

Towards a Methodology to Assess the Impact of LLMs on Recommendations

Marta Salcedo

marta.salcedos@estudiante.uam.es
Universidad Autónoma de Madrid
Madrid, Spain

Alejandro Bellogín

alejandro.bellogin@uam.es
Universidad Autónoma de Madrid
Madrid, Spain

Abstract

Large Language Models (LLMs) are increasingly used in recommender systems, raising important questions about their societal impact, including bias, transparency, and effects on information diversity. In this work in progress, we propose an initial methodological framework to assess the impact of LLM-based recommendations beyond traditional accuracy metrics. Our approach considers multiple dimensions such as calibration and bias propagation, enabling comparisons with conventional systems. We present preliminary observations from an ongoing case study, highlighting both potential benefits and risks. This early work aims to stimulate discussion and guide the development of more responsible and socially-aware recommender systems.

CCS Concepts

• Information systems → Recommender systems; Evaluation of retrieval results.

Keywords

Recommender Systems, Evaluation, Fairness, LLM

ACM Reference Format:

Marta Salcedo and Alejandro Bellogín. 2026. Towards a Methodology to Assess the Impact of LLMs on Recommendations. In *International Conference on Information Technology for Social Good (GoodIT '26)*, September 02–04, 2026, Pisa, Italy. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3794786.3830753>

1 Introduction

Large Language Models (LLMs) are increasingly being integrated into Recommender Systems (RS) [10], enabling more conversational and context-aware interactions. However, their adoption raises important questions about their broader societal impact, including potential biases, effects on information diversity, and implications for transparency and user trust [8]. Despite growing interest in LLM-based recommenders, there is still a lack of systematic approaches to assess these dimensions in a unified way.

In this work-in-progress paper, we present an initial methodological framework to evaluate the impact of LLM-driven recommendations beyond traditional accuracy metrics. Our approach explores a set of complementary dimensions, including calibration, exposure

diversity, bias propagation, and fairness, aiming to capture both system-level and user-centered effects. The proposed framework is model-agnostic and is intended to support comparative analyses between LLM-based and traditional recommendation approaches.

As an initial validation, we instantiate the proposed methodology in a movie recommendation scenario using the MovieLens dataset, which provides both item attributes (e.g., movie genres) and user demographic information (e.g., age, gender, and occupation) required by the considered evaluation dimensions.

We report preliminary results from an ongoing case study in a real-world recommendation scenario, highlighting early observations on how the integration of LLMs can influence recommendation outcomes and user exposure patterns. These initial findings suggest both opportunities for improved contextualization and potential risks related to bias amplification and reduced exposure.

This work represents a first step toward a comprehensive evaluation methodology for responsible LLM-based recommender systems. While individual metrics for diversity, fairness, and calibration have previously been studied in isolation, this work investigates their integration into a unified evaluation methodology for LLM-based recommender systems, an area that remains largely unexplored [19]. Through this contribution, we aim to foster discussion within the community and gather feedback to refine the framework, extend its evaluation dimensions, and better understand the role of LLMs in shaping socially-aware recommendation technologies, in agreement with recent mandates for responsible AI [3]

2 Background and Related Work

Recent advances in Large Language Models have led to their increasing adoption in recommender systems, enabling conversational, context-aware, and generative recommendation capabilities [10]. In this context, platforms such as OpenP5 [24] have emerged to support the development and benchmarking of LLM-based recommendation approaches, providing standardized datasets, model implementations, and evaluation pipelines across multiple recommendation tasks. By framing recommendation as a language modeling problem, OpenP5 facilitates unified experimentation and reproducibility, lowering the barrier to entry for research in generative recommendation [24]. However, despite these advances, evaluation within such frameworks remains largely focused on traditional performance metrics (e.g., accuracy and ranking quality), with limited attention to broader societal implications.

