




## RESEARCH ARTICLE

# CALIBRATED POPULARITY RE-RANKING WITH ALTERNATIVE DIVERGENCE MEASURES FOR POPULARITY BIAS MITIGATION

Emre YALCIN <sup>1,\*</sup>

<sup>1</sup> Computer Engineering Department, Faculty of Engineering, Sivas Cumhuriyet University, Sivas, Turkey  
[eyalcin@cumhuriyet.edu.tr](mailto:eyalcin@cumhuriyet.edu.tr) -  [0000-0003-3818-6712](https://orcid.org/0000-0003-3818-6712)

### Abstract

Popularity bias significantly limits the effectiveness of recommender systems by disproportionately favoring popular items and reducing exposure to diverse, less-known content. This bias negatively impacts personalization and marginalizes niche users and item providers. To address this challenge, calibrated recommendation methods have gained attention, notably the Calibrated Popularity (CP) approach, due to its simplicity, effectiveness, and model-agnostic nature. Originally, CP employs Jensen–Shannon divergence (JSD) to align the popularity distribution of recommended items with users’ historical interaction patterns. However, the choice of divergence measure substantially impacts calibration effectiveness and recommendation diversity. In this study, we systematically explore several alternative divergence measures, including Chi-Square, Wasserstein, Kullback–Leibler, Hellinger, Total Variation, Bhattacharyya, Cosine, and Renyi divergences, within the CP framework. Additionally, we propose a novel divergence-independent evaluation metric, namely Overall Similarity Error, enabling consistent assessment of calibration quality across divergence measures. Experimental results on two benchmark datasets using two collaborative filtering algorithms highlighted critical insights. More aggressive divergences, particularly Chi-Square, significantly enhanced calibration quality, reduced popularity bias, and increased recommendation diversity, albeit with a modest reduction in accuracy. In contrast, smoother divergences, such as JSD, maintained higher accuracy but provided limited improvements in reducing popularity bias. Also, the performed group-based analysis categorizing users into mainstream, balanced, and niche segments based on their historical popularity preferences revealed distinct patterns: balanced users typically achieved higher accuracy due to their evenly distributed preferences; mainstream users showed superior calibration results benefiting from robust signals of popular items; niche users obtained more diverse and personalized recommendations, clearly benefiting from aggressive divergence measures. These results underscore the complexity of addressing popularity bias and highlight the importance of adopting adaptive, user-aware calibration strategies to effectively balance accuracy, diversity, and fairness in recommender systems.

### Keywords

Divergence measures,  
Popularity bias,  
Calibrated recommendation,  
Collaborative filtering,  
Personalization

### Time Scale of Article

Received : 17 April 2025  
Accepted : 03 July 2025  
Online date : 25 September 2025

## 1. INTRODUCTION

Recommender systems (RSs) play a crucial role in navigating the overwhelming volume of content available on digital platforms. From e-commerce and social media to online news and video streaming services, these systems help users discover relevant items efficiently by predicting and ranking content based on user preferences [1]. Traditional recommendation algorithms, such as collaborative filtering,

\*Corresponding Author: [eyalcin@cumhuriyet.edu.tr](mailto:eyalcin@cumhuriyet.edu.tr)

matrix factorization, and deep learning-based approaches, typically aim to maximize accuracy metrics like precision, recall, or mean reciprocal rank. However, an overemphasis on relevance often leads to unintended side effects, one of the most prominent being *popularity bias*.

Popularity bias refers to the tendency of recommendation algorithms to disproportionately favor frequently interacted or globally popular items, often at the expense of less-known but potentially more relevant content [2]. This bias limits exposure to long-tail items and undermines the personalization potential of RSs, especially for users whose interests diverge from mainstream patterns [3]. These users often receive homogenized recommendation lists that fail to reflect their actual preferences. On a system level, popularity bias reduces content diversity, marginalizes niche item providers, and reinforces feedback loops that further entrench popular content. In the long run, this can lead to echo chambers, limit informational diversity and diminish the fairness and sustainability of recommender ecosystems [4].

To address this issue, recent research has introduced the concept of *calibrated recommendation*, which aims to align the distribution of item attributes (e.g., popularity or genre) in recommendation lists with the distributions observed in users' historical interactions [5-7]. Calibration improves not only personalization but also fairness and inclusivity, helping RSs adapt to a broader range of user profiles. By reflecting a user's actual preference structure, including tendencies toward niche or diverse items, calibrated lists increase perceived relevance and user trust. This alignment plays a vital role in delivering ethical and user-centered recommendation experiences.

A recent and widely adopted method in this area is the Calibrated Popularity (CP) framework [6], which re-ranks recommendation lists to strike a balance between accuracy and calibration. It does so by optimizing a joint objective that combines a relevance score with a divergence term, which penalizes deviations between the popularity distribution of the recommendation list and that of the user's past preferences. The original CP formulation uses Jensen–Shannon Divergence (JSD) for this purpose. CP's model-agnostic nature and ability to personalize recommendations by aligning with individual popularity profiles have made it a practical choice for use. However, it remains unclear whether alternative divergence functions might yield better calibration, especially across diverse user groups and datasets.

Divergence measures differ in symmetry, sensitivity to rare events, and penalization characteristics, all of which influence re-ranking behavior and outcomes [8]. For example, Chi-Square Distance emphasizes large proportional differences and may be more effective for users with highly skewed popularity profiles (e.g., those who prefer only tail items). In contrast, Wasserstein Distance captures the effort needed to transform one distribution into another and may offer smoother calibration for users with broad popularity distributions. These differences suggest that divergence choice is not merely a technicality; it can have a meaningful impact on recommendation quality, fairness, and user satisfaction.

In this work, we extend the CP framework by incorporating and systematically evaluating several alternative divergence functions, including Chi-Square, Wasserstein, Kullback–Leibler, Hellinger, Total Variation, Bhattacharyya, Cosine, and Renyi distance. Our goal is to explore whether these divergences provide more precise and context-aware calibration, particularly for long-tail users. We conduct experiments across two prominent benchmark datasets and collaborative filtering algorithms, evaluating each method using relevance, average popularity, diversity, and calibration quality.

Our contributions can be summarized as follows:

1. We propose a novel metric, *Overall Similarity Error (OSE)*, to evaluate calibration quality in a divergence-independent manner. Unlike existing approaches that are tied to specific divergence functions, OSE provides a consistent and interpretable measure of alignment between user history and recommendation output.
2. We extend the CP framework by integrating multiple divergence functions into the re-ranking objective, enabling more flexible and adaptive calibration strategies.

3. We perform a comprehensive experimental comparison of these divergence-based CP variants across two representative collaborative filtering algorithms and benchmark datasets, using multiple evaluation criteria.
4. We conduct group-based analysis to assess how divergence types impact users with varying preferences for popular or niche content, offering practical insights for personalized and fair RS design.

The remainder of this paper is organized as follows: The next section reviews relevant literature on calibrated recommendation methods and divergence-based calibration strategies. Section 3 presents the theoretical background of the CP framework, and the following section introduces the alternative divergence measures considered in this study. Section 5 describes our experimental methodology, including datasets, collaborative filtering algorithms, evaluation metrics, and experimental setups. In Section 5, we also provide experimental results, beginning with an overall comparison of divergence methods, followed by a detailed group-based analysis. Section 6 presents the study's limitations and discusses potential directions for future research. Finally, Section 7 summarizes our main findings and provides concluding remarks.

## 2. RELATED WORK

RSs have been widely studied for their ability to predict user preferences and deliver personalized content, using collaborative filtering, matrix factorization, or deep learning techniques [1]. While these models are typically optimized for accuracy, they frequently exhibit systemic biases, most notably, popularity bias, which disproportionately favors globally popular items while marginalizing niche content. This often results in homogenized recommendation lists, reduced novelty, and unfair experiences for users with long-tail preferences [3].

