Designing and Evaluating Calibration-Driven Recommender Systems for Fairness and Explainability

Paul Atauchi[1], Andre Zanon[1], Leonardo Rocha[2], Marcelo Manzato[1]

[1]University of São Paulo (ICMC/USP), Brazil

[2]Universidade Federal de São João del Rei (UFSJ), Brasil

{paul.atauchi, andre.zanon, manzato}@icmc.usp.br; lcrocha@ufsj.edu.br

Abstract. *Recommender systems are increasingly required to satisfy fairness and transparency constraints. However, the interaction between calibration-based fairness interventions and recommendation explainability remains insufficiently understood. This paper proposes a calibration-driven framework that extends single-category calibration to a multi-category setting, enabling a richer representation of user preferences across item dimensions. The approach integrates fairness-aware calibration with post-hoc knowledge-graph-based explanations into a unified pipeline. We evaluate both recommendation performance and explanation quality, considering alignment with user profiles, accuracy, miscalibration, and popularity bias, as well as explanation metrics such as recency, popularity, and diversity. Results show improved alignment, reduced miscalibration and popularity bias, and preserved accuracy, while explanation quality remains stable. These findings indicate that calibration-based fairness can be achieved without compromising explainability.*

1. Introduction

Recommender systems (RSs) help users navigate large item spaces through personalized suggestions and are widely used in e-commerce, streaming, social networks, and online marketplaces [Aggarwal et al. 2016, Roy and Dutta 2022]. Although traditional approaches focus on accuracy metrics, such as rating prediction and ranking relevance [Aggarwal et al. 2016], these alone are insufficient to ensure trustworthy and satisfactory user experiences. Consequently, recent research has expanded beyond accuracy to include diversity, novelty, serendipity, and fairness [Ping et al. 2025, Zhao et al. 2025], as well as explainability and transparency to support responsible recommendations [Ge et al. 2024].

In particular, explainability has become an important part of responsible Artificial Intelligence (AI), aiming to clarify why specific items are recommended and thereby improving transparency, user trust, and engagement [Zhang and Chen 2020]. In many systems, explanations are generated post-hoc from model outputs [Arévalo and Salmeron 2025]. The problem with this approach is that biases and/or distortions introduced at the recommendation stage may be propagated to the explanation phase, potentially affecting its quality, transparency, and trustworthiness [Liu et al. 2024].

To address this issue, fairness in recommender systems seeks to ensure both equitable user outcomes and fair item exposure [Zhao et al. 2025]. Calibration is one approach to achieving user-centered fairness by ensuring model outputs are representative of underlying data or preferences. Steck [Steck 2018] formalized calibration as a fairness-aware post-processing mechanism that aligns the categorical distribution of recommenda-

tions with users' historical preferences while minimizing accuracy loss. However, existing calibration approaches are largely limited to a single-category (e.g., genre), implicitly assuming one-dimensional user preferences. This simplification fails to capture the inherently multifaceted nature of user interests across multiple dimensions.

This work proposes a calibration-driven recommendation framework for fairness and explainability that investigates the interplay between multi-category calibration and post-hoc explanation quality in recommender systems. Specifically, we extend traditional single-category calibration to a multi-category setting, enabling recommendations to better capture the multifaceted user preferences across multiple categorical attributes. The framework also explicitly links calibration to explanation generation, enabling a systematic analysis of how fairness-driven post-processing interventions affect explanation quality. By jointly examining these components, our approach provides new insights into the relationship between fairness mechanisms and explainability, highlighting trade-offs and synergies often overlooked when these aspects are studied in isolation.

2. Related Work

Early recommender system research primarily focused on accuracy-based objectives, such as rating prediction or ranking performance [Aggarwal et al. 2016]. However, this focus can introduce biases by favoring popular items and repeating common data patterns. As a result, recommendations may not fully reflect user preferences, especially by ignoring less frequent or niche interests. To address this issue, Steck [Steck 2018] proposed calibration, a post-processing method that adjusts recommendations to better match the user's past preferences. This involves balancing relevance with how well the recommended items follow the user's preference distribution, often measured using Kullback-Leibler (KL) divergence. Calibration is now seen as a simple and effective way to improve fairness and representation in recommendation lists.

Several works have extended calibration in different directions. Abdollahpouri et al. [Abdollahpouri et al. 2019] showed that popularity bias over-exposes popular items and reduces personalization. Sacilotti et al. [Sacilotti et al. 2023] proposed a method that combines genre and popularity to reduce this bias while preserving accuracy. Da Silva et al. [da Silva et al. 2021] introduced a flexible post-processing framework that supports different metrics and personalized trade-offs between fairness and accuracy. Souza and Manzato [Souza and Manzato 2024b] proposed a two-stage calibration approach that aligns recommendations with genre and popularity preferences. Da Silva and Durão [da Silva and Durão 2023] presented a framework to jointly optimize relevance and fairness using a trade-off formulation, and studies such as [Lima et al. 2022] further investigate fairness through regularization strategies, analyzing their impact on bias mitigation and recommendation quality. Recent advances also explore efficiency-oriented solutions, such as surrogate submodular formulations combined with lazy greedy algorithms for calibrated recommendation [da Silva et al. 2025]. Moreover, Naghiaei et al. [Naghiaei et al. 2024] expanded calibration by incorporating novelty, diversity, and serendipity through a multi-objective re-ranking approach.

Explainability has emerged as a key research direction for improving transparency and user trust in recommender systems. Existing approaches are typically categorized as intrinsic or post-hoc methods [Zhang and Chen 2020]. Among post-hoc approaches, knowledge graph-based methods have gained particular attention. For instance, Musto et

al. [Musto et al. 2016] proposed ExpLOD, which generates natural language explanations by leveraging Linked Open Data (LOD) to identify shared properties between user-liked and recommendation items. These properties are ranked by relevance and used to generate explanations via template-based generation. Extending this work to ExpLOD_v2, Musto et al. [Musto et al. 2019] incorporate both direct and indirect relationships from LOD, enabling more expressive and generalizable explanations across multiple items. Addressing limitations in exploiting hierarchical information, Du et al. [Du et al. 2022] introduced a property-based explanation model (PEM) that leverages the DBpedia [1] category hierarchy to generate semantically rich and non-redundant explanations.