Traditional RS evaluation has predominantly focused on accuracy-oriented metrics, such as precision, recall, and other ranking-based measures, which quantify how well a system predicts user preferences or retrieves relevant items [15]. These metrics have become the *de facto* standard for comparing algorithms, particularly



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

GoodIT '26, Pisa, Italy

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2483-1/2026/09

<https://doi.org/10.1145/3794786.3830753>

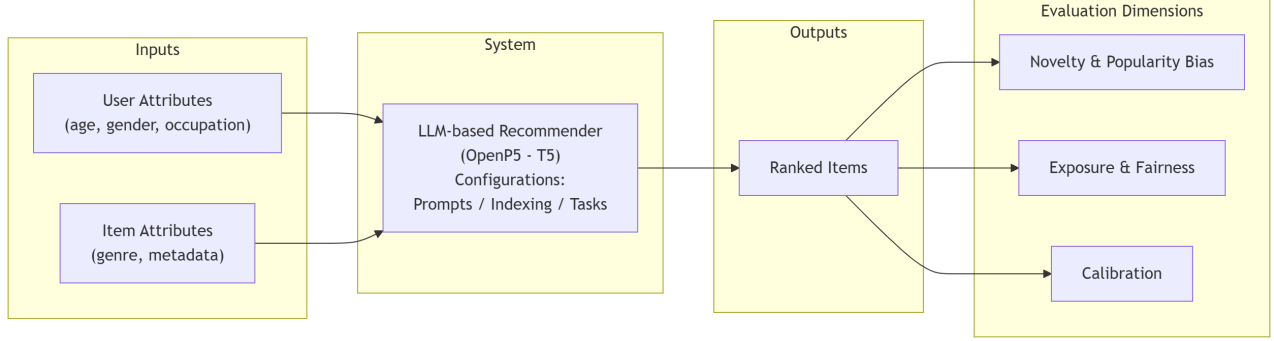


Figure 1: Overview of the proposed methodology as a pipeline diagram.

in offline evaluation settings, where improvements in predictive performance are often prioritized [4]. However, this perspective provides only a partial view of system quality, as it does not capture other important aspects such as diversity, novelty, and long-term user experience [5]. As recommender systems increasingly mediate access to information, the limitations of purely accuracy-driven evaluation have become more apparent [16].

Beyond performance considerations, a growing body of research has highlighted societal risks associated with recommender systems, including bias, fairness issues, and the emergence of filter bubbles [11–13, 17]. Data and algorithmic biases can lead to disproportionate exposure of popular items and under-representation of minority or niche content, raising concerns about fairness for both users and providers. Moreover, personalization mechanisms may contribute to filter bubbles, where users are repeatedly exposed to homogeneous content, potentially reinforcing existing beliefs and limiting exposure to diverse perspectives. These challenges underscore the need for evaluation approaches that go beyond effectiveness and explicitly account for fairness, diversity, transparency, and broader societal impact, motivating the methodology proposed in this work.

3 Proposed Methodology

To systematically assess the societal implications of LLM-based recommender systems, we structure our methodology along three complementary components, which are visually represented in the diagram shown in Figure 1: input information, recommender system, and the evaluation of the recommendation output as a ranking of items. Next, we first provide an overview of the proposed evaluation framework, outlining its goals, scope, and design principles (Section 3.1). We then define a set of key evaluation dimensions that capture different aspects of societal impact, including bias, fairness, and calibration (Section 3.2). Finally, we describe the operationalization of these dimensions through concrete metrics and an experimental setup, enabling a structured and reproducible analysis of system behavior (Section 3.3).

3.1 Overview

This work-in-progress proposes a methodological framework to assess the **societal impact of LLM-based recommender systems** and to diagnose potential risks arising from their generated outputs.

The goal is not only to evaluate effectiveness, but also to identify and characterize biases, exposure imbalances, and mismatches between recommendations and user or item attributes.

Our framework focuses on LLM-based recommender systems implemented under the *Pretrain–Prompt–Predict* paradigm, as instantiated by OpenP5 [24], where recommendation is formulated as a language modeling task. Within this setting, different configurations can be explored, including variations in prompting strategies, item indexing mechanisms, and recommendation tasks (e.g., sequential and straightforward recommendation). Some examples of these prompting strategies, as presented in [24], are shown in Table 1.