To mitigate these issues, calibrated recommendation methods have emerged as promising strategies to strike a balance between accuracy, personalization, and fairness. Calibration seeks to align the distribution of certain item attributes, such as genre, topic, or popularity level, in the recommended list with that of the user's historical interactions. The foundational work by [5] introduced the concept of minimizing Kullback-Leibler (KL) divergence between these distributions (i.e., genre) in a post-processing re-ranking step. Although this calibration may reduce accuracy to some extent, it improves fairness and enhances the perceived personalization of the recommendations.

Building upon this foundation, Kaya and Bridge compared calibrated recommendations with intent-aware models using user sub-profiles [9]. Their findings showed that calibration improves diversity and fairness, albeit with a modest reduction in precision. In a similar vein, Seymen et al. proposed a constrained optimization model, namely Calib-Opt, to dynamically balance relevance and calibration via a fairness-aware objective [10].

More recently, the CP framework, introduced by Abdollahpouri et al. [6], has extended calibration to explicitly tackle popularity bias. CP re-ranks items based on a joint objective function combining predicted relevance and the JSD between the popularity distribution of the recommended items and the user's interaction history. The model-agnostic design of CP makes it highly compatible with existing recommendation pipelines and has proven particularly effective for users with non-mainstream preferences, reducing User Popularity Deviation and improving fairness across user segments.

Despite the effectiveness of JSD, researchers have begun to explore alternative divergence functions for calibration. Da Silva and Durão [11, 12] conducted large-scale evaluations using over 390 variants of calibrated systems, incorporating metrics such as Chi-Square, Hellinger, and Weighted Total Variation. Their findings suggest that different divergence measures yield varying trade-offs between calibration quality, recommendation diversity, and accuracy, underscoring the importance of divergence selection

in calibrated frameworks. Their framework builds directly on Steck’s approach [5], applying calibration over genre-class distributions rather than item-level popularity. While these contributions broaden the calibration literature, they remain confined to attribute-based (genre) calibration and do not explicitly address the alignment of popularity at the individual user level.

Additionally, Cha [13] provides a comparative analysis of divergence measures, highlighting differences in sensitivity, symmetry, and robustness, which are particularly relevant in recommendation contexts where user preference distributions are often sparse or skewed. In contrast, our work builds directly on the CP framework, which calibrates recommendation lists based on user-specific popularity profiles rather than predefined content attributes. We model the popularity distribution of a user’s past interactions using popularity buckets and explore the use of multiple divergence measures within the CP setting.

Beyond calibration itself, the fairness-aware recommendation literature has expanded into taxonomies of fairness goals, such as consumer-side fairness (C-fairness), provider-side fairness (P-fairness), and joint fairness (CP-fairness). As categorized in works by [14-16], fairness in RSs is increasingly seen as a multi-stakeholder issue, one that demands balancing diverse interests across both users and item providers. Also, some prominent studies analyzed how users with different characteristics in terms of personality traits [7] or rating behaviors [17] are unfairly affected by final recommendations. However, calibrated approaches like CP offer one mechanism to address this, particularly in managing item exposure across popularity segments. Additionally, Table 1 summarizes the most prominent calibration-based studies, comparing the divergence functions used, calibration targets, and model integration styles.

**Table 1.** Summary of the most prominent calibration-based approaches

Study	Divergence	Calibration Target	Model Type
Steck [5]	Kullback–Leibler (KL) Divergence	Genre distribution	Model-agnostic
Seymen et al. [10]	KL Divergence (within constraints)	Genre distribution	Model-integrated
Kaya & Bridge [9]	KL Divergence	Genre (via sub-profiles)	Model-agnostic
da Silva & Durão [11]	Chi-Square, Hellinger, TVD	Genre classes	Model-agnostic
Abdollahpouri et al. [6]	Jensen–Shannon Divergence (JSD)	Popularity (head/mid/tail)	Model-agnostic
<b>This study</b>	JSD, KL, Chi-Square, Wasserstein, TVD, etc.	Popularity (head/mid/tail)	Model-agnostic

Despite the richness of this body of work, most prior studies have focused on a limited set of divergence functions without systematically evaluating the impact of alternative divergences, such as Wasserstein, Chi-Square, or Total Variation Distance, across different user segments or collaborative filtering architectures. Among existing calibration frameworks, the CP method stands out for its simplicity, generalizability, and empirical effectiveness. Unlike genre-based or intent-aware models, CP directly addresses popularity bias by optimizing user-specific alignment of popularity. Its model-agnostic structure enables seamless integration with diverse recommendation models, and its user-centered formulation ensures relevance not just in content type but in popularity expectations. For these reasons, CP forms the core of our methodology. We extend it by systematically evaluating alternative divergence functions and introducing the OSE metric for divergence-independent evaluation of calibration quality.

### 3. BACKGROUND ON THE CALIBRATED POPULARITY METHOD

In recent years, fairness-aware recommendation techniques have gained increasing attention as systems are expected not only to be accurate but also to reflect user-specific preferences and promote diversity.

Among these methods, the CP framework [6] presents a post-processing re-ranking approach that aims to mitigate popularity bias while maintaining personalization and ensuring high-quality recommendations. CP focuses explicitly on aligning the popularity distribution of the recommended list with that of the user's historical interactions.

Let  $\mathbf{R}_u = \{\mathbf{i}_1, \mathbf{i}_2, \dots, \mathbf{i}_n\}$  be the initial list of top- $N$  recommended items for user  $u$ , produced by a base recommendation algorithm (e.g., matrix factorization or a neural model). The goal of CP is to re-rank this list to obtain a new list  $\mathbf{L}_u \subseteq \mathbf{R}_u$ , such that it reflects the user's historical exposure to item popularity levels. The user's historical popularity distribution is denoted by  $\mathbf{P}$ , which is typically constructed based on the popularity levels of the items the user has interacted with. The popularity distribution of the re-ranked recommendation list is denoted by  $\mathbf{Q}(\mathbf{L}_u)$ . The core principle of CP is to reduce the divergence between  $\mathbf{Q}(\mathbf{L}_u)$  and  $\mathbf{P}$  while preserving recommendation quality. The optimization problem in CP is defined as in Eq. 1.

$$\mathbf{L}_u^* = \arg \max_{\mathbf{L}_u \subseteq \mathbf{R}_u} [(1 - \lambda) \cdot \mathbf{Rel}(\mathbf{L}_u) - \lambda \cdot \mathbf{D}(\mathbf{P} \parallel \mathbf{Q}(\mathbf{L}_u))] \quad (1)$$

In this expression,  $\mathbf{Rel}(\mathbf{L}_u)$  denotes the total predicted relevance score for items in list  $\mathbf{L}_u$ , based on the output of the base recommender.  $\mathbf{D}(\mathbf{P} \parallel \mathbf{Q}(\mathbf{L}_u))$  represents a divergence function, commonly JSD, that measures the distance between the historical and recommended popularity distributions. The parameter  $\lambda \in [0, 1]$  controls the trade-off between relevance and calibration: setting  $\lambda = 0$  reduces the model to pure relevance-based ranking, while  $\lambda = 1$  focuses solely on popularity alignment.

To construct  $\mathbf{Q}(\mathbf{L}_u)$  and  $\mathbf{P}$ , items are categorized based on their cumulative rating frequency, following a Pareto-based bucketing scheme [18]. Instead of uniformly dividing items by rank or frequency percentiles, the method aggregates the total number of ratings across all items and splits the catalog into three tiers based on cumulative contribution to this total. Specifically, the most popular items that collectively account for the first 20% of all ratings are designated as the *head* or popular items. Items contributing to the next 60% of cumulative ratings are labeled as the *mid*, and the remaining items, which account for the final 20%, are considered the *tail* or niche items. Each item  $i$  is assigned to a bucket label based on this cumulative distribution.

Using this classification, a user's historical interaction profile is mapped to a probability distribution  $\mathbf{P} = [p_1, p_2, \dots, p_k]$ , where each  $p_j$  represents the fraction of interactions falling into bucket  $j$ . Similarly, the distribution  $\mathbf{Q}(\mathbf{L}_u) = [q_1, q_2, \dots, q_k]$ , is computed over the items in the re-ranked recommendation list, enabling divergence computation between the user's past preferences and the system's output in terms of popularity exposure.