Despite these advances, most calibration methods assume that user preferences can be represented along a single categorical dimension, an assumption that recent work has challenged, highlighting the inherently multifaceted nature of user interests. In this direction, [Atauchi et al. 2025a] proposed a multi-category calibration approach that simultaneously considers multiple attributes derived from a knowledge graph (KG), improving alignment with user preferences while maintaining accuracy. Nevertheless, calibration research remains largely focused on preference alignment, overlooking its broader impact on other system components. Although calibration and explainability are typically applied as post-processing steps, their interaction has received limited attention. Prior work [Atauchi et al. 2025b] showed that single-category calibration has minimal impact on explanation quality, suggesting robustness to re-ranking effects. However, the impact of multi-category calibration on explanation quality remains unexplored, leaving open whether improving calibration across multiple dimensions affects it. More recently, advances in large language models have enabled the integration of LLMs with KG, supporting self-explainable recommendation pipelines that combine structured semantic reasoning with natural language generation [Seabra et al. 2025].

Overall, prior research has made significant progress in enhancing fairness through calibration and transparency through explainability. However, these directions have largely evolved independently. Current calibration methods still fail to capture the multidimensional nature of user preferences, and there is limited understanding of how multi-category calibration affects explanation quality. Addressing this gap, we study the impact of multi-category calibration on post-hoc explanation quality.

3. Proposed Framework

This work proposes a calibration-driven recommender system framework that integrates multi-category calibration and explanation generation within a unified framework. The modular and extensible framework supports different recommendation algorithms and explanation paradigms without altering its core structure, which consists of three sequential modules: (i) recommendation generation, (ii) fairness-centered calibration, and (iii) explanation generation. Each module contributes a distinct analytical layer while maintaining a coherent end-to-end pipeline from raw interaction data to explanation quality assessment. Figure 1 provides an overview of the pipeline.

3.1. Recommender System Module

As shown in Figure 1(a), the framework takes as input a processed user-item interaction dataset, and a recommendation algorithm produces a ranked top-N recommendation list

[1] https://www.dbpedia.org/

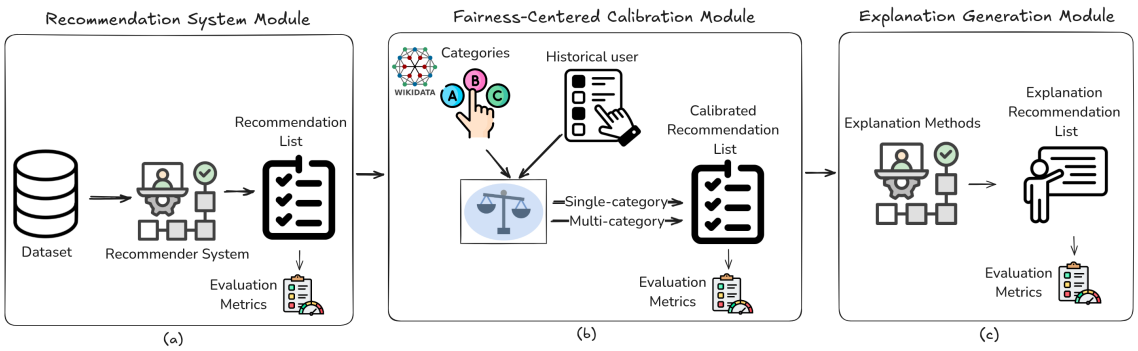

Figure 1. Overview of proposed framework. (a) Recommendation generation. (b) Recommendation calibration. (c) Explanation generation.

I_u for each user u. This module is model-agnostic, requiring only that the model produce a ranked list of items along with relevance scores $s(i, u)$, which are later used during calibration, ensuring that the proposed framework is not tied to a specific recommendation paradigm and can be readily integrated with state-of-the-art or future models.

3.2. Fairness-Centered Calibration Module

Figure 1(b) presents the calibration component. The initial recommendation list is re-ranked to align with the distribution of recommended item categories and the user's historical preference profile. We adopt the calibration objective proposed by Steck[Steck 2018] (Eq. 1) and the multi-category extension previously introduced by Atauchi et al. [Atauchi et al. 2025a] (Eqs 2–3).

3.2.1. Calibration Objective

Calibration is formulated as an optimization problem that balances two competing objectives: recommendation relevance and distributional alignment with user preferences.

Formally, for a user u, the calibrated recommendation list I_u^* is obtained by solving Eq. 1. This equation defines the optimization problem as maximizing recommendation relevance while minimizing the divergence between the category distribution of the recommended list and the user's preference profile. The term $s(I_u)$ quantifies the relevance of the recommendations, whereas $D_{\mathrm{KL}}^{\mathcal{C}}\big(p, q(I_u)\big)$ captures the divergence between the observed category distribution in the generated list and the user's preference distribution p. The parameter $\lambda \in [0, 1]$ acts as a regularization factor, controlling the trade-off between relevance and categorical alignment.

$$I_u^* = \arg\max_{I_u} \left[(1-\lambda)\, s(I_u) - \lambda\, D_{\mathrm{KL}}^{\mathcal{C}}\big(p, q(I_u)\big)\right], \tag{1}$$

3.2.2. Multi-Category Divergence Formulation

This term is defined as a weighted sum of Kullback-Leibler divergences across categories:

$$D_{\mathrm{KL}}^{\mathcal{C}}\big(p, q(I_u)\big) = \sum_{c \in \mathcal{C}} \omega(c)\, D_{\mathrm{KL}}(p^c, q(I_u)^c), \tag{2}$$

C denotes the set of considered item categories, e.g., genre, country of origin, and award received.

