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

Paul Atauchi[1]

[1]Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos, Brazil

Abstract

Recommender systems are increasingly expected not only to provide accurate recommendations, but also to support fairness, transparency, and user trust. Calibration aims to align recommendation distributions with user preference profiles, while explainability helps users understand why specific items are recommended. Although both are commonly applied as post-processing strategies, their interaction remains largely unexplored. To address this gap, this work proposes a unified calibration-driven recommender system framework that integrates multi-category calibration with post-hoc knowledge-graph-based explanations. Unlike traditional calibration approaches that rely on a single category, such as genre, the proposed framework models multifaceted user preferences across multiple semantic dimensions extracted from Wikidata knowledge graphs. The proposed approach is model-agnostic and combines recommendation generation, fairness-centered calibration, and explanation generation within a single pipeline. The research investigates three main questions: (i) how multi-category calibration affects recommendation relevance and beyond-accuracy objectives; (ii) which trade-offs emerge between calibration and the preservation of user interests across categories; and (iii) how calibration influences the quality and alignment of recommendation explanations. Experiments are conducted on the MovieLens and LastFM datasets using BPR-MF and NCF recommendation models, leveraging semantic information extracted from Wikidata knowledge graphs to support both calibration and explanation generation. Recommendation quality is evaluated through metrics related to relevance, calibration, diversity, and popularity bias, while explanation quality is analyzed using metrics associated with recency, diversity, and popularity of explanatory entities. Preliminary results indicate that multi-category calibration improves alignment with users' multidimensional preference profiles while maintaining competitive accuracy and reducing popularity bias. The findings also suggest that post-hoc knowledge-graph-based explanations remain robust under calibration, indicating that fairness-aware calibration and explainability can coexist within a unified recommender system pipeline.

Keywords

Recommender Systems, Multi-Category Calibration, Explainable AI (XAI), Fairness-Aware Recommendation, Linked Open Data (LOD)

1. Context and Motivation

Recommender systems (RSs) have become essential tools of modern digital platforms. As the number of item choices grows, RSs help users navigate large choice spaces by providing personalized suggestions that support more efficient decision-making [1]. Consequently, these systems are widely deployed across domains, including e-commerce, multimedia streaming services, social networks, and online marketplaces [2].

Traditionally, recommender systems have emphasized accuracy objectives such as rating prediction and ranking relevance [1]. However, accuracy alone does not guarantee satisfactory or responsible user experiences. Research has therefore shifted toward beyond-accuracy criteria including diversity, novelty, serendipity, explainability, and fairness [3, 4].

Calibration is one fairness-oriented approach that aims to reflect users' interests in a balanced way. Steck [5] formalized calibration as a fairness-aware post-processing mechanism. However, existing calibration approaches mainly focus on a single categorical dimension (e.g., genre), implicitly assuming that preferences are one-dimensional and failing to capture the multifaceted nature of users' interests across multiple dimensions [6].

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil

paul.atauchi@usp.br (P. Atauchi)

0000-0003-2600-7836 (P. Atauchi)

On the other hand, explainability has emerged as a key concern in responsible AI design, aiming to clarify why specific items are recommended and thereby improving transparency, user trust, and engagement [7, 8, 9]. In many practical systems, explanations are generated post-hoc, conditioned on the outputs of the recommendation model. Consequently, although calibration strategies may mitigate biases at the recommendation level, recommendation explanations may still reflect biased or distorted semantic representations of user preferences.

Existing calibration approaches mainly focus on aligning recommendation lists with user preference distributions under single-category settings, such as genre. However, user preferences are inherently multifaceted and span multiple category dimensions, motivating the need for multi-category calibration strategies. Furthermore, recommendation explanations also convey semantic signals about user interests and should similarly reflect these multi-dimensional preferences. Therefore, this work investigates a unified framework for jointly aligning recommendation and explanation distributions under multifaceted user preference dimensions.

2. Related Work

2.1. The evolution of recommender systems

Recommender systems (RS) have evolved from simple collaborative filtering techniques into highly complex, large-scale decision-making engines [10]. Early work in the field focused primarily on predicting user ratings and estimating item relevance, with accuracy-oriented metrics such as RMSE, Precision@K, and Recall@K serving as the dominant evaluation criteria [1].