The CP re-ranking process involves evaluating multiple candidate permutations of  $\mathbf{L}_u$  and computing the combined objective function for each. In practice, exhaustive enumeration is computationally infeasible; therefore, greedy algorithms or beam search strategies are employed to explore high-potential reorderings efficiently. In our implementation, we adopt a greedy heuristic that starts from the initial recommendation list and iteratively evaluates candidate item swaps to minimize the overall objective, defined as a weighted combination of predicted relevance and divergence. At each step, the pair of items yielding the greatest improvement is swapped. This continues until no further gain is observed or a maximum number of iterations is reached. This strategy offers a computationally practical and model-agnostic approach to approximating the optimal trade-off between accuracy and calibration. Since the CP method operates on the recommendation list generated by any underlying model, it is model-agnostic and easily integrated into existing recommender pipelines without requiring architectural changes.

This method offers a robust solution to the issue of popularity bias by enabling user-centered calibration. Rather than enforcing global diversity or fairness constraints, it directly models each user's historical

preferences regarding item popularity and attempts to replicate that structure in the output. This design yields recommendations that are not only accurate but also aligned with the user's implicit expectations regarding item popularity. The framework's modularity, particularly its use of a tunable trade-off parameter  $\lambda$ , makes it highly adaptable to different use cases, ranging from bias mitigation to fairness enhancement. By adjusting the balance between accuracy and calibration, CP enables practitioners to tailor system behavior to the needs of their user base while retaining flexibility in deployment.

#### 4. DIVERGENCE MEASURES FOR POPULARITY CALIBRATION

In the CP framework, the divergence measure plays a critical role in quantifying the mismatch between the popularity distribution of recommended items and that of the user's historical interactions. Different divergence metrics impose varying sensitivities, symmetries, and penalization behaviors, which can significantly influence the resulting recommendation list. In this study, we consider a range of divergence functions with varied theoretical properties, aiming to explore their suitability for personalized calibration. The following provides a detailed explanation of the considered divergence measures.

- **Jensen–Shannon Divergence (JSD) [5]:** It is the most widely used divergence metric in calibration settings, particularly in the original CP framework. It is a symmetrized and smoothed version of the Kullback–Leibler divergence, defined as:

$$JSD(P||Q) = \frac{1}{2}KL(P||M) + \frac{1}{2}KL(Q||M) \quad (2)$$

Here,  $KL(P||M) = \sum_i p(i) \log \frac{P(i)}{Q(i)}$  is the Kullback–Leibler divergence and  $M = \frac{1}{2}(P + Q)$ .

JSD has a bounded range  $[0, \log 2]$  and is symmetric, making it stable for use in systems where both over- and under-representation of popularity buckets need to be penalized. Its smoothness makes it well-suited for general-purpose calibration, especially for users with mixed preferences. However, its sensitivity might be insufficient for strongly niche profiles, limiting its correction power in highly biased cases.

- **Kullback–Leibler (KL) divergence [19]:** This metric, in contrast, is asymmetric and emphasizes cases where the predicted distribution fails to cover regions of the user's distribution. It is given by:

$$KL(P||Q) = \sum_i p(i) \log \frac{P(i)}{Q(i)} \quad (3)$$

KL is sensitive to small values in  $Q$ , especially when  $Q(i)$  is close to zero. This means it heavily penalizes omissions, which can be valuable for users whose preferences lie in the tail of the popularity spectrum. However, its asymmetry might over-penalize certain imbalances, and it may not be robust when  $Q$  has support gaps. In the CP context, KL could be powerful for niche-preferring users, as it heavily penalizes the omission of tail items. However, this harshness can destabilize calibration for users with broader interests or noisy histories.

- **Total Variation (TV) Distance [20]:** This metric is simple, symmetric, and bounded in the range  $[0, 1]$ , as formulated in Eq. 4. It measures the maximum difference between the two distributions across all buckets. TVD is especially effective in highlighting gross distributional shifts and can be appropriate for use cases where equal attention to under- and over-representation is desired. It is less sensitive to small fluctuations and is generally easy to interpret.

$$TV(P, Q) = \frac{1}{2} \sum_i |P(i) - Q(i)| \quad (4)$$

- **Wasserstein Distance [21]:** This metric also known as Earth Mover's Distance, offers a geometric perspective by measuring the minimum cost of transforming one distribution into another. For one-dimensional discrete distributions, it can be expressed as:

$$W(P, Q) = \sum_i |CDF_P(i) - CDF_Q(i)| \quad (5)$$

where  $CDF_P(i)$  and  $CDF_Q(i)$  are the cumulative distribution functions of  $P$  and  $Q$ , respectively. Wasserstein distance captures the notion of how far *mass* must be moved to match one distribution to another. It might be ideal for users with gradually skewed profiles, where hard thresholding may over-correct. It offers smooth, interpretable adjustment across buckets.

- **Hellinger Distance [22]:** This is another symmetric metric with a probabilistic foundation, defined as:

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_i (\sqrt{P(i)} - \sqrt{Q(i)})^2} \quad (6)$$

Hellinger is always in the range  $[0, 1]$  and behaves similarly to Euclidean distance in the square root space. It provides smooth gradients and is robust to noise, making it a solid choice for systems with sparse or uncertain data. It balances fairness and calibration without aggressive penalization, suitable for general audiences.

- **Chi-Square Distance [23]:** This distance emphasizes large deviations in expected proportions and penalizes cases where the recommendation distribution significantly diverges from user preference proportions. It is defined as in Eq. 7. It is particularly sensitive to over-representations and is often more aggressive than symmetric metrics. It could outperform others in correcting overexposure to popular items, making it effective for fairness-aware calibration, especially for long-tail users.

$$\chi^2(P, Q) = \sum_i \frac{(P(i) - Q(i))^2}{P(i)} \quad (7)$$

- **Cosine Distance [24]:** This metric measures the angular dissimilarity between two vectors, disregarding their magnitude. It is defined as:

$$\text{Cosine}(P, Q) = 1 - \frac{\sum_i P(i) \cdot Q(i)}{\sqrt{\sum_i P(i)^2} \cdot \sqrt{\sum_i Q(i)^2}} \quad (8)$$

This distance focuses on the directionality of the vectors in probability space. In CP re-ranking, Cosine distance is valuable when preserving the structural shape of the user's historical popularity distribution is more important than matching absolute proportions. It is particularly well-suited for users whose interaction profiles are stable in structure but variable in intensity, such as periodic or light-touch users. Because it emphasizes the consistency of relative preferences across popularity buckets, Cosine distance can provide a more forgiving calibration strategy that avoids overreacting to scale differences.

- **Renyi Divergence [24]:** This is a parametric generalization of the KL divergence. It is defined as:

$$D_\alpha^{\text{Renyi}}(P||Q) = \frac{1}{\alpha-1} \log(\sum_i P(i)^\alpha Q(i)^{1-\alpha}) \text{ for } \alpha > 0, \alpha \neq 1 \quad (9)$$

Renyi divergence introduces a tunable parameter  $\alpha$  that adjusts the emphasis on different parts of the distribution. As  $\alpha \rightarrow 1$ , the divergence converges to KL. At low  $\alpha$ , the divergence is more forgiving and behaves similarly to Total Variation, while higher values of  $\alpha$  place more weight on regions where  $Q$  fails to capture  $P$ . This tunability allows CP to adapt calibration strength across different user profiles. In practice, higher  $\alpha$  values may be used for long-tail or fairness-critical users, where strong penalties are needed for underrepresentation of niche items. Conversely, lower values may benefit broad-interest users, where soft alignment is preferable. Note that we set  $\alpha = 0.5$  in our experiments.

- **Bhattacharyya Distance [25]:** It measures the degree of overlap between two distributions and is commonly used in probabilistic classification. It is defined as:

$$D_{Bhat}(P, Q) = -\ln\left(\sum_i \sqrt{P(i) \cdot Q(i)}\right) \quad (10)$$

This symmetric measure rewards high similarity between corresponding elements of the distributions. In the CP setting, Bhattacharyya distance might be effective for moderate or balanced users, where the goal is to maximize overlap rather than punish misalignment. It avoids the sharp penalties associated with asymmetric metrics like KL or Chi-Square, offering a smoother calibration curve. Moreover, its probabilistic nature makes it robust under conditions of data sparsity or noisy bucket definitions, where harsh divergence may destabilize optimization.

