

Beyond Preference Matching: Exploration-Aware Calibration for Diverse Recommendations

Carolina Toledo Ferraz
carolina.ferraz@mackenzie.br
Universidade Presbiteriana Mackenzie
São Paulo, SP

Marcelo Garcia Manzato
mmanzato@icmc.usp.br
Universidade de São Paulo
São Carlos, SP

ABSTRACT

Most calibration approaches in recommender systems focus on matching the distribution of user preferences observed in historical interactions. Although effective for improving preference representation, reproducing historical preference distributions may reinforce already concentrated user profiles, limiting content discovery and recommendation diversity. We propose Exploration-Aware Calibration (EAC), a calibration framework that explicitly incorporates category-level exploration into the recommendation process. Inspired by exploration-exploitation principles, EAC assigns exploration bonuses to underrepresented categories and integrates them into a calibration-based re-ranking strategy. We evaluate the proposed method using multiple recommendation algorithms in both movie and music domains. Results show that EAC consistently improves intra-list diversity, category coverage, and reduces recommendation concentration while preserving competitive levels of recommendation accuracy. Our findings demonstrate that introducing controlled exploration into calibration offers a practical and effective approach for balancing relevance and diversity in recommender systems.

KEYWORDS

recommender systems, calibration, diversity, filter bubbles, re-ranking, exploration-exploitation

1 INTRODUCTION

Recommender systems play a central role in online platforms by helping users navigate increasingly large catalogs of items. Over the years, research has primarily focused on improving recommendation accuracy, leading to significant advances in collaborative filtering, matrix factorization, and deep learning approaches [7, 15]. However, optimizing only for relevance may result in recommendation lists that repeatedly expose users to similar content, potentially reducing diversity and limiting opportunities for content discovery. Such concentration effects have been associated with the emergence of filter bubbles, in which users are predominantly exposed to content aligned with their historical preferences [11, 13].

To address this issue, calibration has emerged as a promising post-processing strategy. The main idea behind calibration is to align the distribution of item categories in recommendation lists with the distribution observed in the user’s historical interactions [21]. For example, if a user’s consumption history contains 50% drama, 30%

comedy, and 20% science fiction, a calibrated recommendation list should approximately preserve this distribution. Previous studies have shown that calibration can improve the representation of user preferences while maintaining competitive recommendation accuracy [1].

Despite its benefits, traditional calibration presents an important limitation. Since it aims to reproduce historical preference distributions, it may inadvertently reinforce existing concentration patterns. In other words, users with highly concentrated profiles may continue to receive recommendations dominated by the same categories, reducing exposure to potentially relevant but underrepresented content. Consequently, calibration may improve preference alignment without necessarily promoting discovery or mitigating recommendation concentration.

In this paper, we propose *Exploration-Aware Calibration* (EAC), a post-processing re-ranking strategy that integrates calibration with category-level exploration. Inspired by the exploration-exploitation principle commonly studied in recommender systems and bandit algorithms [8, 9], EAC introduces an exploration bonus that favors categories that are both underrepresented in the user’s profile and insufficiently covered in the recommendation list being constructed. This mechanism encourages the inclusion of less explored categories while preserving alignment with user preferences.

We evaluate the proposed approach in two distinct recommendation domains, movies and music, using the MovieLens and LastFM datasets. Experiments are conducted with three recommendation algorithms, namely Most Popular, ItemKNN, and Matrix Factorization. The evaluation considers both accuracy-oriented metrics and diversity-related metrics, including category coverage, intra-list diversity, and recommendation concentration. We also evaluate the proposed approach with different sets of users according to their preferences for niche/mainstream items. The results demonstrate that EAC consistently improves diversity and category exposure while preserving competitive levels of recommendation accuracy.

The main contributions of this work are threefold: (i) we introduce a novel exploration-aware calibration strategy that combines preference alignment and category-level exploration; (ii) we propose a re-ranking mechanism inspired by exploration-exploitation principles to promote underrepresented categories; and (iii) we provide an extensive evaluation across multiple recommendation algorithms and domains, analyzing the impact of the proposed approach on accuracy, diversity, and recommendation concentration.

This paper is organized as follows. Section 2 reviews related work on calibration, diversity, and exploration in recommender systems. Section 3 presents the proposed EAC framework, including the problem formulation, the traditional calibration baseline, and the exploration-aware re-ranking procedure. Section 4 describes the

experimental setup, covering the datasets, recommendation models, evaluation metrics, and research questions. Section 5 reports and discusses the experimental results, analyzing overall performance, hyperparameter selection, and user group behavior. Finally, Section 6 presents the final remarks and directions for future work.

2 RELATED WORK

2.1 Calibration in Recommender Systems

Calibration has emerged as an important paradigm for improving the alignment between recommendation lists and users' historical preferences. Rather than optimizing only relevance, calibrated recommendation approaches seek to preserve the distribution of categories observed in users' interaction histories [21]. The underlying assumption is that recommendation lists should not only contain relevant items but also reflect the diversity of interests expressed by the user.

Several studies have extended this idea to different recommendation scenarios. Silva and Durão [17] proposed a calibration framework and decision protocol for generating recommendation lists that balance relevance and category distributions. More recently, Souza and Manzato [20] investigated calibration strategies to mitigate popularity bias and improve fairness, demonstrating that calibration can contribute to more balanced recommendation lists while preserving recommendation quality. Recent works [4, 14, 19] have also explored calibration in different contexts, including popularity-bias mitigation, multi-category calibration, and calibration using large language models. These studies reinforce the importance of calibration as a mechanism for balancing recommendation objectives beyond accuracy.

2.2 Diversity, Filter Bubbles and Exploration

The growing concern regarding recommendation concentration has motivated extensive research on diversity and filter bubbles. Filter bubbles describe situations in which personalization mechanisms repeatedly expose users to content aligned with their previous interests, reducing opportunities for discovery and limiting content diversity [13]. Empirical studies have shown that recommender systems may contribute to narrow exposure patterns even within users' broader interests [11, 22]. Systematic reviews further highlight that recommendation concentration is often associated with popularity bias and insufficient exposure to alternative content categories [3].

To address these issues, prior work has proposed diversification and exploration mechanisms, including re-ranking methods, exposure balancing strategies, personality traits, and exploration-exploitation approaches inspired by multi-armed bandits [2, 9, 12, 18]. Such methods typically promote less exposed items or categories to increase diversity and catalog coverage. However, most exploration-oriented approaches do not explicitly consider calibration objectives and may alter the original preference distribution represented in recommendation lists.