3.2.3. Category Weighting

To mitigate over-calibration effects, each category is weighted by Eq. 3, where $\rho(c) \in [0, 1]$ denotes the degree of similarity between the user's historical interaction profile and category c. In our implementation, $\rho(c)$ is computed using cosine similarity. This weighting reduces calibration pressure on categories that are already well represented in the user profile, generalizing conventional single-category calibration to a multi-category setting.

$$\omega(c) = \begin{cases} 1 - \rho(c), & \text{if } |\mathcal{C}| > 1 \quad \text{(multi-category case)}, \\ 1, & \text{if } |\mathcal{C}| = 1 \quad \text{(single-category case)}. \end{cases} \tag{3}$$

3.3. Explanation Generation Module

This component (see Figure 1(c)) generates post-hoc explanations for the re-ranked list using structured knowledge sources represented as knowledge graphs. This component is decoupled from the calibration module, making it fully model-agnostic. Thus, any post-hoc explanation method can be applied to the re-ranked list. This design enables a fair and modular analysis of how calibration affects different explanation paradigms. Although this work focuses on syntactic KG-based methods, the framework is modular and can incorporate alternative paradigms such as counterfactual or LLM-based explanations.

4. Materials and Methods

4.1. Dataset

Experiments were conducted on two widely used benchmark datasets, MoviLens 100K [Harper and Konstan 2015] and LastFM 2K [Cantador et al. 2011]. Both datasets provide user-item interactions and can be enriched with external knowledge from Wikidata. For each dataset, interactions were split into training and test sets using an 90/10 ratio.

4.2. Methodological Approach

This study evaluates the impact of calibration on the quality of recommendations and explanations. We consider two recommendation approaches, namely Bayesian Personalized Ranking Matrix Factorization (BPR-MF) [Rendle et al. 2009] and Neural Collaborative Filtering (NCF) [He et al. 2017], which are non-neural and neural algorithms, respectively. Both methods are implemented using the case-recommender library with default parameter settings to ensure reproducibility.

The recommendation output is post-processed with a calibration step, enabling comparisons between the original and calibrated rankings. Unlike previous works that used only a single category to align user profiles with recommendations in a post-processing step, we extracted a KG from Wikidata [2] to obtain multiple categories for the items in the datasets. To obtain the KG for MovieLens, we use the IMDb ID shared by the dataset and Wikidata to extract multi-category item data. To extract information about artists in the LastFM dataset, we use their names to retrieve their respective details. A KG connects items to attributes through categories, creating triples $(item, category, attribute)$. In the KG extracted for MovieLens, we found 9,535 items,

[2] `https://www.wikidata.org/`

23 categories, and 69,487 attributes across 295,787 triples. For the KG of LastFM, we found 11,646 items, 33 different categories, and 23,001 attributes across 134,197 triples.

In this work, we focus primarily on syntactic explanation approaches. Methods such as ExpLOD [Musto et al. 2016], ExpLOD_v2 [Musto et al. 2019], and PEM [Du et al. 2022] operate on the structural topology of the KG, identifying explanation paths based on graph connectivity patterns and attribute frequencies, providing interpretable and transparent justifications grounded in the graph topology.

To ensure the robustness of the results, all experiments use a 10-fold cross-validation protocol. Statistical significance testing evaluates differences between calibrated and non-calibrated approaches, as well as differences in explanation quality across methods at both the single-category and multi-category levels. To analyze the balance between relevance and calibration, we evaluated three values of trade-off parameter λ : 0.1, 0.5, and 0.9. For the multi-category calibration setting, we jointly considered the genre, award received, and country of origin attributes extracted from the knowledge graph.

4.3. Metrics

The effectiveness of the proposed approach was evaluated by complementary metrics:

1. Relevance. Recommendation accuracy is evaluated using Mean Average Precision (MAP), which measures how well relevant items are ranked for each user. Higher MAP values indicate that relevant items appear earlier in the recommendation lists.

2. Miscalibration. To assess how well recommendations reflect user preference distributions, we use Mean Rank Miscalibration (MRMC) metric [da Silva et al. 2021]. MRMC measures the divergence between the distribution of recommended items and users' historical preferences across the ranked list. Lower values indicate better calibration. In this work, calibration is analyzed from two perspectives: item categories and item popularity. These are combined into a single score, $F1_{MRMC}$, using the harmonic mean, where higher values indicate better overall calibration.

3. Diversity. It is measured using Intra-list Diversity (ILD), which captures how different the items in a recommendation list are from each other. Higher ILD values indicate more diverse recommendations, reducing redundancy and encouraging exploration.

4. Popularity Bias. We use the Group Average Popularity (GAP) metric [Abdollahpouri et al. 2021] to measure the average popularity of recommendations across user groups. Users are divided into Blockbuster, Diverse, and Niche groups [Souza and Manzato 2024a, Abdollahpouri et al. 2019]. We compare the popularity of recommended items against users' historical preferences and compute the relative change, summarizing the overall deviation using RMSE_POP, where lower values indicate better alignment with users' original popularity preferences [Palomino et al. 2026].

5. Explanation Quality. It is evaluated using three metrics proposed by Balloccu et al. [Balloccu et al. 2022]. Linked Item Recency (LIR) measures whether explanations rely on recent user interactions. Shared Entity Popularity (SEP) evaluates whether explanations involve popular or niche entities. Explanation Type Diversity (ETD) measures the variety of explanation types provided across the recommendation list.

6. Multi-Attribute Utility Theory (MAUT). It aggregates multiple evaluation metrics into a single score by assigning each a weight reflecting its importance. Metrics are

normalized beforehand for comparability, and final utility is computed as a weighted sum.

5. Experimental Results

We divide our results analysis into two parts: calibration and explanation. The calibration results are reported in Tables 1, 2, 3, where the first column indicates the calibrated category, followed by results for the baseline and three trade-off groups (0.1, 0.5, and 0.9), with metrics including MAP, RMSE_POP, and ILD computed per category, as well as the F1 score for each category and the overall MAUT score.

Table notation. For all tables, ↑ (↓) denotes that higher (lower) values are preferred. Symbols ●, ▲, and ▼denote no significant difference, significantly worse, and significantly better results, respectively, relative to multi-category calibration.