Each divergence function provides a unique lens through which calibration can be achieved. Symmetric distances like JSD, TVD, and Hellinger are more balanced and stable, making them suitable for general calibration. Asymmetric or skew-sensitive measures like KL and Chi-Square can offer sharper correction but may need careful tuning. Wasserstein, by modeling redistribution effort, offers an intuitive trade-off between structural flexibility and penalization. In practice, the ideal divergence function may vary depending on user behavior patterns, system objectives, and the granularity of the popularity bucketing.

## 5. EXPERIMENTAL STUDIES

This section describes the datasets, evaluation metrics, and experimental setup employed in this study, and subsequently presents and analyzes the obtained results.

### 5.1. Datasets

To evaluate the effectiveness of different divergence metrics within the CP framework, we conducted experiments on two real-world benchmark datasets from distinct domains: MovieLens-1M (MLM) and Douban Books (DB) [7]. These datasets represent user–item interactions in the domains of movies and books, respectively, allowing for a cross-domain assessment of calibration performance. Both datasets employ a 5-point rating scale (ranging from 1 to 5) to express user preferences and exhibit a sparse rating structure characteristic of large-scale RS data. Compared to MovieLens-1M, Douban Books is notably larger and introduces additional challenges due to its higher sparsity and greater diversity in item popularity distribution. This combination of datasets enables a robust and comparative evaluation of the proposed calibration methods across different data densities, item catalogs, and user behavior patterns. Further details regarding dataset statistics and pre-processing steps are summarized in Table 2.

**Table 2.** Detailed information about datasets

Dataset	Domain	#Users	#Items	#Ratings	Sparsity (%)
MLM	Movie	6,040	3,952	1,000,209	95.7
DB	Book	13,024	22,347	792,062	99.7



## 5.2. Evaluation Metrics

To comprehensively assess the performance of different divergence-based calibration strategies, we employ four evaluation metrics that capture various aspects of recommendation quality: accuracy, calibration alignment, popularity bias, and diversity. These include *Precision@k*, *Overall Similarity Error (OSE)*, *Average Recommendation Popularity (ARP)*, and *Aggregate Diversity* [6].

*Precision@k* is a standard accuracy metric in RSs, measuring the proportion of relevant items among the top- $k$  recommendations. Given a user  $u$  and a set of recommended items  $L_u$ , precision is defined as:

$$\text{Precision@k} = \frac{|\{i \in L_u : i \in R_u^{true}\}|}{k} \quad (11)$$

where  $R_u^{true}$  denotes the ground truth relevant items for user  $u$ . This metric reflects the system's ability to place relevant items in the top-ranked positions.

*OSE* is a novel metric proposed in this study to evaluate the degree of alignment between the popularity distribution of recommended items and the user's historical preference distribution, independently of the divergence function used in re-ranking. Unlike metrics that rely on the same divergence function for both optimization and evaluation, *OSE* provides a divergence-agnostic measure of calibration quality. It is defined as:

$$OSE = \sum_i |P(i) - Q(i)| \quad (12)$$

where  $P(i)$  and  $Q(i)$  represent the proportions of popularity bucket  $i$  in the user's history and the recommended list, respectively. As a simple yet effective calibration indicator, *OSE* enables consistent comparisons across different divergence configurations and user groups, making it a key contribution of this work.

Although the *OSE* is mathematically equivalent to twice the TVD (see Eq. 4), its role in this study is conceptually distinct. Rather than serving as a divergence function for optimization, *OSE* is introduced as a simple, interpretable, and *divergence-independent* evaluation metric. Its purpose is to provide a consistent measure of calibration quality across different divergence-based CP variants, regardless of the specific divergence function used during re-ranking. In contrast to TVD, which has previously been employed as part of the optimization objective, *OSE* is applied solely at the evaluation stage.

*ARP* measures the mean global popularity of the items recommended to users. Formally, it is defined as:

$$ARP = \frac{1}{|U|} \sum_{u \in U} \frac{1}{|L_u|} \sum_{i \in L_u} pop(i) \quad (13)$$

where  $pop(i)$  denotes the popularity of item  $i$ , computed as the ratio of the number of users who have interacted with  $i$  to the number of all users. *ARP* provides insight into popularity bias, with lower *ARP* values indicating a shift toward recommending less globally popular (i.e., more niche) items.

*Aggregate Diversity* measures the overall variety of items recommended across the entire user population. It is defined as the total number of unique items appearing in the recommendation lists across all users:

$$\text{Aggregate Diversity} = \frac{|\cup_{u \in U} L_u|}{|I|} \quad (14)$$

where  $I$  is the set of all items in the catalog. This metric reflects the system’s ability to maximize catalog coverage and support content discovery. High aggregate diversity indicates that the system avoids repetitively recommending the same popular items and instead utilizes a broader portion of the item space.

Together, these metrics provide a comprehensive evaluation of divergence-based calibration performance, capturing local relevance, alignment with user-specific popularity trends, mitigation of global popularity bias, and catalog-level diversity. Our proposed *OSE* metric, in particular, enables an objective and interpretable assessment of calibration quality regardless of the divergence measure employed during optimization.

### 5.3. Experimentation Methodology

We adopt a leave-one-out cross-validation strategy to evaluate the performance of divergence-based calibrated recommendation methods [26, 27]. For each user, a single interaction is randomly withheld as the test instance, while the remaining users and all their interactions are used for training. Using the trained base models, Spherical  $k$ -Means (SKM) [28] and Variational Autoencoder for Collaborative Filtering (VAECF) [29], we generate predicted relevance scores for every item in the dataset for the test user. Then, items are ranked in descending order based on their predicted scores. The top-100 items form the candidate list  $R_u$ , which serves as input to the CP re-ranking procedure.

This process is repeated for all users in the dataset to ensure consistent evaluation. On top of the base predictions, CP re-ranking is applied using each of the nine divergence functions introduced in Section 4. To control the trade-off between relevance and calibration when applying CP strategy, the  $\lambda$  parameter is fixed at 0.5 throughout all experiments. Lastly, the final top-10 recommendation lists are evaluated using the accuracy, calibration, and diversity metrics described in Section 5.2, which are calculated individually for each user and then averaged across the entire user set. Note that all algorithms are implemented using the Cornac library [30], a widely adopted Python framework for research in RS.

### 5.4. Results and Discussion

This section presents and analyzes the experimental outcomes of the proposed approach. The results are divided into two parts: general results, which reflect overall performance across all users, and group-based results, which examine how different divergence functions perform across user segments with varying popularity preferences.

#### 5.4.1. Overall performance of divergence measures

Table 3 summarizes the overall performance of each divergence-based calibration method across the two datasets, MLM and DB, and two collaborative filtering algorithms, VAECF and SKM. The table reports results for four evaluation metrics: *Precision* ( $\uparrow$ ), *OSE* ( $\downarrow$ ), *ARP* ( $\downarrow$ ), and *Aggregate Diversity* ( $\uparrow$ ). Arrows indicate the desired direction of improvement for each metric. For readability, the names of divergence functions are abbreviated (e.g., *jsd*, *kl*, *chisq*). This structured presentation enables a direct comparison of how each divergence function impacts the trade-off between accuracy, calibration quality, popularity bias, and catalog coverage. Note that the best-performing value for each evaluation metric in each setting is highlighted in bold in the table. Additionally, we performed paired  $t$ -tests to determine whether the observed differences between the best and the second-best methods, in terms of a specific criterion, are statistically significant. Accordingly, we highlighted with an asterisk (\*) the results that are significant at the 95% confidence level.

The results presented in Table 3 reveal clear patterns regarding the impact of divergence function choice on the performance of the CP framework. While the original CP method was introduced using the JSD

as its core distance measure, our findings suggest that alternative divergence functions can offer significant improvements, particularly when calibration quality and bias mitigation due to popularity are prioritized.