The proposed approach adopts an **offline evaluation setting**, where recommendation outputs are analyzed using both item metadata (e.g., genre) and user attributes (e.g., gender, age, occupation), depending on data availability. The framework is designed as a **multi-dimensional evaluation pipeline**, enabling a diagnostic analysis of system behavior across different configurations. Importantly, the methodology is implemented in an extensible way using the Elliot evaluation framework [2], allowing the integration of additional metrics and dimensions as needed. Rather than comparing against external baselines, our focus is on **understanding how design choices within LLM-based RS** influence their societal impact.

3.2 Evaluation dimensions

The proposed dimensions were selected to capture three complementary perspectives of societal impact that have been extensively discussed in the recommender systems literature. Novelty and popularity bias address concentration effects and the visibility of long-tail content, which are closely related to diversity and information access. Exposure and fairness focus on how recommendation opportunities are distributed across user and item groups, reflecting concerns about equitable treatment and representation. Finally, calibration evaluates whether recommendation outputs faithfully reflect user preferences and item characteristics, thus relating to user trust and transparency.

Together, these dimensions enable the assessment of who is exposed to content, what type of content is promoted, and how well recommendations align with user profiles, providing a broader view of societal impact than traditional effectiveness metrics alone.

Table 1: Examples of different prompting strategies, as proposed in [24].

Task	Prompt	Input
straightforward	seen	What should we recommend for {dataset} user_{user_id} ?
straightforward	unseen	What is the top recommendation for {dataset} user_{user_id} ?
sequential	seen	Considering {dataset} user_{user_id} has interacted with {dataset} items {history} . What is the next recommendation for the user ?
sequential	unseen	What is the top recommended item for {dataset} user_{user_id} who interacted with {dataset} item {history} ?

Other dimensions, such as privacy, explainability, environmental sustainability, or long-term behavioral effects, are also relevant but are left for future extensions of the framework.

Novelty and Popularity Bias (D1). This dimension evaluates the extent to which the recommender system promotes popular items versus less-known (long-tail) content. A strong reliance on popular items may lead to homogenization of recommendations and reduced discovery, while promoting novelty can enhance user experience and mitigate concentration effects. Analyzing this trade-off is particularly relevant for LLM-based systems, which may implicitly encode popularity signals through pretraining data.

Exposure and Fairness (D2). This dimension captures how recommendation outputs distribute visibility across different groups of items or users. It focuses on identifying imbalances in exposure, such as systematic over-representation of certain item categories or unfair treatment of specific groups. Ensuring fairness in exposure is critical for maintaining equitable access to content and avoiding reinforcing existing disparities in information ecosystems.

User and Item Calibration (D3). Calibration measures the alignment between the characteristics of recommended items and the preferences or attributes of users, as well as the overall distribution of item properties. A well-calibrated recommender system reflects the diversity of user interests and maintains appropriate proportions of item attributes (e.g., genres), while miscalibration may lead to distorted or misleading representations of user preferences.

3.3 Operationalization

The proposed evaluation dimensions are operationalized through a set of established metrics and an experimental protocol tailored to LLM-based recommender systems.

For novelty and popularity bias (D1), we consider metrics such as *Average Percentage of Long Tail Items* (APLT) [1], which measures the proportion of non-popular items in the recommendation lists, *Average Recommendation Popularity* (ARP) [25], capturing the overall popularity level of recommended items, and *Expected Popularity Complement* (EPC), which quantifies novelty as the inverse of item popularity [5].

For exposure and fairness (D2), we adopt a range of metrics that assess ranking fairness and group exposure, including *Popularity-based Ranking Equal Opportunity* (PopREO), *Ranking-based Equal Opportunity* (REO), and *Ranking-based Statistical Parity* (RSP) [26], which evaluate how evenly items from different groups are represented in top-ranked positions. Additionally, *Mean Absolute Deviation* (MAD) and *Bias Disparity* (BD) are used to quantify the extent

of imbalance in recommendation performance across item or user groups [9, 23].

For calibration (D3), we evaluate the extent to which the distribution of attributes in the recommendation lists reflects the corresponding distributions in user profiles or item spaces [7, 21]. This allows us to assess whether recommendations are aligned with user interests and whether they preserve meaningful proportions of item characteristics. We note that we use the original formulation from [21], however, since it is restricted to be computed based on item attributes, we extend the formulation to also consider user attributes by representing items according to the features expressed by the users who rated each specific item.

