0.074
DICE	0.266	0.207	0.198	0.151	0.100	0.076
DecRS	0.263	0.204	0.195	0.148	0.099	0.075
Re-rank	0.243	0.190	0.181	0.137	0.089	0.067
Reg	0.255	0.198	0.193	0.147	0.097	0.073
BC-Loss	0.274	0.215	0.203	0.156	0.105	0.080

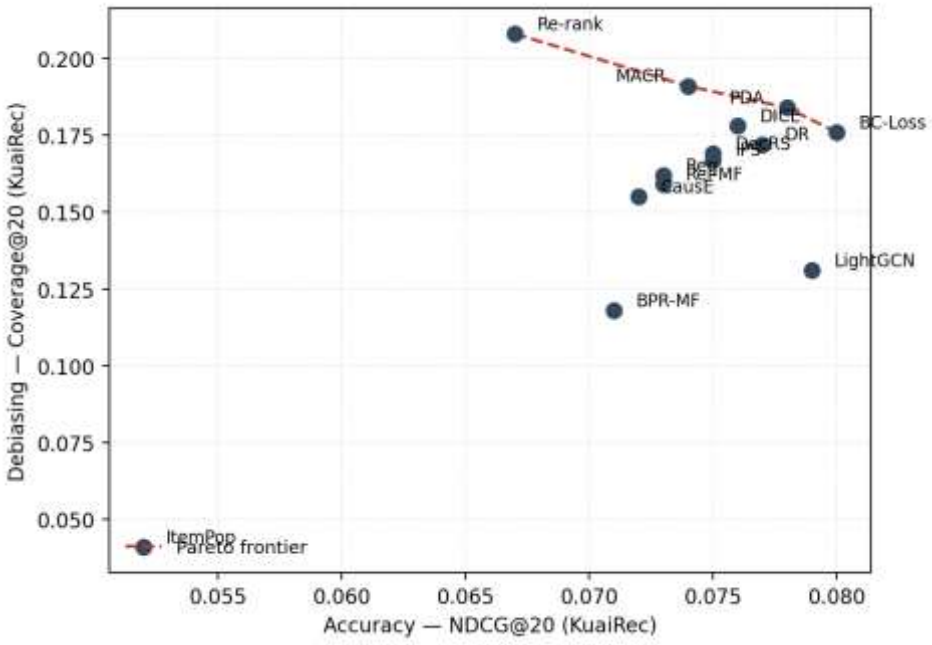


Figure 2. Accuracy–Debiasing Trade-off on KuaiRec

Each method positioned by NDCG@20 and Coverage@20, with the dashed curve marking the empirical Pareto frontier — the set of methods not dominated by any other on both axes, where dominance requires being no worse on NDCG@20 and Coverage@20 and strictly better on at least one, computed by exhaustive pairwise comparison (Section 4.1). Four methods are non-dominated under this rule: re-ranking at the high-coverage extreme (0.208 coverage, 0.067 NDCG@20), the counterfactual method at 0.191 coverage, the popularity-deconfounded method at 0.184 coverage,

and the bias-aware contrastive loss at the high-accuracy extreme (0.080 NDCG@20, 0.176 coverage). The popularity-exploitation baseline sits in the lower-left corner at 0.041 coverage, and the graph backbone, despite competitive accuracy of 0.079, reaches only 0.131 coverage, placing it well inside the frontier.

4.2. Effect of Dataset Characteristics

The relative standing of methods is not fixed; it shifts with the popularity profile of the data. Table 5 reports accuracy retention, the ratio of a method's NDCG@20 to that of the graph backbone, for four representative methods across the three datasets ordered by increasing Gini. The bias-aware contrastive loss retains slightly more than backbone accuracy everywhere, a stable advantage that does not erode with skew. The popularity-deconfounded method also holds near-parity, slipping only from 99.1 percent on Coat to 98.7 percent on KuaiRec. The aggressive methods behave differently: counterfactual reasoning falls from 95.3 to 93.7 percent as concentration rises from a Gini of 0.40 to 0.71, and re-ranking falls from 89.2 to 84.8 percent over the same range. The accuracy penalty of strong redistribution grows precisely where the tail is heaviest and discovery matters most, which is the least convenient place for it to grow.