Across both datasets and collaborative filtering algorithms, the Chi-Square (*chisq*) distance consistently achieves the lowest *OSE* and *ARP* values. This suggests that Chi-Square is the most effective at aligning the popularity distribution of recommended items with the user’s historical preferences, while simultaneously encouraging the recommendation of less globally popular (tail) items. This behavior can be attributed to Chi-Square’s strong penalization of large relative differences, particularly when the expected (historical) preference  $P(i)$  is high but the predicted  $Q(i)$  is low. This aggressive correction mechanism makes it especially well-suited for users with skewed or niche preferences, for whom standard methods often over-recommend popular items. However, the *Precision* scores for Chi-Square are notably lower than those of other divergences, reflecting a trade-off between calibration and accuracy. This is an expected consequence of re-ranking methods that favor distributional alignment over item-level predicted relevance. Despite this, the increase in *Aggregate Diversity* with Chi-Square indicates that such calibration strategies are more effective at surfacing underrepresented items, which is a desirable outcome in fairness-aware and long-tail recommendation settings.

In contrast, JSD, Hellinger, and Bhattacharyya maintain higher precision scores across all settings, but show relatively poor calibration performance. These metrics are symmetric and smoother in their penalization, and hence, less reactive to small but important misalignments in distribution. For example, JSD averages the divergence from both  $P$  and  $Q$  to the mean distribution  $M$ , which leads to more balanced but diluted penalization. This makes it more appropriate for users with mainstream preferences, where strict calibration is less critical. Hellinger and Bhattacharyya share this property, offering gradual gradients that support stable optimization but may fail to enforce strong calibration constraints.

As can be followed by Table 3, Cosine distance, although commonly used in vector similarity tasks, shows moderate calibration improvement compared to JSD, likely due to its emphasis on directionality rather than magnitude. This allows it to preserve the structural shape of the user’s preference profile without overreacting to scale differences. Renyi divergence, with its tunable sensitivity parameter  $\alpha$ , performs as a flexible middle ground, offering significantly improved *OSE* and *ARP* scores over JSD while maintaining reasonable accuracy. Its adaptability allows for targeted penalization depending on the user group. In our setting, it appears to strike an effective balance, especially for users with semi-niche behaviors. Similarly, Total Variation (*tv*) achieves solid calibration with minimal relevance loss, as it equally penalizes all deviations and is not disproportionately affected by rare categories. On the other hand, Wasserstein distance, which measures the cumulative effort to reshape one distribution into another, performs particularly well in terms of *ARP* reduction and diversity. Its ability to model smooth transitions between popularity levels allows it to preserve the general structure of a user’s profile while encouraging exploration away from globally dominant items. This makes it effective for moderate users whose preferences are not sharply concentrated. Finally, *Aggregate Diversity* results confirm that divergence functions which prioritize calibration (e.g., Chi-Square, KL, Wasserstein) also maximize catalog coverage, revealing their potential for RSs aiming to improve content exposure and reduce systemic bias.

**Table 3.** Overall performance of divergence-based CP methods. Arrows indicate the desired direction, and the best-performing value for each evaluation metric in each setting is highlighted in bold.

Dataset	CF Algorithm	Divergence Measure	Precision $\uparrow$	OSE $\downarrow$	ARP $\downarrow$	Aggregate Diversity $\uparrow$
MLM	VAECF	<i>jsd</i>	<b>0.639</b>	1.011	0.314	0.167
		<i>hellinger</i>	0.637	0.958	0.309	0.174
		<i>bhattacharyya</i>	0.636	0.953	0.308	0.173
		<i>cosine</i>	0.636	0.926	0.305	0.177
		<i>tv</i>	0.634	0.882	0.303	0.177
		<i>renyi</i>	0.630	0.847	0.297	0.178
		<i>kl</i>	0.626	0.763	0.289	0.183
		<i>wasserstein</i>	0.626	0.805	0.289	0.183
		<i>chisq</i>	0.608	<b>0.571*</b>	<b>0.267*</b>	<b>0.189*</b>
	SKM	<i>jsd</i>	<b>0.477</b>	1.431	0.427	0.023
		<i>hellinger</i>	0.475	1.408	0.424	0.023
		<i>bhattacharyya</i>	0.474	1.356	0.418	0.026
		<i>cosine</i>	0.472	1.306	0.412	0.026
		<i>tv</i>	0.470	1.282	0.409	0.022
		<i>renyi</i>	0.466	1.218	0.402	0.024
		<i>kl</i>	0.465	1.194	0.399	0.029
		<i>wasserstein</i>	0.463	1.141	0.393	0.029
		<i>chisq</i>	0.430	<b>0.801*</b>	<b>0.340*</b>	<b>0.033*</b>
DB	VAECF	<i>jsd</i>	<b>0.224</b>	1.304	0.115	0.026
		<i>hellinger</i>	0.224	1.293	0.114	0.027
		<i>bhattacharyya</i>	0.224	1.267	0.113	0.028
		<i>cosine</i>	0.224	1.272	0.113	0.030
		<i>tv</i>	0.222	1.240	0.112	0.030
		<i>renyi</i>	0.223	1.210	0.111	0.030
		<i>kl</i>	0.221	1.120	0.108	0.032
		<i>wasserstein</i>	0.221	1.057	0.103	0.037
		<i>chisq</i>	0.212	<b>0.853*</b>	<b>0.092*</b>	<b>0.041*</b>
	SKM	<i>jsd</i>	<b>0.164</b>	1.420	0.121	0.005
		<i>hellinger</i>	0.162	1.391	0.120	0.005
		<i>bhattacharyya</i>	0.158	1.258	0.115	0.006
		<i>cosine</i>	0.164	1.381	0.120	0.006
		<i>tv</i>	0.162	1.357	0.119	0.006
		<i>renyi</i>	0.160	1.328	0.117	0.007
		<i>kl</i>	0.156	1.165	0.112	0.007
		<i>wasserstein</i>	0.153	1.043	0.105	0.007
		<i>chisq</i>	0.139	<b>0.819*</b>	<b>0.091*</b>	<b>0.008*</b>

In summary, although CP was initially formulated with JSD, our experiments demonstrate that selecting a divergence function tailored to the system’s calibration and fairness goals can lead to significantly better outcomes. Chi-Square stands out for its effectiveness in calibration and diversity, whereas Renyi and Wasserstein provide versatile alternatives that balance accuracy and fairness. These findings underscore the importance of considering divergence function selection as a crucial design decision in calibrated RSs.

#### 5.4.2. Group-Based Comparison

To gain deeper insight into how different divergence functions support users with varying popularity preferences, we also perform a group-based evaluation. Users are categorized into three groups based on the distribution of item popularity in their historical interactions. Specifically, we construct each user’s popularity profile across predefined buckets (*head*, *mid*, and *tail*) and group them as follows:

- $G_1$  (*Mainstream Users*): Users whose historical interactions are dominated by popular (*head*) items.

- $G_2$  (*Balanced Users*): Users with a more evenly distributed interaction pattern across popularity levels.
- $G_3$  (*Niche Users*): Users who predominantly interact with less popular (*tail*) items.

This classification enables us to evaluate whether certain divergence functions are better suited to tailoring recommendations for specific user types. In particular, it enables us to assess whether calibrated methods can effectively mitigate popularity bias and enhance personalization for long-tail users ( $G_3$ ) without compromising performance for mainstream users ( $G_1$ ).

To assess how calibrated recommendation methods adapt to different user profiles, we analyze performance across three user groups, i.e.,  $G_1$ ,  $G_2$ , and  $G_3$ . Group-specific results are illustrated in Figs 1 through 4, each corresponding to one dataset–model pair: Fig. 1 (MLM–VAECF), Fig. 2 (MLM–SKM), Fig. 3 (DB–VAECF), and Fig. 4 (DB–SKM). Each figure reports *Precision*, *OSE*, *ARP*, and *Aggregate Diversity* scores, allowing us to identify which functions are most effective for each user group.