Over the past decade, recommender systems have expanded across domains such as e-commerce, multimedia streaming services, social media, and online marketplaces, where they must handle extremely large item catalogs, dynamic user behavior, contextual signals, and sequential interactions [11, 10]. Beyond relevance prediction, modern recommender systems are increasingly expected to support both platform and user objectives, including discovery, long-term satisfaction, engagement, and retention.

As the number of available options has grown, filtering (selecting eligible items) and ranking (ordering recommended items) have become core components of recommendation pipelines. Although advances in representation learning and deep learning have substantially improved predictive performance, many practical systems still prioritize accuracy-oriented optimization, often overlooking broader societal and user-centric concerns [10].

2.2. Beyond-Accuracy objectives in recommender systems

The limitations of accuracy-only evaluation are well documented in recommender systems literature. High accuracy does not necessarily imply a good user experience, as accurate systems may reinforce popularity bias, reduce exposure to novel content, fail to reflect diverse user interests, or produce unfair outcomes across users or item providers [11, 12, 4].

In response, research has increasingly emphasized objectives beyond accuracy. Metrics such as diversity promote variety within recommendation lists; novelty encourages exposure to unseen items; serendipity captures the ability to surprise users with unexpected yet relevant content; and fairness seeks equitable treatment and exposure across users, groups, or content providers [12, 4]. These objectives reflect the understanding that recommender systems influence not only what users consume, but also how their preferences evolve over time.

Among beyond-accuracy objectives, fairness has emerged as a critical dimension. Fairness in recommender systems encompasses multiple perspectives, including user-side fairness (ensuring recommendations reflect users' or groups' interests) and item-side or producer-side fairness (ensuring equitable exposure across items or providers) [4, 13]. Importantly, fairness objectives often interact with accuracy in non-trivial ways, requiring careful trade-off analysis rather than straightforward optimization.

With this context, calibration has been proposed as a principled mechanism for aligning recommendation outputs with user preference distributions. Steck [5] defines calibration as the degree to

which the distribution of recommended items matches the distribution observed in users' historical interactions. This formulation frames calibration as a fairness-aware objective that improves user preference alignment. Subsequent work has extended calibration to related concerns such as exposure control, constrained optimization, and producer fairness [14, 15, 16]. However, most existing approaches remain limited to single-category settings, although real user preferences are multidimensional [6, 17].

In parallel, there is growing recognition that user experience factors such as transparency, trust, and acceptance are essential to responsible recommender system design. Explanation plays a key role in this context by helping users understand and evaluate recommendations [7, 8, 9]. This discussion is also closely connected to the broader field of Explainable Artificial Intelligence (XAI), which seeks to make AI-driven decisions more transparent, interpretable, and trustworthy for users and stakeholders [18, 19]. Existing explanation techniques include feature-based, example-based, model-intrinsic, and post-hoc approaches [18]. Among these, knowledge-graph-based explanation methods have gained considerable attention in recommender systems because they can capture structured relational representations and provide interpretable paths between users and recommended items [20].

In practice, explainable recommendation systems commonly adopt a post-hoc or model-agnostic paradigm, where explanations are generated after the recommendation decision has already been made rather than being intrinsically integrated into the recommendation model itself [8]. In such approaches, the recommendation model is typically treated as a black box, while a separate explanation component is used to justify the recommended items. Prior work has evaluated explanation quality in recommender systems through both user-centric and algorithmic perspectives. User-centric evaluation commonly considers dimensions such as transparency, trustworthiness, persuasiveness, satisfaction, and effectiveness, while algorithmic evaluation focuses on factors including explanation fidelity, readability, textual similarity metrics such as BLEU and ROUGE, and offline explanation quality metrics related to recency, popularity, and diversity [7, 8, 21].

2.3. Towards Joint Recommendation and Explanation Alignment under Multifaceted User Preferences

Recent research increasingly investigates recommender system pipelines that jointly consider recommendation relevance, fairness, and transparency [9].

Beyond recommender systems, the XAI literature has also investigated the relationship between calibration and explanation quality. Löfström et al. [22] showed that calibrating predictive models can improve explanation reliability by aligning explanation confidence with predictive uncertainty. These findings reinforce that calibration and explainability should not be treated as isolated properties.