The evaluation is conducted on the **MovieLens** dataset¹, leveraging both user and item attributes. We consider recommendation tasks consistent with the OpenP5 framework, including sequential and straightforward recommendation. The LLM-based recommender system is instantiated using the T5 model within OpenP5, as it evidenced better results than LLaMa [24], and multiple configurations are explored as in the original work.

The output of each configuration is a **ranked list of items**, which is processed using a standard evaluation pipeline integrated into the Elliot framework, something not supported natively in the original framework provided in [24]. By systematically comparing the rankings produced under different configurations across the defined dimensions, the proposed methodology enables a diagnostic analysis of how LLM-based recommender systems shape recommendation outcomes and their associated societal implications.

4 Preliminary Results

Table 2 shows the results obtained for the multi-dimensional evaluation methodology followed, as explained in the previous section, together with accuracy results from ranking metrics like Precision, Recall, Hit Rate (HR), normalized Discounted Cumulative Gain (nDCG), and Mean Average Precision (MAP) [15], to contextualize and properly assess the tradeoffs achieved with respect to the considered dimensions. All metrics have been computed using the top-10 items recommended for every user, that is, at cutoff 10.

Regarding **accuracy results**, we observe that the sequential recommendation task (where the model is aware of the user history in the prompt, as specified in [24]) obtains higher results than the straightforward task (where the prompt only includes the user id). This is consistent with the experiments reported in [24]. Also in agreement with the original work, using prompts not seen during

¹<https://grouplens.org/datasets/movielens/1m/>

Table 2: Multi-dimensional evaluation results, using the terminology from [24] (Rnd denotes random item indexing, Seq is sequential, and Clb collaborative) and the evaluation metrics described in Section 3.2. Best result overall with †, per task in bold.

Task	Prompt	Idx	Accuracy			D1: novelty & bias					D2: exposure & fairness					D3: Calibration					
			Prec↑	Recall↑	HR↑	nDCG↑	MAP↑	APLT↑	ARP↑	EPC↑	PopREO↓	MAD↓	REO↓	RSP↓	BD-GG↓	BD-GA↓	BD-GO↓	GenreCal↓	AgeCal↓	GenderCal↓	OccCal↓
Sfwd	Seen	Rnd	0.0010	0.0100	0.0100	0.0100	0.0029	0.0000	42.0372	0.0019	1.0000	0.0004	17.3205	8.1372	-0.3424	-0.0122	0.7336	1.7579	†0.0971	0.0322	†0.1996
Sfwd	Seen	Seq	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0128	24.3276	0.0000	†0.0000	†0.0000	4.6621	-0.3651	†-0.1118	0.6159	1.9946	0.1637	0.0306	0.3645
Sfwd	Seen	Clb	0.0030	0.0300	0.0300	0.0226	0.0068	0.0526	24.5914	0.0068	1.0000	0.0047	10.3615	5.1619	-0.2582	0.0988	0.9797	†1.6425	0.1399	0.0169	0.3329
Sfwd	Unseen	Rnd	0.0020	0.0200	0.0200	0.0139	0.0038	0.0000	42.4789	0.0026	1.0000	0.0008	12.8927	8.1204	-0.3443	-0.0097	0.7446	1.7974	0.0976	0.0324	0.2044
Sfwd	Unseen	Seq	0.0010	0.0100	0.0100	0.0063	0.0019	0.0090	23.8257	0.0027	1.0000	0.0025	17.3205	4.9443	-0.2284	-0.0161	0.7711	1.9281	0.1544	0.0246	0.3395
Sfwd	Unseen	Clb	0.0030	0.0300	0.0300	0.0189	0.0058	0.0354	25.9603	0.0056	1.0000	0.0055	10.3615	5.0308	-0.2376	0.0967	0.9612	1.6507	0.1333	†0.0155	0.3008
Sqtl	Seen	Rnd	0.0020	0.0200	0.0200	0.0083	0.0019	0.0000	43.6310	0.0015	1.0000	0.0008	12.8927	8.2586	-0.3876	-0.0912	0.6577	2.0269	0.0994	0.0360	0.2224
Sqtl	Seen	Seq	†0.0060	†0.0600	†0.0600	†0.0430	†0.0119	†0.1327	†16.8902	†0.0081	1.0000	0.0298	7.5899	†2.4463	0.0163	0.1983	0.6632	1.6883	0.2144	0.0398	0.5266
Sqtl	Seen	Clb	0.0030	0.0300	0.0300	0.0156	0.0045	0.0680	19.9374	0.0033	†0.0847	0.0129	11.9694	3.2843	-0.1034	0.5264	1.1328	1.6885	0.1716	0.0284	0.3671
Sqtl	Unseen	Rnd	0.0020	0.0200	0.0200	0.0079	0.0017	0.0000	43.2123	0.0014	1.0000	0.0008	12.8927	8.2845	†-0.3879	-0.0903	0.6127	2.0182	0.0997	0.0363	0.2232
Sqtl	Unseen	Seq	†0.0060	0.0600	0.0600	0.0369	0.0100	0.1294	17.1995	0.0066	1.0000	0.0379	7.5422	2.6499	-0.0907	0.1631	†0.5002	1.7596	0.2383	0.0364	0.5604
Sqtl	Unseen	Clb	0.0050	0.0500	0.0500	0.0284	0.0076	0.0729	19.3182	0.0064	0.2558	0.0105	10.3988	3.2087	-0.1328	0.4768	0.9620	1.7075	0.1664	0.0286	0.3625