The analysis of our group-based evaluation reveals several broad trends regarding the performance of divergence functions across different user groups, as well as their implications for calibration in RSs. Overall, when examining the *Precision* metric, we observe that the balanced user group ( $G_2$ ) consistently outperforms the others, except for the DB-SKM configuration (see Fig. 4). This suggests that users whose historical interactions are evenly distributed across popularity buckets tend to benefit most from the system’s ability to align recommendations with their varied preferences. The balanced nature of  $G_2$  appears to provide an optimal environment where the divergence functions can fine-tune recommendations effectively, resulting in a higher degree of accuracy in matching user interests.

In contrast, when considering the calibration metrics (i.e., *OSE*), the results indicate that the mainstream users ( $G_1$ ) achieve the best calibration outcomes, except for the MLM-VAECF configuration (see Fig. 1). Mainstream users, whose interactions are dominated by popular items, seem to serve as the primary target during the calibration process. The high calibration scores for  $G_1$  suggest that the RS is particularly adept at replicating the established popularity profile of these users. This phenomenon may stem from the fact that popular items generally provide more robust statistical signals, making them easier to optimize during the calibration phase, although it might also suggest that the calibration strategy is inherently biased towards reinforcing existing popularity patterns.

Turning to the *ARP* metric, the findings indicate that the niche user group ( $G_3$ ) consistently exhibits the best performance in nearly all configurations, except for DB-SKM (see Fig. 4). This result is particularly insightful, as  $G_3$  users predominantly interact with less popular, long-tail items. The elevated *ARP* values for  $G_3$  indicate that the system is capable of accurately capturing and prioritizing the specific, less mainstream interests of these users. The enhanced *ARP* performance is also reflected in the aggregate diversity metric, where  $G_3$  typically registers higher diversity scores. Such an outcome suggests that the divergence functions, especially those that encourage diversity, can extend the recommendation lists to include a broader range of items that are more representative of the long tail, thereby better serving niche users.

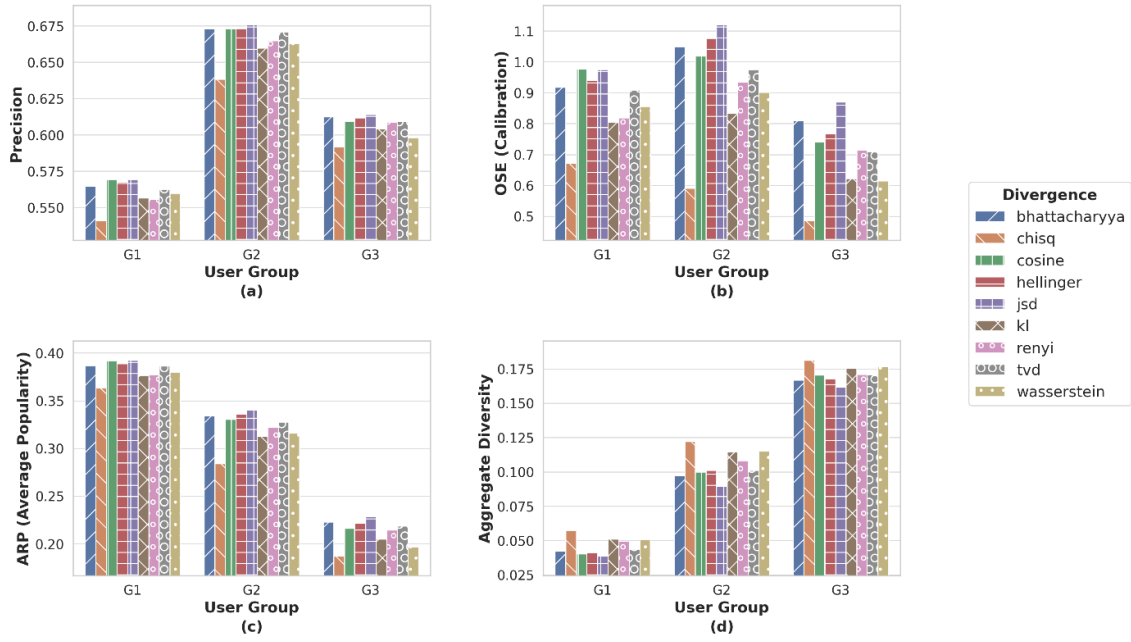


Figure 1. Group-based performance comparison of divergence functions on MLM with VAECF for (a) *Precision*, (b) *OSE*, (c) *ARP*, and (d) *Aggregate Diversity*.

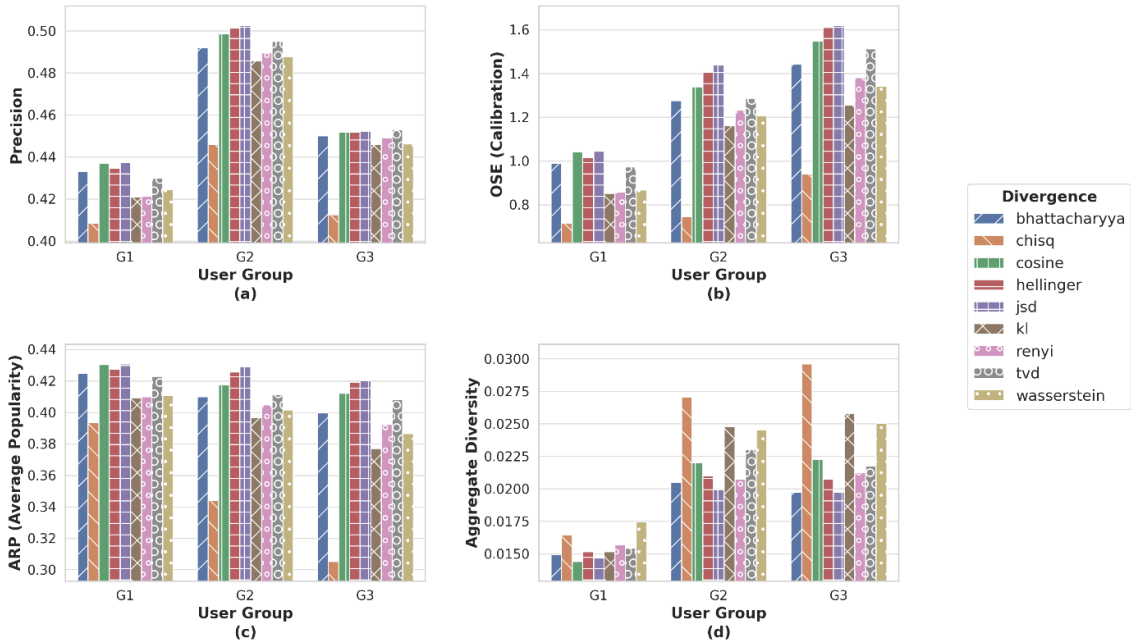
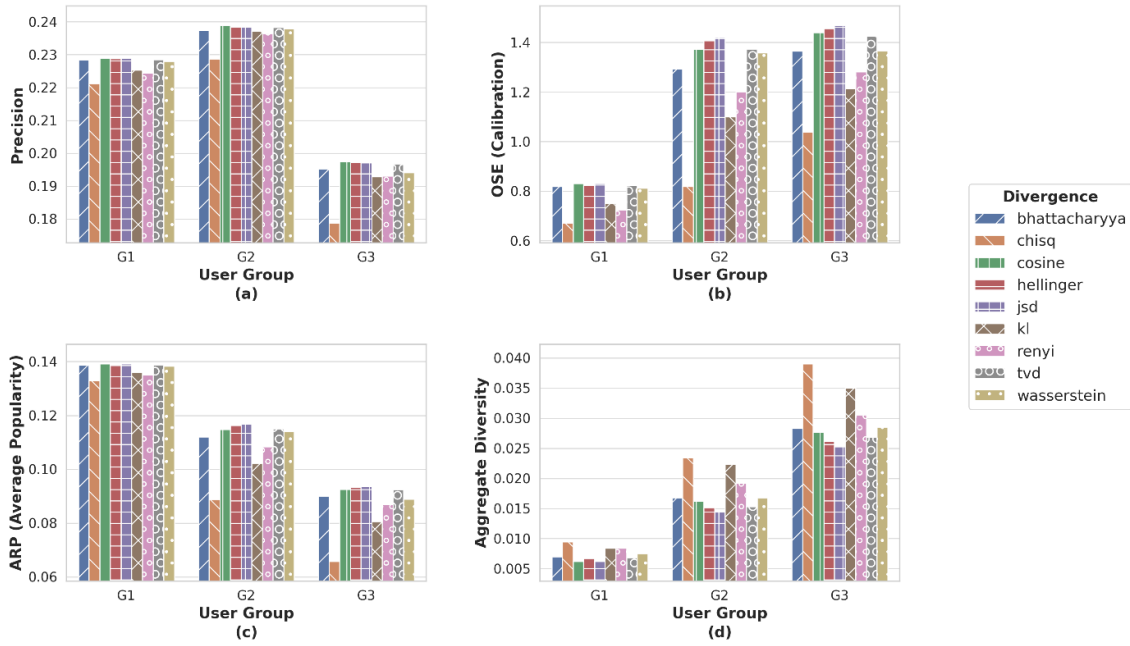
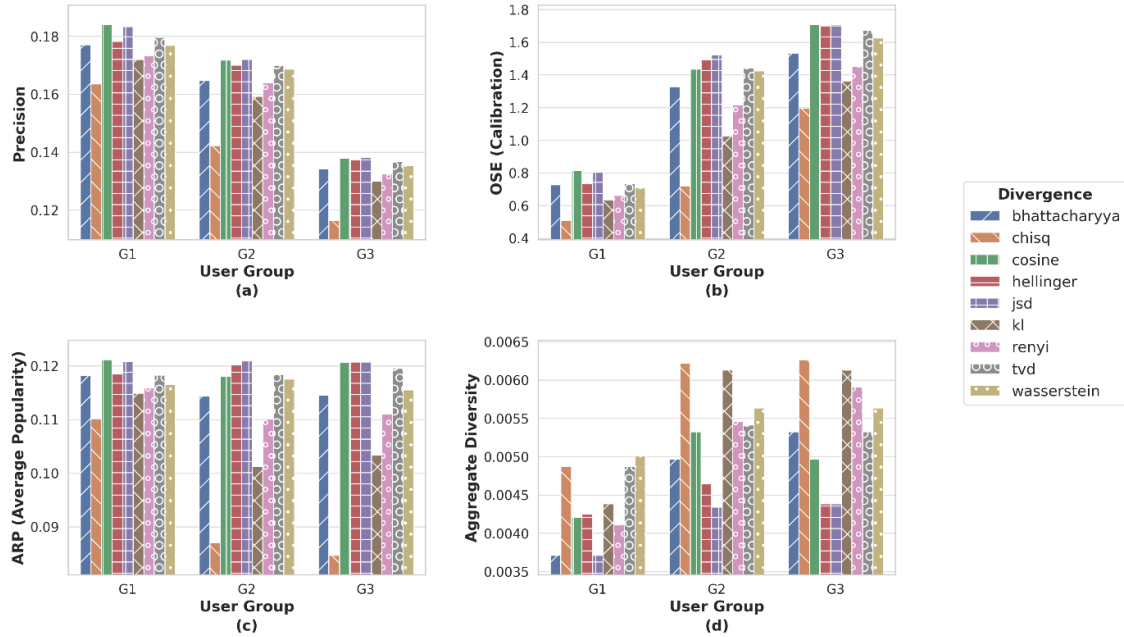