Table 1. Calibration Results for `LastFM` — `BPR-MF`

Category	↑MAP	↓RMSE_POP	↑ILD_G	↑ILD_CO	↑ILD_AR	↑F1_G	↑F1_CO	↑F1_AR	↑MAUT
baseline	0.201	1.079	0.906	0.570	0.774	0.392	0.649	0.484	0.137
genre (0.1)	0.207●	1.079▲	0.906●	0.570▲	0.775▼	0.404▼	0.653▲	0.484▲	0.166▲
country of origin (0.1)	0.204▲	1.079▲	0.907●	0.581▼	0.775▼	0.395▲	0.693▼	0.484▲	0.223▼
award received (0.1)	0.203▲	1.069▼	0.906▲	0.569▲	0.768▲	0.394▲	0.649▲	0.521▼	0.173▲
multicategory (0.1)	0.206	1.074	0.906	0.574	0.773	0.400	0.667	0.503	0.204
genre (0.5)	0.214●	1.059▲	0.909●	0.571▲	0.782▼	0.454▼	0.671▲	0.480▲	0.314▲
country of origin (0.5)	0.199▲	1.069▲	0.909▲	0.607▼	0.777▼	0.397▲	0.751▼	0.484▲	0.390▲
award received (0.5)	0.196▲	0.996▼	0.906▲	0.569▲	0.703▲	0.396▲	0.650▲	0.606▼	0.299▲
multicategory (0.5)	0.216	1.017	0.909	0.589	0.749	0.433	0.715	0.572	0.473
genre (0.9)	0.187▲	0.959▲	0.911▲	0.576▲	0.778▼	0.477▼	0.683▲	0.477▲	0.499▲
country of origin (0.9)	0.195●	1.031▲	0.910▲	0.595▲	0.776▼	0.397▲	0.761▼	0.485▲	0.432▲
award received (0.9)	0.176▲	0.881▼	0.910▲	0.581▲	0.571▲	0.389▲	0.654▲	0.630▼	0.476▲
multicategory (0.9)	0.194	0.902	0.913	0.604	0.670	0.455	0.734	0.605	0.727

Table 2. Calibration Results for `LastFM` — `NCF`

Category	↑MAP	↓RMSE_POP	↑ILD_G	↑ILD_CO	↑ILD_AR	↑F1_G	↑F1_CO	↑F1_AR	↑MAUT
baseline	0.269	0.389	0.959	0.729	0.368	0.324	0.669	0.599	0.659
genre (0.1)	0.270●	0.389●	0.959▼	0.729▲	0.369●	0.325▼	0.669▲	0.599▲	0.663▲
country of origin (0.1)	0.270●	0.389●	0.959▲	0.730▼	0.368●	0.324▲	0.675▼	0.599▲	0.666▼
award received (0.1)	0.270●	0.389●	0.959▲	0.729●	0.368▲	0.324▲	0.669▲	0.601▼	0.663●
multicategory (0.1)	0.270	0.389	0.959	0.729	0.369	0.325	0.670	0.599	0.664
genre (0.5)	0.276●	0.389●	0.960▼	0.728▲	0.371▼	0.337▼	0.672▲	0.598▲	0.682▲
country of origin (0.5)	0.273▲	0.389●	0.959▲	0.733▼	0.368▲	0.326▲	0.710▼	0.599▲	0.708▼
award received (0.5)	0.271▲	0.390▲	0.959▲	0.729▲	0.364▲	0.324▲	0.668▲	0.612▼	0.681▲
multicategory (0.5)	0.275	0.389	0.960	0.730	0.369	0.332	0.683	0.601	0.693
genre (0.9)	0.282▲	0.390●	0.962▼	0.724▲	0.378▼	0.388▼	0.684▲	0.595▲	0.758▲
country of origin (0.9)	0.261▲	0.390●	0.959▲	0.725▲	0.368▲	0.325▲	0.763▼	0.601▲	0.749▲
award received (0.9)	0.264▲	0.395▲	0.959▲	0.731●	0.340▲	0.323▲	0.667▲	0.635▼	0.682▲
multicategory (0.9)	0.290	0.390	0.961	0.731	0.372	0.372	0.728	0.611	0.815

Table 3. Calibration Results for `MovieLens` — `BPR-MF`

Category	↑MAP	↓RMSE_POP	↑ILD_G	↑ILD_CO	↑ILD_AR	↑F1_G	↑F1_CO	↑F1_AR	↑MAUT
baseline	0.286	0.571	0.883	0.233	0.861	0.505	0.837	0.452	0.161
genre (0.1)	0.290▼	0.571▲	0.884▼	0.233●	0.861●	0.515▼	0.837●	0.452▲	0.177●
country of origin (0.1)	0.288▼	0.569●	0.883▲	0.238▼	0.861●	0.505▲	0.841▼	0.452▲	0.191▼
award received (0.1)	0.285●	0.571●	0.883▲	0.233●	0.861▼	0.506▲	0.837●	0.469▼	0.179●
multicategory (0.1)	0.286	0.570	0.883	0.233	0.861	0.508	0.837	0.460	0.176
genre (0.5)	0.297▼	0.572▲	0.889▼	0.229▲	0.863▼	0.562▼	0.836▲	0.452▲	0.249▲
country of origin (0.5)	0.287●	0.562●	0.884▲	0.259▼	0.860●	0.503▲	0.860▼	0.455▲	0.327▼
award received (0.5)	0.277▲	0.559▼	0.883▲	0.233●	0.856▲	0.510▲	0.838▲	0.534▼	0.268▲
multicategory (0.5)	0.290	0.563	0.885	0.234	0.859	0.528	0.839	0.508	0.277
genre (0.9)	0.264▲	0.541▲	0.896▼	0.230▲	0.863▼	0.601▼	0.833▲	0.451▲	0.390▲
country of origin (0.9)	0.269▲	0.551▲	0.887▲	0.256▼	0.860▼	0.500▲	0.870▼	0.459▲	0.390▲
award received (0.9)	0.258▲	0.479▼	0.885▲	0.231●	0.800▲	0.507▲	0.840▲	0.565▼	0.427▲
multicategory (0.9)	0.275	0.511	0.891	0.235	0.830	0.567	0.845	0.558	0.504