2.3 Research Gap

Although calibration can improve recommendation diversity by representing the breadth of users' historical interests, it does not inherently promote exploration beyond those observed preferences.

Consequently, when historical profiles are highly concentrated, calibration may preserve rather than mitigate such concentration. In contrast, diversification and exploration techniques deliberately increase exposure to alternative content, but typically do not explicitly preserve preference calibration.

To the best of our knowledge, limited attention has been given to integrating category-level exploration directly into calibration mechanisms. This paper addresses this gap by proposing *Exploration-Aware Calibration* (EAC), a re-ranking strategy that combines preference alignment with controlled exploration of underrepresented categories. Unlike traditional calibration approaches, EAC explicitly encourages category discovery while maintaining consistency with users' historical interests.

3 METHODOLOGY

This section presents the proposed *Exploration-Aware Calibration* (EAC) framework. We first describe the traditional calibration strategy and then introduce a category-level exploration mechanism designed to increase exposure to underrepresented categories while preserving alignment with users' historical preferences.

3.1 Problem Definition

Let U be the set of users and I the set of items. For a given user $u \in U$, a recommender model produces a ranked list of candidate items based on a relevance score $Rel(u, i)$.

Each item is associated with one or more categories (e.g., movie genres or music tags). Using the user's historical interactions, we estimate a preference distribution over categories, denoted by P_u . Similarly, given a recommendation list L , we define Q_L as the distribution of categories represented in that list.

The objective of calibration is to generate recommendation lists whose category distribution is consistent with the user's historical preferences. At the same time, we seek to increase exposure to categories that are underrepresented in the user's profile, promoting content discovery and reducing recommendation concentration.

3.2 Traditional Calibration

Following previous work on calibrated recommendation [21], calibration is formulated as a re-ranking problem. During the construction of the recommendation list, each candidate item is evaluated according to both its estimated relevance and its contribution to matching the user's category distribution.

The calibration component is computed using the Kullback-Leibler (KL) divergence between the user's profile distribution P_u and the category distribution of the recommendation list after inserting a candidate item:

$$KL(P_u || Q_{L \cup i}) = \sum_{c \in C} P_u(c) \log \frac{P_u(c)}{Q_{L \cup i}(c)} \quad (1)$$

where C denotes the set of categories.

The traditional calibration score is given by:

$$Score_{cal}(i) = (1 - \lambda)Rel(u, i) - \lambda KL(P_u || Q_{L \cup i}) \quad (2)$$

where $\lambda \in [0, 1]$ controls the trade-off between recommendation relevance and calibration.

Although this strategy improves the alignment between recommendation lists and historical preferences, it may preserve existing concentration patterns when user profiles are highly skewed toward a small set of categories [12].

3.3 Exploration-Aware Calibration

To encourage content discovery, we introduce a category-level exploration mechanism inspired by the Upper Confidence Bound (UCB) exploration-exploitation principle [5]. In this paper, UCB is used as an inspiration for defining a deterministic category-level exploration bonus rather than as an online bandit algorithm with reward or regret estimation. The key idea is to provide additional opportunities for categories that are both underrepresented in the user's profile and insufficiently represented in the recommendation list being constructed.

During re-ranking, the algorithm maintains, for each category c , the number of times it has already appeared in the partial recommendation list, denoted by N_c . Let T denote the current size of the recommendation list.

We define an exploration bonus for each category as:

$$UCB_c = (1 - P_u(c)) \sqrt{\frac{\ln(T+1)}{N_c + 1}} \quad (3)$$

The first term, $(1 - P_u(c))$, assigns larger bonuses to categories that appear less frequently in the user's historical profile. The second term decreases as the category becomes more represented in the recommendation list, preventing excessive promotion of the same category.

Since items may belong to multiple categories, the exploration bonus of an item is computed as the average bonus of its associated categories:

$$UCB(i) = \frac{1}{|C_i|} \sum_{c \in C_i} UCB_c \quad (4)$$

where C_i represents the set of categories associated with item i .

3.4 Final Ranking Function

The proposed Exploration-Aware Calibration combines relevance, calibration, and exploration within a single scoring function:

$$Score(i) = (1 - \lambda)Rel(u, i) - \lambda KL(P_u || Q_{L \cup i}) + \beta UCB(i) \quad (5)$$

where $\beta \geq 0$ controls the influence of exploration.

When $\beta = 0$, the proposed framework reduces to traditional calibration. As β increases, the algorithm progressively promotes items associated with underrepresented categories, encouraging greater diversity and category coverage.

3.5 Re-ranking Procedure

The recommendation process starts from a candidate set generated by a base recommender model. The final recommendation list is then constructed iteratively.

At each iteration, the algorithm evaluates all remaining candidate items using the scoring function defined in Equation 5. The item with the highest score is selected and appended to the recommendation list. After selection, the category statistics are updated and the

Algorithm 1 Exploration-Aware Calibration (EAC)

Require: User u , candidate set R_u , user category profile P_u , relevance scores $Rel(u, i)$, list size K , calibration weight λ , exploration weight β

Ensure: Re-ranked recommendation list L_u

```

1:  $L_u \leftarrow \emptyset$ 
2:  $T \leftarrow 0$ 
3:  $N_c \leftarrow 0, \forall c \in C$ 
4: while  $|L_u| < K$  do
5:   for all  $i \in R_u \setminus L_u$  do
6:     Compute  $Q_{L_u \cup \{i\}}$ 
7:      $cal(i) \leftarrow KL(P_u || Q_{L_u \cup \{i\}})$ 
8:     for all  $c \in C_i$  do
9:        $UCB_c \leftarrow (1 - P_u(c)) \sqrt{\frac{\ln(T+1)}{N_c+1}}$ 
10:    end for
11:     $UCB(i) \leftarrow \frac{1}{|C_i|} \sum_{c \in C_i} UCB_c$ 
12:     $Score(i) \leftarrow (1 - \lambda)Rel(u, i) - \lambda cal(i) + \beta UCB(i)$ 
13:  end for
14:   $i^* \leftarrow \arg \max_{i \in R_u \setminus L_u} Score(i)$ 
15:   $L_u \leftarrow L_u \cup \{i^*\}$ 
16:  for all  $c \in C_{i^*}$  do
17:     $N_c \leftarrow N_c + 1$ 
18:  end for
19:   $T \leftarrow T + 1$ 
20: end while
21: return  $L_u$ 

```

process repeats until the desired recommendation length is reached. Algorithm 1 summarizes the proposed re-ranking procedure.