Motivated by this perspective, this research investigates the joint alignment of recommendation and explanation distributions under multifaceted user preferences. More specifically, the proposed research studies how multi-category calibration strategies influence not only recommendation distributions, but also the semantic properties conveyed by post-hoc knowledge-graph-based explanations. The central hypothesis is that recommendation explanations should reflect the same multifaceted preference distributions considered during recommendation calibration.

3. Research Questions

This work is guided by three research questions:

RQ1: How does multi-category calibration affect recommendation relevance and beyond-accuracy objectives compared to non-calibrated and single-category calibrated recommender systems?

Existing calibration approaches are predominantly single-dimensional and assume that user preferences can be represented by a single category (e.g., genre). This research question investigates whether extending calibration across multiple categories influences recommendation quality and beyond-accuracy objectives.

RQ2: What trade-offs emerge between multi-category calibration and the preservation of individual users' multifaceted preference profiles?

While multi-category calibration aims to improve distributional alignment across recommendation categories, enforcing calibration constraints may also alter the representation of individual user interests. This research question investigates how calibration affects the fidelity of personalized preference distributions across multiple semantic dimensions.

RQ3: How does multi-category calibration influence the alignment and quality of recommendation explanations compared to non-calibrated and single-category calibrated recommender systems?

Most explanation methods are developed independently of fairness-aware calibration mechanisms, implicitly assuming unconstrained recommendation lists. This research question investigates whether calibration-driven reranking influences the semantic properties, alignment, and quality of post-hoc recommendation explanations across different explanation paradigms under non-calibrated, single-category calibrated, and multi-category calibrated settings.

4. Methodology

The proposed methodology investigates recommendation and explanation alignment under multifaceted user preferences. More specifically, the experimental pipeline evaluates how multi-category strategies influence: (i) recommendation relevance and beyond-accuracy objectives, (ii) the preservation of individualized multifaceted preference distributions, and (iii) the semantic alignment and quality of post-hoc recommendation explanations. Figure 1 shows the proposed framework adopted in this research. The framework is organized into three main modules: recommendation generation, multi-category calibration, and explanation generation.

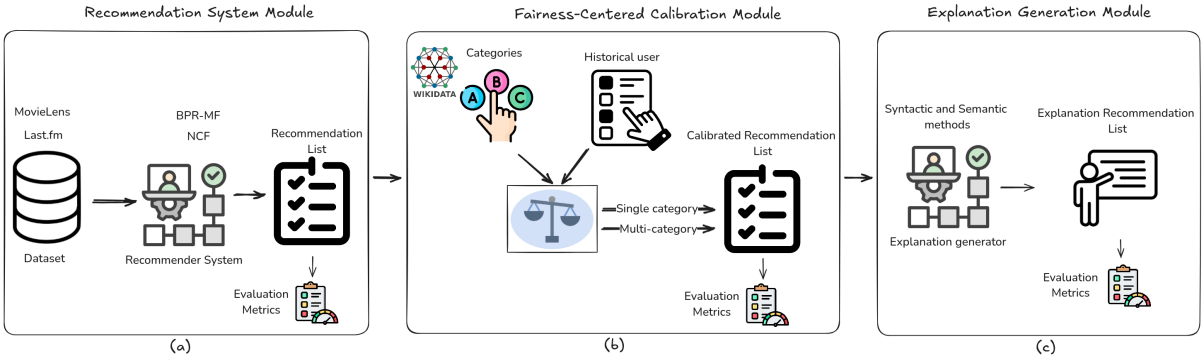

Figure 1: Overview of the proposed framework, which consists of three components: (a) recommendation generation, (b) recommendation calibration, and (c) explanation generation.

4.1. Recommender System Module

As shown in Figure 1(a), the recommender system module receives a processed user-item interaction dataset and produces a ranked top-N recommendation list I_u for each user u. The framework is model-agnostic and supports different recommendation paradigms, enabling the analysis of calibration and explanation alignment under distinct recommendation approaches.

Experiments consider representative collaborative filtering models, including Bayesian Personalized Ranking Matrix Factorization (BPR-MF) [23] and Neural Collaborative Filtering (NCF) [24], representing matrix factorization and neural recommendation architectures. Evaluations are conducted on publicly available datasets from different domains, including MovieLens [25] and LastFM [26].