Table 4. Calibration Results for `MovieLens` — `NCF`

Category	↑MAP	↓RMSE_POP	↑ILD_G	↑ILD_CO	↑ILD_AR	↑F1_G	↑F1_CO	↑F1_AR	↑MAUT
baseline	0.358	0.246	0.907	0.361	0.603	0.502	0.843	0.516	0.593
genre (0.1)	0.359●	0.245●	0.907▼	0.361●	0.603▲	0.504▼	0.843●	0.516▲	0.598●
country of origin (0.1)	0.358●	0.246●	0.907●	0.362▼	0.603▲	0.502▲	0.844▼	0.516▲	0.600▼
award received (0.1)	0.358●	0.245●	0.907●	0.361●	0.604▼	0.502▲	0.843●	0.518▼	0.599●
multicategory (0.1)	0.358	0.245	0.907	0.361	0.603	0.503	0.843	0.517	0.597
genre (0.5)	0.368▼	0.245●	0.908▼	0.360●	0.603▲	0.517▼	0.843▲	0.517▲	0.624▼
country of origin (0.5)	0.357▲	0.246●	0.907▲	0.364▼	0.603▲	0.502▲	0.852▼	0.517▲	0.627▼
award received (0.5)	0.354▲	0.245●	0.907▲	0.359▲	0.609▼	0.502▲	0.843●	0.532▼	0.628▼
multicategory (0.5)	0.362	0.245	0.907	0.361	0.605	0.507	0.844	0.520	0.616
genre (0.9)	0.335▲	0.250▲	0.912▼	0.357●	0.604▲	0.568▼	0.843▲	0.517▲	0.701▲
country of origin (0.9)	0.343▲	0.248▲	0.907▲	0.368▼	0.600▲	0.500▲	0.869▼	0.516▲	0.665▲
award received (0.9)	0.307▲	0.246●	0.906▲	0.352▲	0.627▼	0.498▲	0.844▲	0.558▼	0.667▲
multicategory (0.9)	0.354	0.245	0.909	0.357	0.615	0.531	0.848	0.540	0.720

The calibration results show that multi-category calibration maintains comparable MAP performance across all trade-off settings, relative to the baseline and single-category approaches. Regarding the RMSE_POP metric, we observe comparable values or reduced popularity bias as the trade-off increases, relative to both baseline and single-category approaches. Across both datasets and models, ILD results show a generally consistent pattern for multi-category calibration. At a high trade-off ($\lambda = 0.9$), it yields the most balanced diversity across the genre (G), country of origin (CO), and award received (AR) dimensions. In BPR-MF, calibration strongly increases ILD for the genre and country of origin dimensions, but often degrades diversity in the award received dimension, revealing pronounced cross-category competition. Multi-category calibration effectively mitigates this collapse by redistributing diversity more evenly. Negative significance in ILD_X indicates that single-category calibration on X achieves higher diversity than multi-category calibration, reflecting a specialization advantage rather than true degradation. In all cases,

calibrating directly on a category maximizes its ILD, whereas multi-category calibration distributes optimization across categories and thus reduces the peak value in each individual dimension. This can make multi-category calibration appear significantly worse than single-category results, even when absolute ILD remains similar to or better than the baseline. The exception is the award received dimension, which behaves less consistently: it shows weak interactions with other categories and achieves reliable improvements mainly when directly optimized. For the F1 score, the pattern is highly consistent and well defined. For each category X, single-category calibration on X yields the highest F1_X, forming a clear diagonal in which each category only outperforms the others in its own metric. Negative significance appears only in this diagonal, meaning multi-category performs worse than the corresponding single-category baseline for that category. In contrast, multi-category shows significant improvement in off-diagonal cases, boosting F1 scores for the other categories. This indicates a strict specialization behavior, in which each category is best optimized independently, while multi-category calibration promotes cross-category improvements rather than maximizing a single dimension.

The explanation results are reported in Tables 5 and 6, on which the algorithm under analysis is in the first column, followed by results for the baseline and the same trade-off groups, and report LIR, ETD, and SEP, together with the overall MAUT score. The results show generally stable behavior across datasets, models, and explanation algorithms. Overall, LIR remains stable across all methods, indicating that the recency of explanatory properties is preserved regardless of the calibration strategy. For ETD, which measures diversity, multi-category calibration generally yields values comparable to or higher than those of single-category approaches, especially at higher trade-offs, demonstrating its ability to promote diverse explanations without relying on a single attribute. However, this pattern has some notable exceptions. In particular, when calibrating on country of origin in MovieLens with BPR-MF using ExpLOD, and on genre in MovieLens with BPR-MF using ExpLOD_v2, single-category calibration outperforms multi-category calibration. A similar effect is observed in LastFM with BPR-MF across all explanation algorithms (ExpLOD, ExpLOD_v2, and PEM) when calibrating on the country of origin, where multi-category shows negative significance. These cases indicate that, for certain dataset–model–algorithm combinations, single-category calibration can better maximize explanation diversity for specific attributes. This reflects localized specialization effects, where multi-category balancing slightly reduces peak ETD performance. Regarding SEP, where lower values are preferred, multi-category calibration maintains competitive performance, often avoiding the increases in popularity bias observed in some single-category configurations. When compared directly, single-category calibration tends to optimize specific aspects of explanation quality, but these improvements are localized and not consistently maintained across other metrics. In contrast, multi-category calibration provides more balanced improvements, avoiding strong trade-offs and achieving robust performance across LIR, ETD, and SEP simultaneously. This advantage is particularly evident in MAUT, where multi-category calibration consistently yields the highest or near-highest values, especially at higher trade-off levels, demonstrating that jointly optimizing multiple categories leads to superior overall explanation quality. Overall, the results indicate that multi-category calibration is a more effective and stable approach, delivering well-balanced explanations while maintaining or improving upon the performance of both baseline and single-category methods.

Table 5. Explanation Results for `LastFM` using `BPR-MF` and `NCF`.