The resulting recommendation lists preserve alignment with users' historical preferences through calibration while simultaneously encouraging exposure to less explored categories through the exploration bonus.

4 EXPERIMENTAL SETUP

4.1 Datasets

The experiments were conducted in two recommendation domains: movies and music. For the movie domain, we used the MovieLens 100K dataset¹, a widely adopted benchmark in recommender systems research. For the music domain, we used the HetRec 2011 LastFM 2K dataset², which contains user–artist interactions and user-generated tags.

MovieLens provides explicit ratings and structured movie genres. LastFM provides implicit feedback through user–artist listening interactions and semantic descriptors in the form of tags. Thus, in MovieLens, categories correspond to movie genres, whereas in LastFM, categories correspond to music tags.

4.2 Data Preparation

For MovieLens, ratings greater than or equal to four were considered positive feedback. User category profiles were constructed from the genres of positively consumed movies in the training split,

¹<https://grouplens.org/datasets/movielens/100k/>

²<https://grouplens.org/datasets/hetrec-2011/>

Table 1: Datasets used in the experiments.

Dataset	Users	Items	Interactions	Categories
MovieLens 100K	876	1,447	54,839	19 genres
LastFM	1,869	11,288	85,395	500 tags

weighting each genre by its proportion among the user’s training movies.

For LastFM, all user–artist interactions were considered positive feedback, since the dataset represents implicit user preferences. User category profiles were constructed from the artists consumed in the training split, weighting each tag by its proportion among the tags of the user’s training artists, with each artist contributing equally regardless of listening frequency. To reduce sparsity and noise, only the top-500 most frequently assigned tags were retained, and artists without valid tag information after filtering were removed.

4.3 Experimental Protocol

For MovieLens, we adopted a temporal split per user. The five most recent positive interactions were assigned to the test set, the five preceding interactions were used for validation, and the remaining interactions were used for training.

Since LastFM does not provide timestamps for the user–artist interactions used in this work, we adopted a user-based random split with fixed seed. For each user, 5 interactions were assigned to the test set, 5 to validation, and the remaining interactions to training.

For both datasets, the training set was used to train the recommender models, the validation set was used to select the best λ and β configurations via a Multi-Attribute Utility Theory (MAUT) score [6] combining NDCG, ILD, Category Coverage, and GBI (equal weights, metrics normalized over the $\lambda \times \beta$ grid), and the test set to report the final results (see Subsection 4.7 for all metrics definitions).

In all experiments, items observed in the training set were removed from the recommendation candidates. Recommendation lists were evaluated at cutoff $K = 10$.

The calibration parameter was varied as: $\lambda \in \{0.1, 0.3, 0.5, 0.7\}$, and the exploration parameter was varied as: $\beta \in \{0.00, 0.01, 0.05, 0.10, 0.20\}$. When $\beta = 0$, EAC becomes equivalent to traditional calibration.

4.4 Recommendation Models

We evaluated the proposed method using three recommendation algorithms.

- **Most Popular (MP):** a non-personalized baseline that ranks items according to their interaction frequency in the training data.
- **ItemKNN:** an item-based nearest-neighbor collaborative filtering method that recommends items similar to those previously consumed by the target user [16]. The number of neighbors was empirically fixed to $k = 50$.
- **Matrix Factorization (MF):** a latent-factor collaborative filtering model that represents users and items in a shared

latent space and estimates preferences through vector interactions [7]. The number of latent factors was empirically fixed to $f = 50$.

These algorithms were selected because they represent different recommendation behaviors, ranging from a popularity-based baseline to neighborhood-based and latent-factor collaborative filtering models.

4.5 Compared Methods

For each recommendation model, we compared three strategies:

- **Original:** the recommendation list produced directly by the base model, without post-processing.
- **Traditional Calibration:** a post-processing re-ranking strategy that balances relevance and category calibration using KL divergence, following the calibrated recommendation principle proposed by Steck [21].
- **Exploration-Aware Calibration (EAC):** the proposed method, which extends traditional calibration by incorporating a category-level exploration bonus inspired by the UCB principle.

4.6 User Groups

To analyze whether EAC benefits different user types equally, we segment users into three groups based on their tendency to consume popular items. We first identify popular items as the top 20% of items ranked by number of training interactions. For each user u , we compute the *popularity ratio*:

$$\text{pop_ratio}(u) = \frac{|\mathcal{H}_u \cap \mathcal{P}|}{|\mathcal{H}_u|}, \quad (6)$$

where $\mathcal{H}_u$ is the set of items consumed by u in the training set and $\mathcal{P}$ is the set of popular items. Users are then classified by comparing their *pop_ratio* to the first (Q_1) and third (Q_3) quartiles of the distribution:

- **Niche** ($\text{pop_ratio} < Q_1$): users whose consumption is concentrated on less popular, niche items.
- **Balanced** ($Q_1 \leq \text{pop_ratio} \leq Q_3$): users with an intermediate mix of popular and niche items.
- **Mainstream** ($\text{pop_ratio} > Q_3$): users who predominantly consume popular items.

4.7 Evaluation Metrics

We evaluate recommendation quality from four perspectives.

Accuracy. We report Normalized Discounted Cumulative Gain (NDCG@ K), which measures ranking quality by assigning higher credit to relevant items appearing at earlier positions:

$$\text{NDCG}@K = \frac{\text{DCG}@K}{\text{IDCG}@K}, \quad \text{DCG}@K = \sum_{i=1}^K \frac{\mathbb{1}[r_i \in \mathcal{R}_u]}{\log_2(i+1)}, \quad (7)$$

where r_i is the item at position i , $\mathcal{R}_u$ is the set of relevant items for user u , and IDCG@ K is the ideal DCG achieved by a perfect ranking.

Diversity and coverage. We report Intra-List Diversity (ILD@ K), computed as the average pairwise dissimilarity between items in the list based on their category vectors:

Table 2: Comparison of Original, Traditional Calibration ($\lambda^*, \beta=0$), and EAC ($\lambda^*=0.7, \beta^*=0.20$) on the test set ($K=10$) for both datasets. Bold = best per model. MAUT is computed from the four displayed metrics. See note below for symbol legend.