training to evaluate the model’s zero-shot generalization capabilities (*unseen* prompts) obtains worse results. In terms of the type of item indexing, random indexing (*Rnd*) always performs the worst, as in [24]; however, for this dataset the authors almost always obtained better results for collaborative indexing (*Clb*) than for sequential indexing (*Seq*), something we also observed but only for the straightforward task. This highlights the importance of reproducibility and the use of independent evaluation frameworks in recommendation, a well-known problem in the community [20].

Moving to the analysis of the first dimension captured by our framework (**novelty and popularity bias**), we consider the metrics APLT, ARP, and EPC that evidence that sequential indexing for the sequential task produce optimal results. It is quite interesting to note that random indexing produces in every case highly skewed recommendations towards popular items (zero APLT, very high ARP).

In terms of D2 (**exposure and fairness**), it is important to consider that exposing every item without considering user preferences (like a random recommender would do) entails optimal exposure and a fair item distribution, but no user satisfaction. Because of this, the best results found for the item sequential indexing in the straightforward task are not meaningful, since its accuracy is zero. However, the sequential task, especially with unseen prompts, obtains much better results in terms of exposure (REO, RSP) and fairness (MAD, BD-GG, BD-GO). Fairness metrics need to have user or item groups defined to account for imbalances. In this dataset we considered genre for items, and age, gender, or occupation for users. Hence, bias disparity (BD) would indicate the extent a recommendation model mirrors the bias found in the data with respect to a combination of these groups for users and items, meaning a value close to 0 implies it neither amplifies nor mitigates existing biases, whereas a positive (negative) value suggests the model introduces a higher (lower) bias than what is present in the original data. In our case, GG corresponds to the genre-gender combination, GA to genre-age, and GO to genre-occupation. We observe the GG case is the easiest to mitigate its original bias, whereas GO represents the opposite, as all the obtained values are positive.

A plausible explanation is that sequential indexing preserves semantic and behavioral relationships between items more effectively than random indexing, reducing the tendency of the model to concentrate recommendations on a limited subset of popular items. This broader exposure naturally benefits fairness-oriented metrics,

as recommendation opportunities become more evenly distributed across groups.

Regarding D3 (**calibration**), optimal methods according to accuracy seem to also be aligned in terms of genre calibration. For age and occupation, random indexing performs better in both tasks; whereas for gender, collaborative indexing produces better results. We also observe that age and gender calibration obtain lower results, evidencing that this type of calibration is *easier* to achieve with the recommendation approaches.