	`BPR-MF`				`NCF`			
Explanation algorithm / (category)	↑LIR	↑ETD	↓SEP	↑MAUT	↑LIR	↑ETD	↓SEP	↑MAUT
ExpLOD								
baseline	0.0179	0.6084	0.6127	0.4056	0.0170	0.7766	0.6157	0.2746
genre (0.1)	0.0177●	0.6084▲	0.6265●	0.3239●	0.0170●	0.7771●	0.6336●	0.2406●
country of origin (0.1)	0.0180▼	0.6169▼	0.6205●	0.4547▼	0.0170●	0.7770●	0.6127●	0.3451●
award received (0.1)	0.0177●	0.6078▲	0.6228▲	0.3158▲	0.0170●	0.7766▲	0.6245●	0.2768●
multicategory (0.1)	0.0178	0.6107	0.6167	0.3806	0.0170	0.7771	0.6171	0.3145
genre (0.5)	0.0172●	0.6056▲	0.5998●	0.3117●	0.0173●	0.7797●	0.6434●	0.3491●
country of origin (0.5)	0.0181▼	0.6372▼	0.6241●	0.5810▼	0.0172●	0.7796●	0.6143●	0.4087●
award received (0.5)	0.0177●	0.6077▲	0.6193●	0.3309●	0.0170●	0.7767▲	0.6064●	0.3451●
multicategory (0.5)	0.0174	0.6178	0.6109	0.3875	0.0171	0.7792	0.6227	0.3482
genre (0.9)	0.0171●	0.5978▲	0.5924●	0.2634▲	0.0177●	0.7871●	0.6212●	0.7407●
country of origin (0.9)	0.0182▼	0.6429▼	0.6272●	0.6452▼	0.0176●	0.7837●	0.6318●	0.5595●
award received (0.9)	0.0179●	0.6139●	0.6051●	0.4564●	0.0173●	0.7769▲	0.6191●	0.3855▲
multicategory (0.9)	0.0174	0.6174	0.6105	0.4084	0.0175	0.7869	0.6294	0.6391
ExpLOD_v2								
baseline	0.0181	0.6075	0.6741	0.3499	0.0161	0.7869	0.6085	0.2567
genre (0.1)	0.0180●	0.6081▲	0.6855▲	0.2789▲	0.0161●	0.7875●	0.6065●	0.2550●
country of origin (0.1)	0.0183●	0.6178▼	0.6339●	0.4580●	0.0161●	0.7875●	0.6185●	0.2582●
award received (0.1)	0.0181●	0.6081▲	0.6267●	0.3671●	0.0161●	0.7869▲	0.5787●	0.2812●
multicategory (0.1)	0.0181	0.6113	0.6703	0.3715	0.0160	0.7874	0.5745	0.3229
genre (0.5)	0.0179●	0.6051▲	0.6734●	0.3016●	0.0162●	0.7901●	0.6013●	0.3212●
country of origin (0.5)	0.0184●	0.6420▼	0.6701●	0.5518▼	0.0162●	0.7919▼	0.6215●	0.3430●
award received (0.5)	0.0181●	0.6130▲	0.5711●	0.4709●	0.0159▲	0.7868▲	0.6373●	0.1297▲
multicategory (0.5)	0.0178	0.6220	0.6732	0.3707	0.0161	0.7904	0.6269	0.2884
genre (0.9)	0.0170●	0.5953▲	0.6756●	0.1226●	0.0165●	0.8024▲	0.5980●	0.6495●
country of origin (0.9)	0.0185▼	0.6555▼	0.6657●	0.6545▼	0.0169●	0.7998▲	0.5828●	0.6773●
award received (0.9)	0.0178●	0.6185▼	0.6158●	0.3971●	0.0160▲	0.7875▲	0.5907●	0.2536▲
multicategory (0.9)	0.0171	0.6106	0.6008	0.2426	0.0167	0.8038	0.5830	0.7428
PEM								
baseline	0.0142	0.9397	0.1767	0.3727	0.0140	0.9579	0.2752	0.3091
genre (0.1)	0.0141●	0.9409●	0.1700●	0.4177●	0.0141▼	0.9581●	0.2810●	0.3147●
country of origin (0.1)	0.0142●	0.9413●	0.1734▲	0.4279●	0.0140●	0.9582●	0.2765●	0.2933●
award received (0.1)	0.0140●	0.9401▲	0.1744●	0.3745▲	0.0140●	0.9578▲	0.2782●	0.2965●
multicategory (0.1)	0.0142	0.9412	0.1677	0.4443	0.0139	0.9581	0.2793	0.2852
genre (0.5)	0.0137●	0.9444●	0.1513▼	0.5201●	0.0141●	0.9595●	0.2660●	0.4426●
country of origin (0.5)	0.0141●	0.9458▼	0.1713●	0.5149●	0.0139●	0.9593●	0.2783●	0.3042●
award received (0.5)	0.0139●	0.9383▲	0.1859▲	0.2688▲	0.0140●	0.9579▲	0.2614●	0.3788●
multicategory (0.5)	0.0137	0.9435	0.1681	0.4419	0.0141	0.9592	0.2733	0.3746
genre (0.9)	0.0138●	0.9512▼	0.1221▼	0.7904▼	0.0146●	0.9645●	0.2599●	0.7892●
country of origin (0.9)	0.0143●	0.9467▼	0.1737▲	0.5654●	0.0142●	0.9637●	0.2662●	0.6090●
award received (0.9)	0.0140●	0.9360▲	0.2015▲	0.1693▲	0.0142●	0.9566▲	0.2650●	0.3586▲
multicategory (0.9)	0.0143	0.9444	0.1612	0.5682	0.0144	0.9640	0.2576	0.7276

Table 6. Explanation Results for `MovieLens` using `BPR-MF` and `NCF`.