Dataset	Model	Method	NDCG	ILD	CatCov	GBI	MAUT	
ML100K	Most Popular	Original	0.0589	0.6879	0.5395	0.2660	0.2500	
		Trad. ($\lambda=0.5$)	0.0519 ↓	0.7520 ↑	0.6765 ↑	0.1876 ↑	0.5467 ↑	
		EAC ($\lambda=0.7, \beta=0.20$)	0.0482 ↓↓	0.8138 ↑↑	0.7439 ↑↑	0.1467 ↑↑	0.7500 ↑↑	
	ItemKNN	Original	0.0762	0.7075	0.5247	0.2799	0.2500	
		Trad. ($\lambda=0.5$)	0.0663 ↓	0.7448 ↑	0.6440 ↑	0.2155 ↑	0.4551 ↑	
		EAC ($\lambda=0.7, \beta=0.20$)	0.0615 ↓↓	0.8193 ↑↑	0.7207 ↑↑	0.1631 ↑↑	0.7500 ↑↑	
	MF	Original	0.0696	0.7127	0.5141	0.3001	0.2500	
		Trad. ($\lambda=0.7$)	0.0565 ↓	0.7541 ↑	0.6626 ↑	0.2126 ↑	0.4459 ↑	
		EAC ($\lambda=0.7, \beta=0.20$)	0.0540 ↓*	0.8355 ↑↑	0.7352 ↑↑	0.1580 ↑↑	0.7500 ↑↑	
	LastFM	Most Popular	Original	0.0660	0.5246	0.5251	0.1449	0.0000
			Trad. ($\lambda=0.7$)	0.1008 ↑	0.5957 ↑	0.5736 ↑	0.1296 ↑	0.9256 ↑
			EAC ($\lambda=0.7, \beta=0.20$)	0.0991 ↑↓	0.6026 ↑↑	0.5792 ↑↑	0.1278 ↑↑	0.9880 ↑↑
ItemKNN		Original	0.1531	0.6424	0.2996	0.2718	0.2500	
		Trad. ($\lambda=0.7$)	0.1480 ↓	0.6679 ↑	0.4238 ↑	0.1871 ↑	0.6989 ↑	
		EAC ($\lambda=0.7, \beta=0.20$)	0.1439 ↓↓	0.7001 ↑↑	0.4298 ↑↑	0.1826 ↑↑	0.7500 ↑↑	
MF		Original	0.1390	0.6642	0.3640	0.2086	0.2500	
		Trad. ($\lambda=0.7$)	0.1336 ↓	0.6770 ↑	0.4763 ↑	0.1616 ↑	0.6562 ↑	
		EAC ($\lambda=0.7, \beta=0.20$)	0.1311 ↓↓	0.6894 ↑↑	0.4885 ↑↑	0.1566 ↑↑	0.7500 ↑↑	

Note: Any arrow (↑/↓) denotes a statistically significant difference (Wilcoxon signed-rank, $p < 0.05$); * = not significant. Traditional rows: one symbol [vs. Orig.]; EAC rows: two symbols [vs. Orig.][vs. Trad.]. Arrow color encodes Cohen's $|d|$ effect size: black negligible ($|d| < 0.2$), blue small ($0.2 \leq |d| < 0.5$), green medium ($0.5 \leq |d| < 0.8$), red large ($|d| \geq 0.8$).

GBI is lower-is-better (↑ = significant decrease).

$$\text{ILD@K} = \frac{2}{K(K-1)} \sum_{i=1}^K \sum_{j=i+1}^K \left(1 - \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|} \right), \quad (8)$$

where $\mathbf{v}_i$ is the binary category vector of item i . We also report Category Coverage (CatCov), defined as the average fraction of distinct categories covered within a single user's recommendation list:

$$\text{CatCov} = \frac{1}{|U|} \sum_{u \in U} \frac{|C(L_u)|}{|C|}, \quad (9)$$

where $C(L_u)$ is the set of distinct categories represented in user u 's list L_u , and C is the complete set of categories.

Recommendation concentration. We use the Genre Bubble Index (GBI), adapted from [10], for both datasets. In MovieLens, genres are movie genres; in LastFM, genres refer to the semantic categories represented by music tags. GBI is defined as:

$$\text{GBI}(L_u) = 1 - \frac{H(L_u)}{H_{\max}}, \quad (10)$$

where $H(L_u)$ is the Shannon entropy of the category distribution of list L_u and $H_{\max} = \ln |C|$ is the maximum possible entropy. Values closer to one indicate higher concentration; values closer to zero indicate more diverse category exposure. We report the mean GBI over all users.

Overall utility (MAUT). To summarize multi-objective performance, we report a Multi-Attribute Utility Theory (MAUT) score computed as the equal-weight mean of the four metrics above, each normalized to $[0, 1]$ over the set {Original, Traditional, EAC} within each (dataset, model) block:

$$\text{MAUT} = \frac{1}{4} \sum_{m \in \mathcal{M}} \hat{m}, \quad \hat{m} = \begin{cases} \frac{m - m_{\min}}{m_{\max} - m_{\min}} & \text{if higher is better,} \\ \frac{m_{\max} - m}{m_{\max} - m_{\min}} & \text{if lower is better (GBI),} \end{cases} \quad (11)$$

where $\mathcal{M} = \{\text{NDCG, ILD, CatCov, GBI}\}$.

Equal weights were adopted to avoid prioritizing any single evaluation dimension; alternative weighting schemes may lead to different trade-offs and are left for future investigation.

4.8 Research Questions

The experiments were designed to answer the following research questions:

- **RQ1:** Does Exploration-Aware Calibration reduce recommendation concentration compared to the original recommendations and to traditional calibration?
- **RQ2:** Does Exploration-Aware Calibration improve diversity and category coverage?
- **RQ3:** What is the impact of Exploration-Aware Calibration on recommendation accuracy?
- **RQ4:** Are the effects of Exploration-Aware Calibration consistent across recommendation algorithms and domains?
- **RQ5:** Does EAC benefit users consistently across different preference profiles (niche, balanced, and mainstream users)?
- **RQ6:** Is the magnitude of EAC's individual gain related to user profile concentration?

4.9 Statistical Analysis

To assess whether differences between methods are statistically significant, we applied the Wilcoxon signed-rank test on per-user metric scores. Effect sizes were computed using Cohen's d . Statistical significance was assessed at $\alpha = 0.05$.