Figure 2. Group-based performance comparison of divergence functions on MLM with SKM for (a) *Precision*, (b) *OSE*, (c) *ARP*, and (d) *Aggregate Diversity*.



**Figure 3.** Group-based performance comparison of divergence functions on DB with VAECF for (a) *Precision*, (b) *OSE*, (c) *ARP*, and (d) *Aggregate Diversity*.



**Figure 4.** Group-based performance comparison of divergence functions on DB with SKM for (a) *Precision*, (b) *OSE*, (c) *ARP*, and (d) *Aggregate Diversity*.

In the context of divergence metrics, comparing the two main approaches, Bhattacharyya and Chi-Square offers additional insights. The Bhattacharyya divergence tends to generate recommendations that closely mirror the historical popularity profiles of users. This is particularly advantageous for achieving high precision in groups like  $G_2$ , where a balanced interest distribution allows for accurate recommendations. However, this alignment sometimes comes at the expense of overall diversity, particularly for mainstream users, since the system may overly concentrate on popular items. On the other hand, the Chi-Square divergence, although sometimes showing a slight decrease in *Precision*,

appears to foster greater recommendation diversity. It manages to mitigate the inherent popularity bias by promoting a wider range of items, which is especially beneficial for mainstream users in the calibration process and for niche users who benefit from exposure to long-tail items.

When evaluating the performance across different dataset–model combinations, it is evident that model architecture and dataset characteristics further modulate the behavior of divergence functions. For instance, while the precision performance of  $G_2$  remains robust in most configurations, the calibration effectiveness favoring  $G_1$  is more pronounced in settings other than MLM-VAECF (see Fig. 1). Similarly, *ARP* and *Aggregate diversity* metrics consistently underscore the strength of the recommendations for  $G_3$ , except in the DB-SKM scenario (see Fig. 4), indicating that the interplay between model, data, and divergence strategy can yield variable outcomes.

In summary, the observed trends highlight the need for user-group-specific strategies in RSs. The superior precision of the balanced group ( $G_2$ ) across most settings suggests that systems can achieve high accuracy by leveraging the naturally diverse interests of these users. Meanwhile, the calibration results for mainstream users ( $G_1$ ) indicate that reinforcing popular trends can be effective, albeit with the risk of perpetuating popularity bias. Finally, the strong *ARP* and *diversity* outcomes for niche users ( $G_3$ ) highlight the potential of divergence functions to enhance personalization by broadening the recommendation spectrum for users with non-mainstream tastes. These insights advocate for a more nuanced approach in the design of divergence-based RSs; one that tailors calibration and diversity enhancement strategies according to the unique popularity profiles of different user segments.

## 6. LIMITATIONS AND FUTURE WORK

While this study offers a comprehensive evaluation of divergence-based calibrated re-ranking strategies, several limitations should be acknowledged. First, a key challenge arises from the inherent data sparsity present in real-world datasets. For instance, the DB dataset used in our experiments exhibits an extreme sparsity level of approximately 99.7%. Although our approach demonstrates reasonable performance under such conditions, the limited number of user interactions substantially restricts the statistical reliability of observed item popularity distributions. This constraint hinders the effectiveness of calibration, particularly when divergence-based methods depend on estimating fine-grained popularity profiles at the individual level. Divergence functions with smoother gradient behavior (e.g., Hellinger or Bhattacharyya) may be somewhat more robust in these settings; however, data sparsity remains a structural bottleneck for all calibration-aware approaches. Developing specialized techniques that explicitly account for sparse-user behavior, potentially through hybrid or data-augmentation strategies, remains an open avenue for future research.

Second, our approach applies calibrated re-ranking as a post-processing step to two specific collaborative filtering models (VAECF and SKM). While this design highlights the model-agnostic nature of the CP framework, it also limits generalizability to other algorithm families such as graph-based or sequence-aware recommenders. Exploring the interaction between divergence-based calibration and a broader class of recommender backbones could yield further insights. Third, the  $\lambda$  parameter, which controls the trade-off between accuracy and calibration, was fixed at 0.5 in all experiments. Although this choice provides a neutral baseline for comparison, it may not represent the optimal setting for every divergence function, dataset, or user group. Future work could incorporate adaptive or learned weighting strategies to tune this parameter dynamically.

These limitations point toward important directions for extending this work, particularly in enhancing robustness under sparse data conditions, expanding model coverage, and improving calibration adaptability.



## 7. CONCLUSION

Popularity bias remains one of the most significant challenges in RSs, often resulting in reduced content diversity, limited personalization, and unfair treatment of niche items and users. To address this issue, we conducted a comprehensive investigation into calibrated popularity-based re-ranking methods by systematically evaluating alternative divergence measures within the prominent and well-known CP framework. Our evaluation, utilizing two distinct datasets (MovieLens-1M and Douban Book) and two representative collaborative filtering models (VAECF and Spherical  $k$ -Means), encompassed multiple divergence functions, including the Jensen–Shannon, Chi-Square, and Wasserstein divergences, among others. By considering both overall and user-group-specific performances, we aimed to understand how different divergence metrics affect the balance between recommendation accuracy, diversity, and the alignment with users' historical popularity preferences.