	`BPR-MF`				`NCF`			
Explanation algorithm / (category)	↑LIR	↑ETD	↓SEP	↑MAUT	↑LIR	↑ETD	↓SEP	↑MAUT
ExpLOD								
baseline	0.0857	0.6907	0.5936	0.2440	0.1032	0.8395	0.5382	0.4304
genre (0.1)	0.0853●	0.6918●	0.5929●	0.2076●	0.1031●	0.8398▼	0.5336●	0.4465●
country of origin (0.1)	0.0849▲	0.6923▼	0.5697●	0.2167●	0.1033●	0.8394●	0.5259●	0.4621●
award received (0.1)	0.0854●	0.6898▲	0.5960●	0.2080●	0.1031●	0.8394●	0.5567●	0.3683●
multicategory (0.1)	0.0853	0.6908	0.6013	0.1964	0.1031	0.8395	0.5351	0.4309
genre (0.5)	0.0861●	0.6946▼	0.5759▲	0.2755●	0.1036●	0.8414●	0.5386●	0.5042●
country of origin (0.5)	0.0858●	0.7020▼	0.5301●	0.4497●	0.1035●	0.8397●	0.5534●	0.4143●
award received (0.5)	0.0862●	0.6805▲	0.5017●	0.3336●	0.1025▲	0.8397●	0.5216●	0.4126●
multicategory (0.5)	0.0865	0.6897	0.5609	0.3387	0.1032	0.8401	0.5444	0.4165
genre (0.9)	0.0886●	0.7140▼	0.5326●	0.6469▼	0.1047▼	0.8396●	0.5356●	0.5312●
country of origin (0.9)	0.0871●	0.7169▼	0.4531●	0.7145▼	0.1032●	0.8429●	0.5357●	0.5331●
award received (0.9)	0.0870●	0.6820●	0.5291●	0.3797●	0.1005▲	0.8400●	0.4874●	0.3294●
multicategory (0.9)	0.0877	0.6864	0.5483	0.4416	0.1033	0.8420	0.5343	0.5379
ExpLOD_v2								
baseline	0.0789	0.6787	0.5258	0.2257	0.0914	0.8096	0.4907	0.3132
genre (0.1)	0.0784●	0.6835▼	0.5300●	0.1827●	0.0913●	0.8102●	0.4958●	0.2955●
country of origin (0.1)	0.0784●	0.6813●	0.5268●	0.1941●	0.0914●	0.8097●	0.4692●	0.3823●
award received (0.1)	0.0787●	0.6803●	0.5363●	0.1689●	0.0912●	0.8099●	0.4962●	0.2915●
multicategory (0.1)	0.0787	0.6805	0.5239	0.1968	0.0913	0.8098	0.4824	0.3441
genre (0.5)	0.0813●	0.7001▼	0.5032●	0.4282▼	0.0918●	0.8150▼	0.4927●	0.4448●
country of origin (0.5)	0.0799●	0.6883●	0.5217●	0.2994●	0.0916●	0.8109●	0.4726●	0.4198●
award received (0.5)	0.0805●	0.6752▲	0.4985●	0.3070●	0.0914●	0.8103●	0.4921●	0.3026●
multicategory (0.5)	0.0800	0.6877	0.5216	0.2823	0.0914	0.8115	0.4738	0.3961
genre (0.9)	0.0851●	0.7370▼	0.4800●	0.8626▼	0.0925●	0.8239▼	0.4783●	0.6767●
country of origin (0.9)	0.0818●	0.7005●	0.5155●	0.4797●	0.0921●	0.8155●	0.4976●	0.4634●
award received (0.9)	0.0825●	0.6767▲	0.5062●	0.3553▲	0.0903▲	0.8135▲	0.4929●	0.3008▲
multicategory (0.9)	0.0833	0.7037	0.5136	0.5479	0.0925	0.8183	0.4777	0.5884
PEM								
baseline	0.0293	0.9542	0.1483	0.4092	0.0332	0.9871	0.1683	0.3778
genre (0.1)	0.0288●	0.9542●	0.1481●	0.3856●	0.0335●	0.9871●	0.1600●	0.4759●
country of origin (0.1)	0.0288●	0.9540●	0.1477●	0.3867●	0.0335●	0.9872●	0.1545▼	0.5309▼
award received (0.1)	0.0286●	0.9540●	0.1490●	0.3417▲	0.0334●	0.9872●	0.1635●	0.4433●
multicategory (0.1)	0.0288	0.9540	0.1453	0.4217	0.0334	0.9871	0.1637	0.4316
genre (0.5)	0.0282●	0.9575●	0.1493●	0.3908●	0.0339●	0.9874●	0.1645●	0.5073●
country of origin (0.5)	0.0288●	0.9557●	0.1465●	0.4186●	0.0332●	0.9873●	0.1613●	0.4681●
award received (0.5)	0.0292●	0.9552●	0.1543●	0.3548●	0.0339●	0.9869●	0.1658●	0.4635●
multicategory (0.5)	0.0288	0.9561	0.1488	0.4199	0.0335	0.9871	0.1590	0.4928
genre (0.9)	0.0280●	0.9661▼	0.1504●	0.5562●	0.0333●	0.9889▼	0.1611●	0.6277●
country of origin (0.9)	0.0286●	0.9610●	0.1419▼	0.5960●	0.0327▲	0.9876●	0.1610●	0.4465●
award received (0.9)	0.0294●	0.9592▲	0.1503●	0.4792●	0.0346●	0.9855▲	0.1636●	0.4290●
multicategory (0.9)	0.0282	0.9622	0.1537	0.4480	0.0343	0.9873	0.1598	0.5920

5.1. Discussion

The results highlight a trade-off between specialization and balance in calibration strategies. Single-category calibration consistently maximizes performance on the targeted dimension, as evidenced by the clear diagonal pattern in F1 and the peak ILD values within each category. However, these gains are localized and often come at the expense of other dimensions, revealing a strong cross-category competition. In contrast, multi-category calibration promotes a more balanced optimization across categories. While it does not achieve the highest performance in any single dimension, it mitigates degradation in non-target categories and distributes diversity more evenly. This behavior is particularly evident at higher trade-off values, where multi-category calibration produces the most balanced ILD across dimensions and consistently achieves strong MAUT performance. A similar pattern is observed in the explanation results. Single-category calibration can outperform multi-category approaches in specific cases, particularly when optimizing explanation diversity (ETD) for a given attribute. However, these improvements are not consistently maintained across other metrics. Multi-category calibration, on the other hand, provides stable performance across LIR, ETD, and SEP, avoids strong trade-offs, and yields more robust overall explanation quality.

These findings suggest that multi-category calibration is better suited to real-world recommendation scenarios requiring simultaneous consideration of fairness, diversity, and interpretability. In contrast, single-category calibration may be preferable in applications that prioritize a specific attribute and can tolerate imbalances in others. Overall, the results show that jointly optimizing multiple categories yields more stable, well-balanced systems, even at the expense of peak performance in individual dimensions.

6. Conclusion and Future Work

This paper presents a calibration-driven recommender system framework that jointly analyzes fairness and explainability through multi-category calibration. By extending traditional single-category calibration to multiple dimensions, the proposed method better captures the multifaceted nature of user preferences. Experimental results show that multi-category calibration improves alignment with user profiles, reduces miscalibration and popularity bias, and maintains competitive accuracy when compared to baseline and single-category approaches.