5 RESULTS

5.1 Hyperparameter Selection

Figure 1 shows the MAUT score across the $\lambda \times \beta$ grid on the validation set. All six model-dataset combinations selected $\lambda^* = 0.7$ and $\beta^* = 0.20$ for EAC, the highest values in the respective grids. For Traditional Calibration ($\beta = 0$), $\lambda^* = 0.5$ was selected for ML100K/MostPopular and ML100K/ItemKNN; all other combinations selected $\lambda^* = 0.7$. The consistent preference for the maximum β suggests that greater exploration weight is generally beneficial; however, since $\beta^* = 0.20$ lies at the grid boundary and MAUT continues to increase toward it, larger values of β may yield further gains. We leave the exploration of broader β ranges for future work.

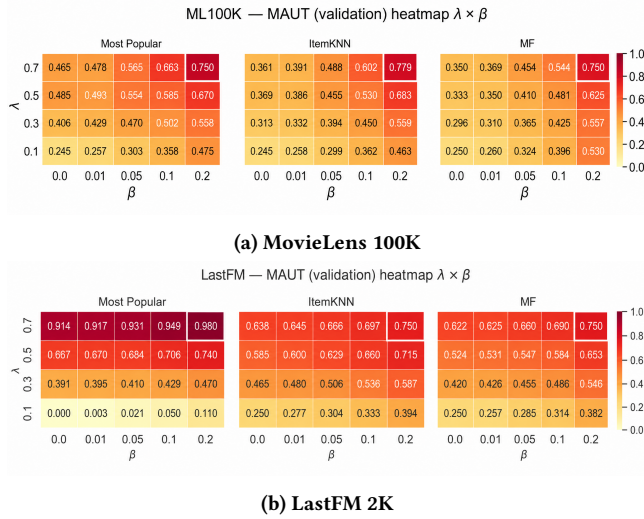


Figure 1: MAUT score on the validation set for each recommender model.

5.2 Overall Results

Table 2 reports test-set results for all methods on both datasets. Arrows and colors encode statistical significance and effect sizes as described in the table note.

RQ1 – Recommendation concentration. EAC significantly reduced GBI in all six model-dataset combinations. Effect sizes versus the original baseline are large in every case ($|d| = 0.86$ – 2.98): for example, MostPopular on ML100K drops from $\text{GBI} = 0.266$ to 0.147 . EAC also improves over Traditional Calibration in all six cases, with effect sizes ranging from negligible to medium ($|d| = 0.07$ – 0.79). In summary, **EAC reduces recommendation concentration consistently and substantially.**

RQ2 – Diversity and coverage. ILD increased significantly in all six combinations versus the original baseline ($|d| = 0.44$ – 2.80). Gains over Traditional Calibration are also significant in all cases ($|d| = 0.22$ – 1.09), confirming that the UCB exploration bonus improves diversity beyond what calibration alone achieves. Category Coverage improved in all configurations: on ML100K it rises from ≈ 0.52 – 0.54 (Original) to ≈ 0.72 – 0.74 (EAC), and on LastFM from ≈ 0.30 – 0.53 (Original) to ≈ 0.43 – 0.58 (EAC). Considering both datasets, the overall MAUT score rises from an average of ≈ 0.21 (Original) to ≈ 0.79 (EAC), reflecting broad multi-objective benefit. Thus, **EAC consistently improves diversity and coverage.**

RQ3 – Accuracy. NDCG decreased in five of the six combinations versus the original baseline, with negligible effect in all cases ($|d| = 0.05$ – 0.14); in one of those cases (MF on ML100K), the decrease versus Traditional Calibration is not statistically significant ($\bullet$ symbol in Table 2). The sole exception is LastFM/MostPopular, where NDCG *increased* ($|d| = 0.23$): the Most Popular baseline on LastFM clusters recommendations around a few dominant tags, and calibration realigns lists with individual profiles, incidentally improving relevance. Overall, accuracy costs are modest, with effect sizes uniformly below the negligible threshold. In summary, **EAC preserves competitive accuracy.**

RQ4 – Consistency. Improvements in GBI, CatCov and ILD are significant for all three recommenders (MostPopular, ItemKNN, MF) and on both datasets. The pattern holds regardless of the strength of the base model, from the non-personalized MostPopular to the latent-factor MF. Consequently, **EAC's benefits are consistent in terms of diversity and coverage across algorithms and domains.**

5.3 User Group Analysis

Table 3 breaks down results by user group (niche, balanced, mainstream). EAC improves GBI and ILD for all groups in both datasets, confirming that the method does not exclusively benefit a specific user segment. Regarding CatCov, the metric shows the recommendations across user groups are categorically diverse. On ML100K, the largest absolute CatCov gain across groups appears for MF/Niche (Original: $0.496 \rightarrow$ EAC: 0.741); on LastFM the gains are substantial for ItemKNN across all groups (Original: 0.14 – $0.39 \rightarrow$ EAC: 0.31 – 0.49). Niche users show the strongest absolute ILD gains (e.g., MF/Niche on ML100K: $0.699 \rightarrow 0.838$), as the UCB bonus effectively spreads recommendations across underrepresented categories. Accuracy differences across groups are predominantly non-significant or exhibit negligible effect, indicating that EAC does not disproportionately penalize any user segment. Thus, in response to RQ5, **EAC improves diversity and reduces concentration for all user groups without disproportionate accuracy cost.**

5.4 Profile Concentration and Individual EAC Gain

To investigate whether EAC benefits scale with individual user profile concentration, we compute for each user the maximum category proportion in their historical profile, $\max_c P_u(c)$, as a continuous measure of preference concentration. We then correlate this with the per-user gain in GBI ($\Delta \text{GBI} = \text{GBI}_{\text{EAC}} - \text{GBI}_{\text{Orig}}$) and ILD ($\Delta \text{ILD} = \text{ILD}_{\text{EAC}} - \text{ILD}_{\text{Orig}}$), averaged across the three

Table 3: Results by user group (test set, $K = 10$). Bold = best per (dataset, model, group). MAUT from four displayed metrics, normalized within each (dataset, model, group) block. See note below for symbol legend.