To support multi-category calibration and explanation generation, item metadata and semantic relations are enriched using knowledge graphs extracted from Wikidata [1].

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

4.2. Fairness-Centered Calibration Module

Following recommendation generation, a post-processing calibration module is applied to align recommendation distributions with multifaceted user preference distributions. Existing calibration approaches predominantly focus on single-dimensional preference alignment, typically considering only one category, such as genre [5]. In contrast, the proposed framework investigates multi-category calibration strategies that simultaneously consider multiple semantic dimensions.

The calibration module extends traditional calibration formulations by incorporating multiple semantic categories derived from knowledge graphs into the calibration objective, as illustrated in Figure 1(b). The proposed formulation aims to align the category distributions in recommendation lists with the multifaceted preference distributions extracted from users' historical interactions. Experimental evaluation is conducted across three recommendation settings: non-calibrated, single-category calibrated, and multi-category calibrated recommendation settings, considering both recommendation relevance and beyond-accuracy objectives, including fairness, diversity, popularity, and distributional alignment metrics such as MRMC [27].

4.3. Explanation Generation Module

The third module investigates how calibration affects recommendation explanations (see Figure 1(c)). In many recommender systems, explanations are generated post-hoc from recommendation outputs [8, 9]. Consequently, although calibration strategies may mitigate biases at the recommendation level, additional biases and semantic distortions may still emerge during explanation generation due to the explanation mechanism itself.

The framework considers post-hoc knowledge-graph-based explanation approaches, including syntactic techniques such as ExpLOD [28], ExpLOD v2 [29], and PEM [30], which operate on the structural topology of the knowledge graph by identifying explanation paths based on graph connectivity patterns and attribute frequencies. It also includes semantic approaches such as ComplEx [31], which capture latent relational semantics through vector-space representations, as well as generative approaches that leverage graph structures and learned relational patterns to produce explanations [32].

The evaluation investigates whether multi-category calibration influences not only recommendation distributions, but also explanation alignment and explanation quality. More specifically, the analysis focuses on whether generated explanations preserve semantic consistency with multifaceted user preferences under different calibration settings.

Explanation evaluation may incorporate metrics related to explanatory quality and alignment. Comparative analyses are conducted across non-calibrated, single-category calibrated, and multi-category calibrated recommendation settings. Furthermore, the relationship between recommendation calibration, explanation quality, and explanation alignment is motivated by prior findings from the XAI literature, which suggest that calibration can affect explanation quality [22].

4.4. Experimental Analysis

The experimental analysis investigates the trade-offs among recommendation relevance, fairness-oriented calibration, preservation of multifaceted user preferences, explanation quality, and explanation alignment. The analysis is guided by the proposed research questions and focuses on understanding how calibration constraints affect both recommendation outcomes and generated explanations.

5. Preliminary Results and Contributions to Date

This research is grounded in three prior studies that investigate the relationship between calibration, fairness, and explainability in recommender systems.

Study 1: Single-Category Calibration and Explanation Quality (RQ3)

The first study examines whether single-category calibration, applied as a post-processing reranking step, affects the quality of post-hoc knowledge-graph-based explanations. Across both BPR-MF and NCF recommendation models on MovieLens and LastFM, the results suggest that post-hoc explanation approaches remain relatively robust under single-category calibration. This indicates that fairness-driven calibration can be integrated without substantially degrading explanation quality. However, because the observed changes in explanation metrics were relatively small, the results also suggest that single-category calibration may have limited influence on the semantic properties captured by explanation approaches, motivating the exploration of multi-category calibration strategies. This study has been published in [33].

Study 2: Multi-Category Calibration and Preference Alignment (RQ1, RQ2)

The second study extends traditional single-category calibration to the multi-category setting. The study introduced a calibration strategy based on multiple knowledge graph attributes and evaluated its effects on recommendation quality, calibration, and preference alignment. Experimental results indicated improvements in the alignment between recommendation lists and user preference distributions while maintaining competitive recommendation accuracy across different recommendation models. The study also observed reductions in popularity bias and more balanced recommendation distribution across categories. Results from this study have been published in [34].