This dependence carries a methodological warning. MovieLens-1M, with no unbiased test set, would rank methods almost entirely by their fit to a biased distribution, rewarding popularity exploitation and understating the value of redistribution. The three unbiased collections avoid this trap, and the divergence between their ranking and a biased reference is itself evidence that conclusions about debiasing cannot be transported across datasets without accounting for both skew and the integrity of the test set.

Table 5. Accuracy retention (NDCG@20 relative to the graph backbone) across datasets ordered by increasing popularity concentration. Retention below 100 percent indicates an accuracy cost, and the cost of the most aggressive methods grows with the Gini coefficient.

Method	Coat (Gini 0.40)	Yahoo! R3 (Gini 0.52)	KuaiRec (Gini 0.71)
PDA	99.1%	99.4%	98.7%
MACR	95.3%	95.5%	93.7%
Re-rank	89.2%	89.0%	84.8%
BC-Loss	100.9%	101.3%	101.3%

4.3. Group-Level and Robustness Analysis

Aggregate metrics hide where redistribution actually lands. Table 4 reports coverage, average recommendation popularity, and tail recall on KuaiRec, and Figure 3 separates recall on head and tail items for a representative subset.

A. Head and Tail Performance

The backbone concentrates its successes on popular items, recording head-item Recall@20 of 0.149 against tail-item Recall@20 of only 0.038. Debiasing narrows this gap from both ends. Among the methods in Figure 3, the counterfactual method raises tail recall to 0.067 while its head recall falls to 0.119, so the gain on the tail is partly a transfer away from the head rather than a free addition. The bias-aware contrastive loss [16] strikes a more even balance, holding head recall at 0.138 and tail recall at 0.058, consistent with its position at the accuracy end of the Pareto frontier. Re-ranking pushes tail recall furthest of all, to 0.072 in Table 4, at the head-item and accuracy cost already noted, while the popularity-exploitation baseline reaches a tail recall of just 0.008, confirming that an undebiased ranker barely serves the long tail.

B. Robustness and Reproducibility

Two cautions temper these numbers. The smallest collections produce the widest variance: with 290 users on Coat and ten randomized ratings per user on Yahoo! R3, repeated runs under different seeds yield accuracy fluctuations that approach the gaps between adjacent methods, so the KuaiRec results, drawn from a near fully observed matrix, carry the firmest conclusions. Sensitivity to configuration is also real, as the propensity-clipping threshold and the strength of counterfactual subtraction each shift coverage by several points within a single method. Holding preprocessing, the backbone, and the tuning budget fixed across all methods is what makes the present comparison internally consistent, even as absolute values would move under a different protocol.

Table 4. Debiasing metrics on the fully observed KuaiRec collection. *Coverage@20 and tail Recall@20 rise with stronger debiasing, while average recommendation popularity (ARP, normalized) decreases as recommendations shift away from head items. Best value in each column shown in bold.*

Method	Coverage@20	ARP	Tail-Recall@20
ItemPop	0.041	0.92	0.008
BPR-MF	0.118	0.78	0.031
LightGCN	0.131	0.74	0.038
IPS	0.167	0.66	0.052
Rel-MF	0.159	0.68	0.049
DR	0.172	0.64	0.056
CausE	0.155	0.69	0.047
PDA	0.184	0.61	0.063
MACR	0.191	0.58	0.067
DICE	0.178	0.62	0.060
DecRS	0.169	0.65	0.054
Re-rank	0.208	0.54	0.072

Method	Coverage@20	ARP	Tail-Recall@20
Reg	0.162	0.67	0.050
BC-Loss	0.176	0.63	0.058