Dataset	Model	Group	Method	NDCG	ILD	CatCov	GBI	MAUT
ML100K	Most Popular	Niche	Original	0.0340	0.6860	0.5439	0.2633	0.2500
			Trad. ($\lambda=0.5$)	0.0296 ↓	0.7507 ↑	0.6787 ↑	0.1868 ↑	0.5485 ↑
			EAC ($\lambda=0.7, \beta=0.20$)	0.0264 **	0.8136 ↑↑	0.7556 ↑↑	0.1419 ↑↑	0.7500 ↑↑
		Balanced	Original	0.0637	0.6912	0.5353	0.2694	0.2500
			Trad. ($\lambda=0.5$)	0.0620 *	0.7620 ↑	0.6965 ↑	0.1781 ↑	0.6848 ↑
			EAC ($\lambda=0.7, \beta=0.20$)	0.0568 * ↓	0.8194 ↑↑	0.7592 ↑↑	0.1394 ↑↑	0.7500 ↑↑
		Mainstream	Original	0.0746	0.6827	0.5437	0.2616	0.2500
			Trad. ($\lambda=0.5$)	0.0538 ↓	0.7323 ↑	0.6321 ↑	0.2085 ↑	0.3988 *
			EAC ($\lambda=0.7, \beta=0.20$)	0.0527 ↓ *	0.8023 ↑↑	0.6996 ↑↑	0.1671 ↑↑	0.7500 ↑↑
	ItemKNN	Niche	Original	0.0590	0.6984	0.5124	0.2904	0.2500
			Trad. ($\lambda=0.5$)	0.0488 ↓	0.7421 ↑	0.6484 ↑	0.2151 ↑	0.4530 ↑
			EAC ($\lambda=0.7, \beta=0.20$)	0.0453 ↓ ↓	0.8208 ↑↑	0.7306 ↑↑	0.1594 ↑↑	0.7500 ↑↑
		Balanced	Original	0.0750	0.7188	0.5348	0.2720	0.2500
			Trad. ($\lambda=0.5$)	0.0677 ↓	0.7614 ↑	0.6667 ↑	0.1988 ↑	0.4863 ↑
			EAC ($\lambda=0.7, \beta=0.20$)	0.0645 ↓ *	0.8268 ↑↑	0.7400 ↑↑	0.1515 ↑↑	0.7500 ↑↑
		Mainstream	Original	0.0966	0.6934	0.5161	0.2857	0.2500
			Trad. ($\lambda=0.5$)	0.0816 ↓	0.7128 ↑	0.5919 ↑	0.2511 ↑	0.3573 *
			EAC ($\lambda=0.7, \beta=0.20$)	0.0721 ↓ *	0.8019 ↑↑	0.6698 ↑↑	0.1915 ↑↑	0.7500 ↑↑
	MF	Niche	Original	0.0667	0.6994	0.4965	0.3164	0.2500
			Trad. ($\lambda=0.7$)	0.0476 ↓	0.7635 ↑	0.6756 ↑	0.2053 ↑	0.5155 ↑
			EAC ($\lambda=0.7, \beta=0.20$)	0.0437 ↓ *	0.8381 ↑↑	0.7414 ↑↑	0.1568 ↑↑	0.7500 ↑↑
		Balanced	Original	0.0682	0.7253	0.5228	0.2913	0.2500
			Trad. ($\lambda=0.7$)	0.0570 ↓	0.7734 ↑	0.6911 ↑	0.1918 ↑	0.4745 ↑
			EAC ($\lambda=0.7, \beta=0.20$)	0.0567 ↓ *	0.8363 ↑↑	0.7510 ↑↑	0.1490 ↑↑	0.7500 ↑↑
		Mainstream	Original	0.0757	0.7001	0.5142	0.3015	0.2500
			Trad. ($\lambda=0.7$)	0.0645 ↓	0.7038 ↑	0.5894 ↑	0.2637 ↑	0.2701 *
			EAC ($\lambda=0.7, \beta=0.20$)	0.0590 ↓ *	0.8310 ↑↑	0.6956 ↑↑	0.1784 ↑↑	0.7500 ↑↑
LastFM	Most Popular	Niche	Original	0.0123	0.5069	0.5249	0.1454	0.0000
			Trad. ($\lambda=0.7$)	0.0329 ↑	0.6013 ↑	0.5743 ↑	0.1291 ↑	0.9335 ↑
			EAC ($\lambda=0.7, \beta=0.20$)	0.0306 ↑ ↓	0.6084 ↑↑	0.5796 ↑↑	0.1273 ↑↑	0.9722 ↑↑
		Balanced	Original	0.0608	0.5224	0.5235	0.1453	0.0000
			Trad. ($\lambda=0.7$)	0.1004 ↑	0.5948 ↑	0.5750 ↑	0.1291 ↑	0.9330 ↑
			EAC ($\lambda=0.7, \beta=0.20$)	0.0994 ↑ ↓	0.6014 ↑↑	0.5802 ↑↑	0.1275 ↑↑	0.9933 ↑↑
		Mainstream	Original	0.1316	0.5472	0.5288	0.1436	0.0000
			Trad. ($\lambda=0.7$)	0.1707 ↑	0.5919 ↑	0.5697 ↑	0.1310 ↑	0.8932 ↑
			EAC ($\lambda=0.7, \beta=0.20$)	0.1684 ↑ ↓	0.5993 ↑↑	0.5765 ↑↑	0.1289 ↑↑	0.9854 ↑↑
	ItemKNN	Niche	Original	0.0617	0.7594	0.1367	0.4242	0.0000
			Trad. ($\lambda=0.7$)	0.0756 ↑	0.7836 ↑	0.3031 ↑	0.2473 ↑	0.8132 ↑
			EAC ($\lambda=0.7, \beta=0.20$)	0.0710 * ↓	0.8340 ↑↑	0.3081 ↑↑	0.2396 ↑↑	0.9172 ↑↑
		Balanced	Original	0.1523	0.6272	0.3375	0.2323	0.2500
			Trad. ($\lambda=0.7$)	0.1473 *	0.6482 ↑	0.4561 ↑	0.1695 ↑	0.6897 ↑
			EAC ($\lambda=0.7, \beta=0.20$)	0.1433 ↓ ↓	0.6791 ↑↑	0.4610 ↑↑	0.1662 ↑↑	0.7500 ↑↑
		Mainstream	Original	0.2478	0.5549	0.3870	0.1986	0.2500
			Trad. ($\lambda=0.7$)	0.2236 ↓	0.5910 ↑	0.4796 ↑	0.1625 ↑	0.6597 *
			EAC ($\lambda=0.7, \beta=0.20$)	0.2196 ↓ ↓	0.6073 ↑↑	0.4889 ↑↑	0.1586 ↑↑	0.7500 * ↓
	MF	Niche	Original	0.0650	0.6705	0.3748	0.2052	0.0280
			Trad. ($\lambda=0.7$)	0.0678 *	0.6956 ↑	0.4603 ↑	0.1669 ↑	0.8449 ↑
			EAC ($\lambda=0.7, \beta=0.20$)	0.0646 * ↓	0.7096 ↑↑	0.4734 ↑↑	0.1611 ↑↑	0.7500 ↑↑
		Balanced	Original	0.1408	0.6671	0.3618	0.2094	0.2500
			Trad. ($\lambda=0.7$)	0.1314 ↓	0.6776 ↑	0.4828 ↑	0.1591 ↑	0.6293 *
			EAC ($\lambda=0.7, \beta=0.20$)	0.1288 ↓ ↓	0.6894 ↑↑	0.4939 ↑↑	0.1546 ↑↑	0.7500 ↑↑
		Mainstream	Original	0.2107	0.6517	0.3577	0.2104	0.2500
			Trad. ($\lambda=0.7$)	0.2051 *	0.6566 ↑	0.4794 ↑	0.1612 ↑	0.5769 *
			EAC ($\lambda=0.7, \beta=0.20$)	0.2036 * ↓	0.6688 ↑↑	0.4928 ↑↑	0.1560 ↑↑	0.7500 * ↑