Study 3: Multi-Category Calibration and Explanation Quality (RQ3)

The third study further investigated RQ3 through the integration of fairness-aware calibration and post-hoc explainability within a unified recommendation framework. The findings indicate that multi-category calibration can achieve more balanced trade-offs between relevance, calibration, diversity, and popularity bias when compared to single-category approaches. Additionally, explanation quality metrics, including recency, diversity, and popularity attribute measures, remained relatively stable across different calibration settings. In several experimental scenarios, multi-category calibration achieved higher overall utility scores.

Overall, these preliminary findings provide evidence that calibration-based fairness strategies can improve recommendation alignment and mitigate popularity bias without compromising explanation quality. They also reinforce the central motivation of this doctoral research, which investigates fairness and explainability as interconnected rather than isolated objectives in responsible recommender systems.

6. Future Work and Expected Contributions

While preliminary studies establish the viability of a unified calibration-explanation pipeline, several directions remain open. Current experiments are limited to movie and music recommendation domains. The framework will be extended to domains such as e-commerce and job recommendation to investigate whether the observed calibration-explanation behavior generalizes across settings with different levels of item metadata density in Wikidata. Furthermore, since calibration strategies rely on historical user interactions to estimate user preference distributions, scenarios with limited interaction data may affect the reliability of preference estimation in both single-category and multi-category settings. This motivates future investigations exploring additional information sources for calibration.

Current explanation quality evaluation relies on offline metrics (e.g., LIR, SEP, ETD), which capture structural and system-side properties of explanations but do not directly assess user perception or whether explanations remain aligned with the user preference distributions considered during calibration. Therefore, future work includes conducting user studies to evaluate whether calibrated recommendations and their generated explanations are perceived as more trustworthy, helpful, and fair

by real users, as well as investigating and developing offline metrics capable of measuring explanation alignment under multi-category calibration settings.

By addressing these directions, this research aims to deliver a comprehensive and empirically grounded framework that bridges fairness-aware and explainable recommender systems.

Acknowledgments

I would like to sincerely thank Prof. Marcelo Manzato, my advisor, and Prof. Leonardo Rocha, my co-supervisor, for their guidance and support. This research was funded by CNPq and CAPES.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT-5.3 exclusively for grammar checking, spelling correction, and minor rephrasing. All original content was written by the author.

References

[1] C. C. Aggarwal, et al., Recommender systems, volume 1, Springer, 2016.

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

[3] M. Naghiaei, M. Dehghan, H. A. Rahmani, J. Azizi, M. Aliannejadi, Personalized beyond-accuracy calibration in recommendation, in: Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval, 2024, pp. 107–116.

[4] Y. Li, H. Chen, S. Xu, Y. Ge, J. Tan, S. Liu, Y. Zhang, Fairness in recommendation: Foundations, methods, and applications, ACM Transactions on Intelligent Systems and Technology 14 (2023) 1–48.

[5] H. Steck, Calibrated recommendations, in: Proceedings of the 12th ACM conference on recommender systems, 2018, pp. 154–162.

[6] D. C. da Silva, D. Jannach, Calibrated recommendations: Survey and future directions, ACM Transactions on Recommender Systems (2025).

[7] N. Tintarev, J. Masthoff, Explaining recommendations: Design and evaluation, in: Recommender systems handbook, Springer, 2015, pp. 353–382.

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

[9] Y. Ge, S. Liu, Z. Fu, J. Tan, Z. Li, S. Xu, Y. Li, Y. Xian, Y. Zhang, A survey on trustworthy recommender systems, ACM Transactions on Recommender Systems 3 (2024) 1–68.

[10] S. Raza, M. Rahman, S. Kamawal, A. Toroghi, A. Raval, F. Navah, A. Kazemeini, A comprehensive review of recommender systems: Transitioning from theory to practice, Computer Science Review 59 (2026) 100849.

[11] J. A. Konstan, J. Riedl, Recommender systems: from algorithms to user experience, User modeling and user-adapted interaction 22 (2012) 101–123.

[12] P. Castells, N. Hurley, S. Vargas, Novelty and diversity in recommender systems, in: Recommender systems handbook, Springer, 2021, pp. 603–646.

[13] Y. Wang, W. Ma, M. Zhang, Y. Liu, S. Ma, A survey on the fairness of recommender systems, ACM Transactions on Information Systems 41 (2023) 1–43.