The stronger calibration results observed for genre may be explained by the direct availability of genre information in user interaction histories, whereas demographic attributes such as occupation or age are only indirectly reflected in rating behavior, making them harder for the model to capture.

In summary, these results suggest that configurations maximizing effectiveness are not necessarily those that maximize fairness or calibration, reinforcing the need for multi-dimensional assessments when deploying LLM-based recommender systems.

5 Discussion and Broader Impact

The growing integration of Large Language Models into recommender systems introduces not only new technical capabilities but also significant societal implications that remain insufficiently understood. The contribution of this work is not the proposal of new metrics, but the definition of a reproducible evaluation pipeline that combines existing measures into a coherent framework for assessing societal impact. Our findings highlight the need to move beyond accuracy-driven evaluation and incorporate dimensions such as fairness, novelty, and calibration with respect to relevant user or item aspects to ensure that LLM-based recommendations contribute positively to social good.

5.1 Operational considerations

The results obtained in this work highlight several practical considerations for the deployment of LLM-based recommender systems in real-world settings. First, the observed trade-offs between accuracy, novelty, and fairness suggest that optimizing a single objective is insufficient, and system designers must explicitly balance multiple, often conflicting criteria depending on the target application [16]. In this context, recent work suggests that fairness can be influenced through prompt design [18] and strategic prompting [6], enabling practitioners to steer recommendation outputs toward more equitable exposure patterns without modifying the underlying model.

Table 3: Results using fairness-oriented prompts. Better results than baseline prompts (Table 2) marked with ‡.

Prompt	Idx	Prec↑	Recall↑	Accuracy			D1: novelty & bias				D2: exposure & fairness				D3: Calibration					
				HR↑	nDCG↑	MAP↑	APLT↑	ARP↓	EPC↑	PopREO↓	MAD↓	REO↓	RSP↓	BD-GG↓	BD-GA↓	BD-GO↓	GenreCal↓	AgeCal↓	GenderCal↓	OccCal↓
Seen	Rnd	0.0020	0.0200	0.0200	0.0136	0.0036	0.0000	41.5988	0.0023	1.0000	0.0011	12.8927	7.7050	-0.3184	-0.0076	0.8835	1.8002	0.0969	0.0312	0.2013
Seen	Seq	0.0050	0.0500	0.0500	0.0400	0.0116	0.1273	‡16.8654	‡0.0087	1.0000	0.0239	8.2935	2.6609	0.0600	0.1730	0.6626	1.6971	0.2114	0.0357	0.5329
Seen	Cib	0.0030	0.0300	0.0300	0.0169	0.0050	0.0730	19.8038	0.0038	0.0847	0.0140	11.9694	3.1801	-0.1069	0.5079	1.0451	1.7203	0.1741	0.0281	0.3760
Unseen	Rnd	0.0010	0.0100	0.0100	0.0100	0.0029	0.0000	42.5856	0.0019	1.0000	0.0004	17.3205	7.9225	-0.3339	-0.0218	0.8895	1.7660	‡0.0936	0.0297	0.2011
Unseen	Seq	0.0050	0.0500	0.0500	0.0312	0.0078	0.1236	17.1842	0.0051	1.0000	0.0213	8.3507	‡2.2440	-0.0748	0.1802	0.7142	1.7680	0.1906	0.0353	0.5098
Unseen	Cib	0.0040	0.0400	0.0400	0.0192	0.0047	0.0787	20.4439	0.0042	0.1172	0.0222	9.8979	3.7914	-0.1683	0.3999	1.0912	1.7239	0.1704	0.0282	0.3656

In fact, we present in Table 3 a first exploration on this idea. We note that at least one metric in every dimension (ARP, RSP, and AgeCal) obtains better results than when using the baseline prompts, so this evidences there is room for improvement and, in particular, that fairness improvements do not necessarily imply a loss in accuracy or diversity [22]. These results provide preliminary evidence that fairness-oriented prompts can influence several dimensions without severely degrading effectiveness [6, 18]. Although these improvements are not uniform across configurations, the results suggest that prompt engineering may represent a practical intervention mechanism for mitigating undesirable recommendation behaviors. A systematic exploration of prompt design, however, remains outside the scope of the present work and constitutes an important direction for future research.