Note: Any arrow (↑/↓) denotes a statistically significant difference (Wilcoxon signed-rank, $p < 0.05$); * = not significant. Traditional rows: one symbol [vs. Orig.]; EAC rows: two symbols [vs. Orig.][vs. Trad.]. Arrow color encodes Cohen's $|d|$ effect size: black negligible ($|d| < 0.2$), blue small ($0.2 \leq |d| < 0.5$), green medium ($0.5 \leq |d| < 0.8$), red large ($|d| \geq 0.8$).

GBI is lower-is-better (↑ = significant decrease).

recommender models. Figure 2 presents the resulting scatter plots with Spearman correlation coefficients.

For LastFM, both metrics show the expected pattern: users with more concentrated profiles benefit more from EAC, with Spearman $\rho = -0.47$ for ΔGBI ($p < 0.001$) and $\rho = 0.37$ for ΔILD ($p < 0.001$). This confirms that the UCB exploration bonus is particularly effective when historical preferences are skewed, as underrepresented categories receive the highest bonuses.

For ML100K, the correlation for ΔILD is not significant ($\rho = 0.01$, $p = 0.68$), and the ΔGBI correlation reverses sign ($\rho = 0.31$, $p < 0.001$), indicating that highly concentrated users in this dataset benefit *less* from GBI reduction. We attribute this to two factors. First, ML100K has only 19 genres, limiting the diversity achievable in a list of $K=10$ items. Second, with $\lambda^* = 0.7$, the strong calibration term forces recommendation lists to reflect the user's dominant genre, partially counteracting the exploration bonus when the profile is highly skewed. Thus, answering RQ6, **profile concentration predicts EAC gain on LastFM but not on ML100K, where the limited genre space moderates the effect.**

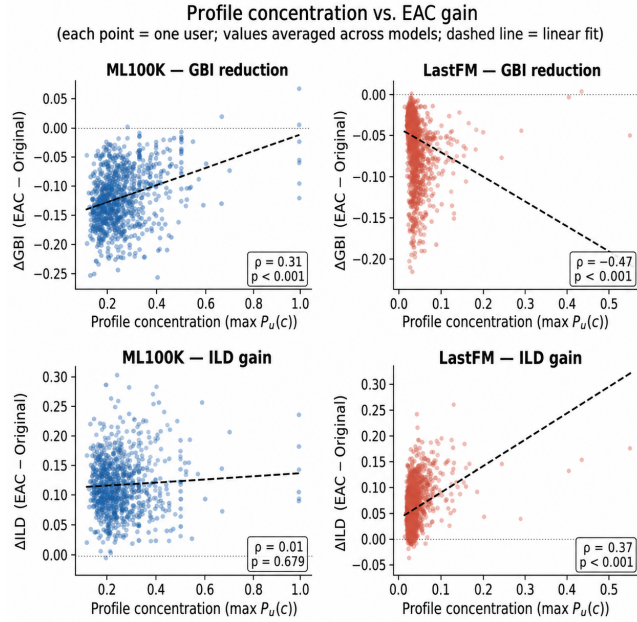


Figure 2: Per-user profile concentration ($\max_c P_u(c)$) versus EAC gain in GBI (top) and ILD (bottom) for ML100K (left) and LastFM (right). Each point represents one user; values are averaged across the three recommender models. Dashed lines show linear fits; Spearman ρ and p -value are reported in each panel.

6 FINAL REMARKS

This paper introduced *Exploration-Aware Calibration* (EAC), a post-processing re-ranking framework that extends traditional calibration with a UCB-inspired exploration bonus. By assigning higher bonuses to categories underrepresented in both the user's historical profile and the recommendation list being constructed, EAC

promotes content discovery without abandoning preference alignment. Experiments across two domains, three recommendation algorithms, and three user groups confirm that EAC consistently reduces recommendation concentration (GBI) and improves intra-list diversity (ILD) and category coverage, while preserving competitive accuracy levels — addressing all six research questions. The multi-objective MAUT score shows improvements for niche, balanced, and mainstream users, demonstrating that the method does not disproportionately benefit any particular user segment.

Three limitations deserve attention. First, the optimal exploration weight $\beta^* = 0.20$ lies at the boundary of the evaluated grid, suggesting that larger values may yield further diversity gains. Second, the correlation between user profile concentration and EAC benefit is dataset-dependent: on ML100K, the limited genre space (19 genres, $K = 10$ items) constrains the achievable diversity, weakening the expected concentration–benefit relationship. Third, the evaluation is entirely offline and therefore cannot determine whether the observed diversity gains translate into improved user satisfaction, engagement, or long-term discovery. Future work should explore broader ranges of β , dynamic per-user exploration rates that adapt to profile sparsity and genre-space size, the integration of EAC into end-to-end neural recommenders, and the long-term effects of exploration-aware re-ranking through user studies or online evaluation.

ETHICAL CONSIDERATIONS

This study uses only publicly available secondary datasets (MovieLens 100K and HetRec 2011 LastFM 2K) and did not involve direct interaction with human participants. Only the information provided in the public datasets was used, with no attempt to identify or re-identify users. Therefore, the study is consistent with Brazilian Resolution CNS No. 510/2016 and did not require evaluation by a Research Ethics Committee.