Importantly, our findings show that fairness-oriented re-ranking does not harm explanation quality. Across datasets, recommendation algorithms, and explanation methods, metrics such as LIR, ETD, and SEP remain stable or improve under calibration. While single-category calibration optimizes individual dimensions, multi-category calibration achieves more balanced and robust performance across recommendation and explanation metrics, making it well suited to real-world multi-objective scenarios.

For future work, we plan to explore additional explanation paradigms beyond syntactic knowledge-graph-based methods, investigate new attribute dimensions to better capture real user preferences, and evaluate the robustness of calibration under knowledge graph incompleteness and attribute noise. Finally, user-centric evaluations will be essential for better understanding how calibration affects perceived fairness, trust, explanation usefulness, and overall user experience in practical applications.

Acknowledgments

Supported by CAPES, CNPq, Fapesp, AWS, Fapemig and INCT-TILD-IAR.

References

Abdollahpouri, H., Mansoury, M., Burke, R., and Mobasher, B. (2019). The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286*.

Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B., and Malthouse, E. (2021). User-centered evaluation of popularity bias in recommender systems. In *Proc. of the 29th ACM conference on user modeling, adaptation and personalization*.

Aggarwal, C. C. et al. (2016). *Recommender systems*, volume 1. Springer.

Arévalo, I. and Salmeron, J. L. (2025). Model-agnostic post-hoc explainability for recommender systems. *Expert Systems with Applications*, page 129608.

Atauchi, P., Zanon, A., Rocha, L., and Manzato, M. (2025a). Aprimorando a personalização de sistemas de recomendação por calibração multicategoria. In *Proceedings of the 31st WebMedia*.

Atauchi, P. D. F., Zanon, A. L., Rocha, L. C. D. d., and Manzato, M. G. (2025b). Do calibrated recommendations affect explanations? a study on post-hoc adjustments. *Journal on Interactive Systems*, 16(1):441–460.

Balloccu, G., Boratto, L., Fenu, G., and Marras, M. (2022). Post processing recommender systems with knowledge graphs for recency, popularity, and diversity of explanations. In *Proceedings of the 45th International ACM SIGIR*, SIGIR '22.

Cantador, I., Brusilovsky, P., and Kuflik, T. (2011). 2nd workshop on information heterogeneity and fusion in recommender systems (hetrec 2011). In *Proceedings of the 5th ACM conference on Recommender systems*, RecSys 2011.

da Silva, D. C. and Durão, F. A. (2023). Introducing a framework and a decision protocol to calibrated recommender systems: Dc da silva, fa durão. *Applied Intelligence*, 53(19):22044–22072.

da Silva, D. C., Manzato, M. G., and Durão, F. A. (2021). Exploiting personalized calibration and metrics for fairness recommendation. *Expert Systems with Applications*, 181:115112.

da Silva, D. C., Pires, J. M., and Durão, F. A. (2025). Exploiting surrogate submodular and cost-effective lazy forward algorithms for calibrated recommendations. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 98–111. SBC.

Du, Y., Ranwez, S., Sutton-Charani, N., and Ranwez, V. (2022). Post-hoc recommendation explanations through an efficient exploitation of the dbpedia category hierarchy. *Knowledge-Based Systems*, 245:108560.

Ge, Y., Liu, S., Fu, Z., Tan, J., Li, Z., Xu, S., Li, Y., Xian, Y., and Zhang, Y. (2024). A survey on trustworthy recommender systems. *ACM TORS*, 3(2):1–68.

Harper, F. M. and Konstan, J. A. (2015). The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19.

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., and Chua, T.-S. (2017). Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*.

Lima, R. O., Comarella, G., Belém, F., and Reis, J. C. (2022). Justiça em sistemas de recomendação: Uma análise de técnicas de regularização. In *SBBD*.

Liu, X., Yu, T., Xie, K., Wu, J., and Li, S. (2024). Interact with the explanations: Causal debiased explainable recommendation system. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 472–481.

Musto, C., Narducci, F., Lops, P., De Gemmis, M., and Semeraro, G. (2016). Explod: A framework for explaining recommendations based on the linked open data cloud. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 151–154.

Musto, C., Narducci, F., Lops, P., De Gemmis, M., and Semeraro, G. (2019). Linked open data-based explanations for transparent recommender systems. *International Journal of Human-Computer Studies*, 121:93–107.

Naghiaei, M., Dehghan, M., Rahmani, H. A., Azizi, J., and Aliannejadi, M. (2024). Personalized beyond-accuracy calibration in recommendation. In *Proceedings of the 2024 ACM SIGIR*.

Palomino, Y. Q., Souza, R. F., Atauchi, P. D. F., and Manzato, M. G. (2026). Calibration of personality-based recommender systems: Balancing relevance and diversity. In *Proceedings of the 41st ACM/SIGAPP Symposium on Applied Computing (SAC '26)*.

Ping, Y., Li, Y., and Zhu, J. (2025). Beyond accuracy measures: the effect of diversity, novelty and serendipity in recommender systems on user engagement. *Electronic Commerce Research*, 25(3):2177–2204.

Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. (2009). Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09.

Roy, D. and Dutta, M. (2022). A systematic review and research perspective on recommender systems. *Journal of Big Data*, 9(1):59.

Sacilotti, A., Souza, R. F. d., and Manzato, M. G. (2023). Counteracting popularity-bias and improving diversity through calibrated recommendations. In *Proceedings*.

Seabra, A., Cavalcante, C., and Lifschitz, S. (2025). Towards a self-explainable recommender system using\\large language models and knowledge graphs. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 479–489. SBC.

Souza, R. and Manzato, M. (2024a). 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*, pages 9–10, Porto Alegre, RS, Brasil. SBC.

Souza, R. and Manzato, M. (2024b). A two-stage calibration approach for mitigating bias and fairness in recommender systems. In *Proceedings of the 39th ACM SAC*.

Steck, H. (2018). Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems*, pages 154–162.

Zhang, Y. and Chen, X. (2020). Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, 14(1):1–101.

Zhao, Y., Wang, Y., Liu, Y., Cheng, X., Aggarwal, C. C., and Derr, T. (2025). Fairness and diversity in recommender systems: a survey. *ACM Transactions on Intelligent Systems and Technology*, 16(1):1–28.