Second, the choice of configuration parameters, such as item indexing strategies or prompt formulation, has a significant impact on both system performance and societal outcomes, reinforcing the need for systematic evaluation frameworks like the one proposed in this work.

Finally, these considerations point toward the importance of integrating multi-dimensional evaluation pipelines into the development lifecycle, enabling continuous monitoring of bias, fairness, and calibration alongside traditional effectiveness metrics when deploying LLM-based recommender systems in socially sensitive contexts.

Although the considered dimensions are measured offline, they capture mechanisms that have been associated with broader societal effects in prior work. Popularity bias and limited novelty may reduce exposure to diverse content, exposure imbalances can affect representation opportunities for different groups, and calibration influences users' ability to receive recommendations that reflect their interests. Consequently, these metrics should be interpreted as risk indicators rather than direct measurements of societal impact. Future work should validate these relationships through user studies and longitudinal analyses.

5.2 Limitations and future work

This work presents several limitations that open important avenues for future research, particularly in relation to risks and governance of LLM-based recommender systems. First, the proposed methodology is currently validated in a limited experimental setting, which may not fully capture the diversity of real-world scenarios or the dynamics of high-impact domains such as news recommendation, where biased or unbalanced exposure may contribute to misinformation, polarization, or societal harm [22]. Second, our findings highlight the inherent trade-offs between accuracy, novelty, fairness, and calibration, suggesting that improving one dimension may negatively affect others; managing these trade-offs remains

an open challenge, especially in applications where societal consequences are significant [14]. While recent work suggests that fairness can be influenced through prompt design, such interventions raise additional concerns regarding robustness, transparency, and the risk of unintended side effects, requiring careful governance and validation.

More broadly, the lack of standardized practices for auditing and controlling LLM-based recommenders points to the need for systematic monitoring frameworks and policy guidelines that ensure accountability and alignment with social good objectives [3, 11]. Future work will therefore extend the proposed methodology to more diverse domains (including news and social platforms), incorporate longitudinal evaluations, and explore mechanisms for integrating fairness-aware prompting and governance strategies into the design and deployment lifecycle of LLM-based recommender systems.

Besides, our evaluation primarily focuses on system-level metrics, without incorporating user studies to assess perceived fairness, trust, or behavioral impact. Additionally, the set of considered dimensions (e.g., fairness, calibration, and bias) is not exhaustive, and other relevant aspects such as long-term user well-being or environmental costs of LLM-based systems remain unexplored. Future work will extend the framework to additional datasets and domains, integrate user-centric evaluations, and broaden the set of impact dimensions to provide a more comprehensive assessment of the societal implications of LLM-driven recommender systems. In particular, future work will investigate the relationship between offline societal-impact metrics and observed user outcomes through controlled user studies, real-world deployment scenarios, and longitudinal analyses in domains such as news recommendation and online information consumption.

Acknowledgments

This work was supported by Grant PID2022-139131NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF, a way of making Europe".

References

- [1] Himan Abdollahpour, Robin Burke, and Bamshad Mobasher. 2017. Controlling Popularity Bias in Learning-to-Rank Recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys 2017, Como, Italy, August 27–31, 2017*, Paolo Cremonesi, Francesco Ricci, Shlomo Berkovsky, and Alexander Tuzhilin (Eds.). ACM, 42–46. doi:10.1145/3109859.3109912
- [2] Vito Walter Anelli, Alejandro Bellogin, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merri, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. 2021. Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11–15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 2405–2414. doi:10.1145/3404835.3463245