Our findings suggest that the choice of divergence measures has a significant impact on the recommendation outcomes. The Chi-Square divergence consistently outperformed other measures in terms of calibration quality and mitigation of popularity bias, although it exhibited lower precision scores. In contrast, symmetric and smoother divergences such as Jensen–Shannon, Bhattacharyya, and Hellinger generally achieved higher precision but showed relatively modest calibration improvements. Divergence measures, such as Wasserstein and Renyi, presented balanced profiles, providing substantial calibration and diversity improvements while maintaining reasonable recommendation accuracy.

Furthermore, our detailed group-based analysis revealed insightful patterns concerning user segmentation based on popularity preferences. Balanced users, with evenly distributed historical interactions, typically achieved the highest precision across most settings, suggesting that these users benefit most when recommendation strategies precisely align with their diverse interest distributions. Mainstream users, whose preferences concentrate on popular items, demonstrated the best calibration performance, underscoring the models' effectiveness in capturing strong popularity signals prevalent in their interaction histories. Meanwhile, niche users, who primarily engage with less popular items, benefited considerably from divergence measures that promote higher diversity and reduced recommendation popularity, thereby effectively enhancing personalization by capturing their specific interests.

Overall, these insights underscore the complexity inherent in popularity calibration and highlight the importance of carefully selecting or combining divergence metrics based on system-level objectives and user characteristics. Practitioners aiming to mitigate popularity bias in real-world RSs should adopt flexible, user-segment-aware calibration strategies, balancing precision, diversity, and fairness to meet diverse user expectations effectively.

## ACKNOWLEDGMENT

During the preparation of certain sections of this manuscript, the author utilized ChatGPT to improve grammar, enhance clarity, and refine language. All content generated with the assistance of this tool was subsequently reviewed, revised, and approved by the authors, who take full responsibility for the final version of the manuscript.

## CONFLICT OF INTEREST

The author stated that there are no conflicts of interest regarding the publication of this article.

## CRedit AUTHOR STATEMENT

Emre Yalcin: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing – Original Draft, Visualization.

## REFERENCES

- [1] Zangerle E, Bauer C. Evaluating recommender systems: survey and framework. *ACM Comput Surv* 2022; 55(8): 1-38.
- [2] Ahanger AB, Aalam SW, Bhat MR, Assad A. Popularity bias in recommender systems-a review. In: *Int Conf Emerging Technol Comput Eng*; February 2022; Cham, Switzerland. Cham: Springer. pp. 431-444.
- [3] Abdollahpouri H, Mansoury M, Burke R, Mobasher B. The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286*.
- [4] Mansoury M, Abdollahpouri H, Pechenizkiy M, Mobasher B, Burke R. Feedback loop and bias amplification in recommender systems. In: *29th ACM Int Conf Inf Knowl Manag*; 2020; New York, NY, USA. New York: ACM. pp. 2145-2148.
- [5] Steck H. Calibrated recommendations. In: *12th ACM Conf Recommender Syst*; 2018; New York, NY, USA. New York: ACM. pp. 154-162.
- [6] Abdollahpouri H, Mansoury M, Burke R, Mobasher B, Malthouse E. User-centered evaluation of popularity bias in recommender systems. In: *29th ACM Conf User Modeling, Adaptation and Personalization*; 2021; New York, NY, USA. New York: ACM. pp. 119-129.
- [7] Yalcin E, Bilge A. Evaluating unfairness of popularity bias in recommender systems: A comprehensive user-centric analysis. *Inf Process Manag* 2022; 59(6): 103100.
- [8] Pardo L. *Statistical Inference Based on Divergence Measures*. Boca Raton, FL, USA: Chapman and Hall/CRC, 2018.
- [9] Kaya M, Bridge D. A comparison of calibrated and intent-aware recommendations. In: *13th ACM Conf Recommender Syst*; 2019; New York, NY, USA. New York: ACM. pp. 151-159.
- [10] Seymen S, Abdollahpouri H, Malthouse EC. A constrained optimization approach for calibrated recommendations. In: *15th ACM Conf Recommender Syst*; 2021; New York, NY, USA. New York: ACM. pp. 607-612.
- [11] da Silva DC, Manzato MG, Durão FA. Exploiting personalized calibration and metrics for fairness recommendation. *Expert Syst Appl* 2021; 181: 115112.
- [12] da Silva DC, Durão FA. Benchmarking fairness measures for calibrated recommendation systems on movies domain. *Expert Syst Appl* 2025; 269: 126380.
- [13] Cha SH. Comprehensive survey on distance/similarity measures between probability density functions. *Int J Math Model Methods Appl Sci* 2007; 1(2): 1.
- [14] Burke R. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093*.
- [15] Deldjoo Y, Jannach D, Bellogin A, Difonzo A, Zanzonelli D. Fairness in recommender systems: research landscape and future directions. *User Model User-Adap Interact* 2024; 34(1): 59-108.
- [16] Vassøy B, Langseth H. Consumer-side fairness in recommender systems: a systematic survey of methods and evaluation. *Artif Intell Rev* 2024; 57(4): 101.

- [17] Yalcin E, Bilge A. Popularity bias in personality perspective: An analysis of how personality traits expose individuals to the unfair recommendation. *Concurrency Comput Pract Exp* 2023; 35(9): e7647.
- [18] Dunford R, Su Q, Tamang E, Wintour A. The pareto principle. *Plymouth Stud Sci* 2014; 7(1): 140-148.
- [19] Cui J, Tian Z, Zhong Z, Qi X, Yu B, Zhang H. Decoupled kullback-leibler divergence loss. *Adv Neural Inf Process Syst* 2024; 37: 74461-74486.
- [20] Feng W, Liu L, Liu T. On deterministically approximating total variation distance. In: *ACM-SIAM Symp Discrete Algorithms (SODA)*; 2024; New York, NY, USA. New York: SIAM. pp. 1766-1791.
- [21] Nietert S, Goldfeld Z, Sadhu R, Kato K. Statistical, robustness, and computational guarantees for sliced wasserstein distances. *Adv Neural Inf Process Syst* 2022; 35: 28179-28193.
- [22] Li X, Liu Z, Han X, Liu N, Yuan W. An intuitionistic fuzzy version of hellinger distance measure and its application to decision-making process. *Symmetry* 2023; 15(2): 500.
- [23] Shen C, Panda S, Vogelstein JT. The chi-square test of distance correlation. *J Comput Graph Stat* 2022; 31(1): 254-262.
- [24] Huang Y, Xiao F, Cao Z, Lin CT. Higher order fractal belief Rényi divergence with its applications in pattern classification. *IEEE Trans Pattern Anal Mach Intell* 2023; 45(12): 14709-14726.
- [25] Baidari I, Honnikoll N. Bhattacharyya distance based concept drift detection method for evolving data stream. *Expert Syst Appl* 2021; 183: 115303.
- [26] Geroldinger A, Lusa L, Nold M, Heinze G. Leave-one-out cross-validation, penalization, and differential bias of some prediction model performance measures—a simulation study. *Diagn Progn Res* 2023; 7(1): 9.
- [27] Yalcin E, Bilge A. Treating adverse effects of blockbuster bias on beyond-accuracy quality of personalized recommendations. *Eng Sci Technol Int J* 2022; 33: 101083.
- [28] Salah A, Rogovschi N, Nadif M. A dynamic collaborative filtering system via a weighted clustering approach. *Neurocomputing* 2016; 175: 206-215.
- [29] Liang D, Krishnan RG, Hoffman MD, Jebara T. Variational autoencoders for collaborative filtering. In: *WWW* 2018; 2018; Lyon, France. New York: ACM. pp. 689-698.
- [30] Salah A, Truong QT, Lauw HW. Cornac: A comparative framework for multimodal recommender systems. *J Mach Learn Res* 2020; 21(95): 1-5.