ACKNOWLEDGMENTS

The authors acknowledge the financial support from FAPESP, CAPES, CNPq, and INCT-TILD-IAR. The authors used AI-assisted tools (Claude/Anthropic and ChatGPT/OpenAI) during the preparation of this work to support code development, data analysis, and manuscript editing. After using these tools, the authors reviewed and edited all content as needed, and take full responsibility for the content of the publication. No AI-generated text was used without revision by the authors.

REFERENCES

- [1] Himan Abdollahpour, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2020. Multi-stakeholder Recommendation: Survey and Research Directions. *User Modeling and User-Adapted Interaction* 30, 1 (2020), 127–158. <https://doi.org/10.1007/s11257-019-09256-1>
- [2] Arda Antikacioglu, Tanvi Bajpai, and R. Ravi. 2019. A new system-wide diversity measure for recommendations with efficient algorithms. *arXiv:1812.03030 [cs.LG]* <https://arxiv.org/abs/1812.03030>
- [3] Qazi Mohammad Areeb, Mohammad Nadeem, Shahab Saqib Sohail, Raza Imam, Faiyaz Doctor, Yassine Himeur, Amir Hussain, and Abbas Amira. 2023. Filter Bubbles in Recommender Systems: Fact or Fallacy – A Systematic Review. *arXiv:2307.01221 [cs.LG]* <https://arxiv.org/abs/2307.01221>
- [4] Paul Atauchi, André Zanon, Leonardo Rocha, and Marcelo Manzato. 2025. Aprimorando a Personalização de Sistemas de Recomendação por Calibração

- Multicategoria. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web* (Rio de Janeiro/RJ). SBC, Porto Alegre, RS, Brasil, 506–510. <https://doi.org/10.5753/webmedia.2025.16145>
- [5] P. Auer, N. Cesa-Bianchi, and P. Fischer. 2002. Finite-Time Analysis of the Multiarmed Bandit Problem. *Machine Learning* 47 (2002), 235–256. <https://doi.org/10.1023/A:1013689704352>
- [6] Ralph L. Keeney and Howard Raiffa. 1976. *Decisions with Multiple Objectives: Preferences and Value Trade-offs*. Wiley. <https://www.cambridge.org/core/books/decisions-with-multiple-objectives/6B9B3CBC5765C7A47B48A9CB4892A89B>
- [7] Y. Koren, R. Bell, and C. Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37. <https://doi.org/10.1109/MC.2009.263>
- [8] T. Lattimore and C. Szepesvári. 2020. *Bandit Algorithms*. Cambridge University Press. <https://tor-lattimore.com/downloads/book/book.pdf>
- [9] L. Li, W. Chu, J. Langford, and R. E. Schapire. 2010. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *Proceedings of the 19th International Conference on World Wide Web*. 661–670. <https://doi.org/10.1145/1772690.1772758>
- [10] Gabriel Machado Lunardi, Guilherme Medeiros Machado, Vinicius Maran, and José Palazzo M. de Oliveira. 2020. A metric for Filter Bubble measurement in recommender algorithms considering the news domain. *Applied Soft Computing* 97 (2020), 106771. <https://doi.org/10.1016/j.asoc.2020.106771>
- [11] T. T. Nguyen, P. Hui, F. M. Harper, L. Terveen, and J. A. Konstan. 2014. Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity. In *Proceedings of the 23rd International Conference on World Wide Web*. 677–686. <https://doi.org/10.1145/2566486.2568012>
- [12] Y. Q. Palomino, R. F. de Souza, P. D. F. Atauchi, and M. G. Manzato. 2026. Calibration of Personality-Based Recommender Systems: Balancing Relevance and Diversity. In *Proceedings of the 41st ACM/SIGAPP Symposium on Applied Computing (SAC '26)*. ACM. <https://doi.org/10.1145/3748522.3779927>
- [13] E. Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press. <https://dl.acm.org/doi/10.5555/2029079>
- [14] Gabriel Prenassi, Rodrigo Souza, Henrique Sekido, Guilherme Fonseca, Marcelo Manzato, and Leonardo Rocha. 2025. Calibração de Sistemas de Recomendação com LLMs: Otimização de Prompts para Balancear Precisão, Diversidade e Justiça. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web* (Rio de Janeiro/RJ). SBC, Porto Alegre, RS, Brasil, 103–111. <https://doi.org/10.5753/webmedia.2025.15613>
- [15] F. Ricci, L. Rokach, and B. Shapira. 2015. *Recommender Systems Handbook* (2nd ed.). Springer. <https://link.springer.com/book/10.1007/978-1-4899-7637-6>
- [16] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. 2001. Item-Based Collaborative Filtering Recommendation Algorithms. In *Proceedings of the 10th International Conference on World Wide Web*. ACM, 285–295. <https://doi.org/10.1145/371920.372071>
- [17] D. C. Silva and F. A. Durão. 2022. Introducing a Framework and a Decision Protocol to Calibrate Recommender Systems. arXiv:2204.03706 [cs.IR] <https://arxiv.org/abs/2204.03706>
- [18] Nicollas Silva, Heitor Werneck, Thiago Silva, Adriano C. M. Pereira, and Leonardo Rocha. 2021. A contextual approach to improve the user's experience in interactive recommendation systems. In *Proceedings of the Brazilian Symposium on Multimedia and the Web* (Belo Horizonte, Minas Gerais, Brazil) (WebMedia '21). Association for Computing Machinery, New York, NY, USA, 89–96. <https://doi.org/10.1145/3470482.3479621>
- [19] Rodrigo Souza and Marcelo Manzato. 2024. Explorando Formas de Calibração e Redução do Viés de Popularidade em Sistemas de Recomendação. In *Anais Estendidos do XXX Simpósio Brasileiro de Sistemas Multimídia e Web* (Juiz de Fora/MG). SBC, Porto Alegre, RS, Brasil, 9–10. https://doi.org/10.5753/webmedia_estendido.2024.244380
- [20] R. F. Souza and M. G. Manzato. 2024. Enhancing Calibration and Reducing Popularity Bias in Recommender Systems. *Lecture Notes in Business Information Processing* 519 (2024), 3–24. https://doi.org/10.1007/978-3-031-64755-0_1
- [21] H. Steck. 2018. Calibrated Recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 154–162. <https://doi.org/10.1145/3240323.3240372>
- [22] Nicholas Sukiennik, Chen Gao, and Nian Li. 2024. Uncovering the Deep Filter Bubble: Narrow Exposure in Short-Video Recommendation. In *Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) (WWW '24). Association for Computing Machinery, New York, NY, USA, 4727–4735. <https://doi.org/10.1145/3589334.3648159>