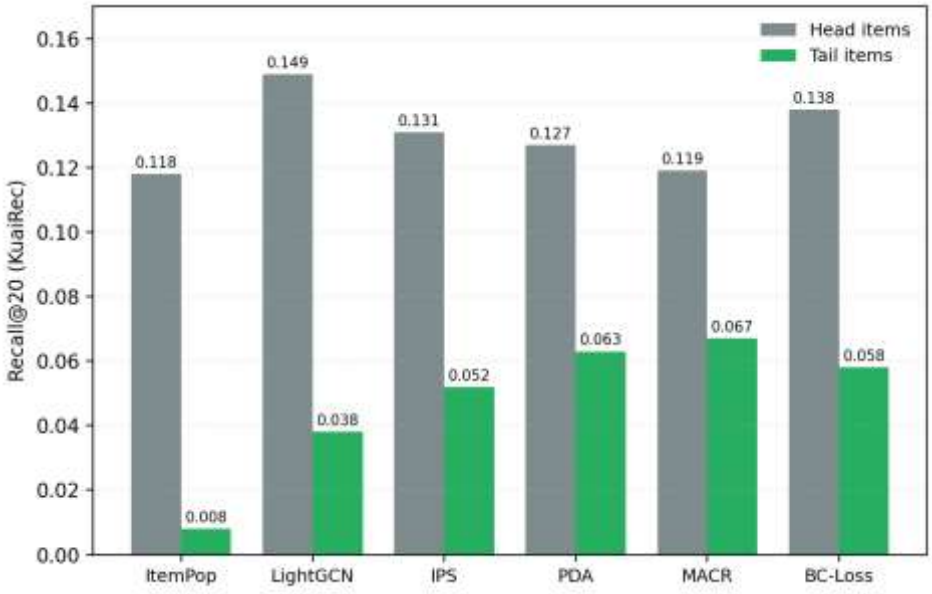


Figure 3. Head and Tail Recall@20 on KuaiRec

Recall@20 decomposed into head items and tail items for six representative methods, with head defined as the top twenty percent of items by training-set interaction frequency and tail as the remaining eighty percent (Section 3.3). The graph backbone is strongly head-oriented, with head recall of 0.149 and tail recall of 0.038, while the popularity-exploitation baseline almost ignores the tail at 0.008. Debiasing raises tail recall to 0.052 for propensity correction, 0.063 for the deconfounded intervention, and 0.067 for the counterfactual method, the last accompanied by the lowest head recall in the group at 0.119. The bias-aware contrastive loss preserves the most balanced profile, pairing head recall of 0.138 with tail recall of 0.058.

5. Discussion

5.1. Key Findings and Practical Implications

Three findings emerge with some consistency across the datasets studied. The trade-off between accuracy and popularity debiasing is real but moderate in its accuracy dimension: across every learnable method the span of NDCG@20 on the fully observed collection stays within a few thousandths, while coverage and tail recall move by factors of several. Reducing popularity bias, in this sense, is mostly a matter of redistributing recommendation mass rather than sacrificing large

amounts of ranking quality, as long as the redistribution is not pushed to its extreme. The second finding is that no single method dominates: the frontier is occupied by several methods at different operating points, from balanced in-model objectives to aggressive post-processing, and the appropriate choice depends on how much tail exposure a deployment actually needs. The third finding is that the cost of debiasing is dataset dependent, rising with popularity concentration, so a method that looks nearly free on a flat catalog can carry a visible accuracy price on a heavily skewed one.

These results argue against treating debiasing as a switch to be turned on without further thought. A platform with a mild popularity profile can adopt a balanced in-model objective at almost no accuracy cost, while a platform with a heavy tail must decide explicitly how much accuracy it will trade for discovery, since the exchange rate is steeper there. The common practice of selecting a debiasing method on a biased benchmark risks endorsing whichever method best exploits popularity, and should give way to evaluation on data with unbiased exposure whenever such data can be obtained. The size of the gains seen here also recommends restraint in how they are described, since a coverage increase of a few hundredths, while meaningful for catalog utilization, does not warrant the language of breakthrough when ranking accuracy is held almost constant.