[14] S. Seymen, H. Abdollahpouri, E. C. Malthouse, A constrained optimization approach for calibrated recommendations, in: Proceedings of the 15th ACM Conference on Recommender Systems, 2021, pp. 607–612.

[15] H. Abdollahpouri, Z. Nazari, A. Gain, C. Gibson, M. Dimakopoulou, J. Anderton, B. Carterette, M. Lalmas, T. Jebara, Calibrated recommendations as a minimum-cost flow problem, in: Proceed-

ings of the Sixteenth ACM International Conference on Web Search and Data Mining, 2023, pp. 571–579.
[16] A. Sacilotti, R. F. d. Souza, M. G. Manzato, Counteracting popularity-bias and improving diversity through calibrated recommendations, in: Proceedings, 2023.
[17] Z. Li, Q. Chen, L. Zou, A. Sun, C. Li, Multi-interest recommendation: A survey, ACM Transactions on Information Systems 44 (2026) 1–38.
[18] A. Abusitta, M. Q. Li, B. C. Fung, Survey on explainable ai: Techniques, challenges and open issues, Expert Systems with Applications 255 (2024) 124710.
[19] K. Kalasampath, K. Spoorthi, S. Sajeev, S. S. Kuppa, K. Ajay, A. Maruthamuthu, A literature review on applications of explainable artificial intelligence (xai), IEEE access 13 (2025) 41111–41140.
[20] T. Markchom, H. Liang, J. Ferryman, Review of explainable graph-based recommender systems, ACM Computing Surveys 58 (2025) 1–35.
[21] G. Balloccu, L. Boratto, G. Fenu, M. Marras, Post processing recommender systems with knowledge graphs for recency, popularity, and diversity of explanations, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 646–656.
[22] H. Löfström, T. Löfström, U. Johansson, C. Sönströd, Investigating the impact of calibration on the quality of explanations, Annals of Mathematics and Artificial Intelligence (2023) 1–18.
[23] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, Bpr: Bayesian personalized ranking from implicit feedback, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI '09, AUAI Press, Arlington, Virginia, USA, 2009, p. 452–461.
[24] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T.-S. Chua, Neural collaborative filtering, in: Proceedings of the 26th international conference on world wide web, 2017, pp. 173–182.
[25] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, Acm transactions on interactive intelligent systems (tiis) 5 (2015) 1–19.
[26] I. Cantador, P. Brusilovsky, T. Kuflik, 2nd workshop on information heterogeneity and fusion in recommender systems (hetrec 2011), in: Proceedings of the 5th ACM conference on Recommender systems, RecSys 2011, ACM, New York, NY, USA, 2011.
[27] D. C. da Silva, M. G. Manzato, F. A. Durão, Exploiting personalized calibration and metrics for fairness recommendation, Expert Systems with Applications 181 (2021) 115112.
[28] C. Musto, F. Narducci, P. Lops, M. De Gemmis, G. Semeraro, Explod: A framework for explaining recommendations based on the linked open data cloud, in: Proceedings of the 10th ACM Conference on Recommender Systems, 2016, pp. 151–154.
[29] C. Musto, F. Narducci, P. Lops, M. De Gemmis, G. Semeraro, Linked open data-based explanations for transparent recommender systems, International Journal of Human-Computer Studies 121 (2019) 93–107.
[30] Y. Du, S. Ranwez, N. Sutton-Charani, V. Ranwez, Post-hoc recommendation explanations through an efficient exploitation of the dbpedia category hierarchy, Knowledge-Based Systems 245 (2022) 108560.
[31] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, G. Bouchard, Complex embeddings for simple link prediction, in: International conference on machine learning, PMLR, 2016, pp. 2071–2080.
[32] A. Said, On explaining recommendations with large language models: a review, Frontiers in big Data 7 (2025) 1505284.
[33] P. D. F. Atauchi, A. L. Zanon, L. C. D. d. Rocha, M. G. Manzato, Do calibrated recommendations affect explanations? a study on post-hoc adjustments, Journal on Interactive Systems 16 (2025) 441–460.
[34] P. Atauchi, A. Zanon, L. Rocha, M. Manzato, Aprimorando a personalização de sistemas de recomendação por calibração multicategoria, in: Proceedings of the 31st Brazilian Symposium on Multimedia and the Web, SBC, Porto Alegre, RS, Brasil, 2025, pp. 506–510.