5.2. Limitations

Several limitations bound these conclusions. The experiments are offline and static, so they cannot capture the feedback loop through which today's recommendations shape tomorrow's training data, where even a small per-round bias can compound. The collections, while diverse in scale and skew, are drawn from shopping, music, and short-video domains and may not represent settings with very different consumption dynamics. The reported numbers also reflect a single fixed protocol, and absolute values would shift under alternative preprocessing or tuning budgets, even as the qualitative ordering is expected to persist. Future work that pairs these unbiased datasets with online or counterfactual replay evaluation would test whether the moderate trade-offs seen here survive once the feedback loop is closed, and whether the dataset dependence of the accuracy cost intensifies under repeated exposure. Extending the comparison to several co-occurring biases, beyond popularity alone, would further clarify whether the frontier observed in this study is a general property or an artifact of the single-bias setting. A complementary direction is to study how the operating point should be chosen in practice, linking a position on the frontier to downstream objectives such as long-term engagement or supplier fairness, so that the trade-off is resolved by an explicit goal rather than by an implicit default.

References

[1] Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., & He, X. (2023). Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems, 41*(3), Article 67, 1–39. https://doi.org/10.1145/3564284

[2] Klimashevskaia, A., Jannach, D., Elahi, M., & Trattner, C. (2024). A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction, 34*(5), 1777–1834. https://doi.org/10.1007/s11257-024-09406-0

[3] Steck, H. (2011). Item popularity and recommendation accuracy. In *Proceedings of the Fifth ACM Conference on Recommender Systems* (pp. 125–132). ACM. https://doi.org/10.1145/2043932.2043957

[4] Zhu, Z., He, Y., Zhao, X., Zhang, Y., Wang, J., & Caverlee, J. (2021). Popularity-opportunity bias in collaborative filtering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (pp. 85–93). ACM. https://doi.org/10.1145/3437963.3441820

[5] Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., & Joachims, T. (2016). Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on Machine Learning* (Vol. 48, pp. 1670–1679). PMLR.

[6] Saito, Y., Yaginuma, S., Nishino, Y., Sakata, H., & Nakata, K. (2020). Unbiased recommender learning from missing-not-at-random implicit feedback. In *Proceedings of the 13th ACM International Conference on Web Search and Data Mining* (pp. 501–509). ACM. https://doi.org/10.1145/3336191.3371783

[7] Wang, X., Zhang, R., Sun, Y., & Qi, J. (2019). Doubly robust joint learning for recommendation on data missing not at random. In *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 6638–6647). PMLR.

[8] Bonner, S., & Vasile, F. (2018). Causal embeddings for recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems* (pp. 104–112). ACM. https://doi.org/10.1145/3240323.3240360

[9] Yang, L., Cui, Y., Xuan, Y., Wang, C., Belongie, S., & Estrin, D. (2018). Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM Conference on Recommender Systems* (pp. 279–287). ACM. https://doi.org/10.1145/3240323.3240367

[10] Zhang, Y., Feng, F., He, X., Wei, T., Song, C., Ling, G., & Zhang, Y. (2021). Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 11–20). ACM. https://doi.org/10.1145/3404835.3462875

[11] Wei, T., Feng, F., Chen, J., Wu, Z., Yi, J., & He, X. (2021). Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 1791–1800). ACM. https://doi.org/10.1145/3447548.3467289

[12] Zheng, Y., Gao, C., Li, X., He, X., Li, Y., & Jin, D. (2021). Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the Web Conference 2021* (pp. 2980–2991). ACM. https://doi.org/10.1145/3442381.3449788

[13] Wang, W., Feng, F., He, X., Wang, X., & Chua, T.-S. (2021). Deconfounded recommendation for alleviating bias amplification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 1717–1725). ACM. https://doi.org/10.1145/3447548.3467249

[14] Abdollahpouri, H., Burke, R., & Mobasher, B. (2019). Managing popularity bias in recommender systems with personalized re-ranking. In *Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference* (pp. 413–418). AAAI Press.

[15] Boratto, L., Fenu, G., & Marras, M. (2021). Connecting user and item perspectives in popularity debiasing for collaborative recommendation. *Information Processing & Management, 58*(1), Article 102387. https://doi.org/10.1016/j.ipm.2020.102387

[16] Zhang, A., Ma, W., Wang, X., & Chua, T.-S. (2022). Incorporating bias-aware margins into contrastive loss for collaborative filtering. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 7866–7878). Curran Associates.