- [3] Ricardo Baeza-Yates and Usama M. Fayyad. 2024. Responsible AI: An Urgent Mandate. *IEEE Intell. Syst.* 39, 1 (2024), 12–17. doi:10.1109/MIS.2023.3343488
- [4] Christine Bauer, Eva Zangerle, and Alan Said. 2024. Exploring the Landscape of Recommender Systems Evaluation: Practices and Perspectives. *Trans. Recomm. Syst.* 2, 1 (2024), 11:1–11:31. doi:10.1145/3629170
- [5] Pablo Castells, Neil Hurley, and Saúl Vargas. 2022. Novelty and Diversity in Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, 603–646. doi:10.1007/978-1-0716-2197-4_16
- [6] Yian Chen, Lana Do, Liheng Yi, Ricardo Baeza-Yates, and John A. Guerra-Gomez. 2024. Advancing AI Incidents Classification: Leveraging LLMs with Strategic Prompting. In *Human-Computer Interaction - 10th Iberoamerican Conference, HCI-COLLAB 2024, Pereira, Colombia, June 4-7, 2024, Revised Selected Papers (Communications in Computer and Information Science)*, Vanessa Agredo Delgado, Pablo H. Ruiz, and Carlos Augusto Meneses Escobar (Eds.). Springer, 39–51. doi:10.1007/978-3-031-91328-0_4
- [7] Diego Corrêa da Silva and Dietmar Jannach. 2026. Calibrated Recommendations: Survey and Future Directions. *Trans. Recomm. Syst.* 4, 3 (2026), 43:1–43:32. doi:10.1145/3789266
- [8] Yashar Deldjoo. 2026. Understanding Biases in ChatGPT-based Recommender Systems: Provider Fairness, Temporal Stability, and Recency. *Trans. Recomm. Syst.* 4, 2 (2026), 17:1–17:35. doi:10.1145/3690655
- [9] Yashar Deldjoo, Vito Walter Anelli, Hamed Zamani, Alejandro Bellogín, and Tommaso Di Noia. 2021. A flexible framework for evaluating user and item fairness in recommender systems. *User Model. User Adapt. Interact.* 31, 3 (2021), 457–511. doi:10.1007/S11257-020-09285-1
- [10] Yashar Deldjoo, Zhankui He, Julian J. McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, René Vidal, Maheswaran Sathiamoorthy, Atoosa Kasirzadeh, and Silvia Milano. 2024. A Review of Modern Recommender Systems Using Generative Models (Gen-RecSys). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, Ricardo Baeza-Yates and Francesco Bonchi (Eds.). ACM, 6448–6458. doi:10.1145/3637528.3671474
- [11] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogín, Alessandro Difonzo, and Dario Zanzonelli. 2024. Fairness in recommender systems: research landscape and future directions. *User Model. User Adapt. Interact.* 34, 1 (2024), 59–108. doi:10.1007/S11257-023-09364-Z
- [12] Tomislav Duricic, Dominik Kowald, Emanuel Lacic, and Elisabeth Lex. 2024. Beyond-accuracy: a review on diversity, serendipity, and fairness in recommender systems based on graph neural networks. *Frontiers Big Data* 6 (2024). doi:10.3389/FDATA.2023.1251072
- [13] Miriam Fernández, Alejandro Bellogín, and Iván Cantador. 2024. Analysing the Effect of Recommendation Algorithms on the Spread of Misinformation. In *Proceedings of the 16th ACM Web Science Conference, WEBSCI 2024, Stuttgart, Germany, May 21-24, 2024*, Luca Maria Aiello, Yelena Mejova, Oshani Seneviratne, Jun Sun, Sierra Kaiser, and Steffen Staab (Eds.). ACM, 159–169. doi:10.1145/3614419.3644003
- [14] Héctor G.-de-Alba, Andres Tellez, José Emmanuel Gómez-Rocha, Cipriano A. Santos, Juan Antonio Orozco, and Ricardo Baeza-Yates. 2026. Towards a trade-off of interpretability, accuracy and scalability: Enhanced formulations in linear classification models. *Comput. Oper. Res.* 189 (2026), 107411. doi:10.1016/J.COR.2026.107411
- [15] Asela Gunawardana, Guy Shani, and Sivan Yogev. 2022. Evaluating Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, 547–601. doi:10.1007/978-1-0716-2197-4_15
- [16] Vincenzo Paparella, Dario Di Palma, Vito Walter Anelli, and Tommaso Di Noia. 2023. Broadening the Scope: Evaluating the Potential of Recommender Systems beyond prioritizing Accuracy. In *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023*, Jie Zhang, Li Chen, Shlomo Berkovsky, Min Zhang, Tommaso Di Noia, Justin Basilico, Luiz Pizzato, and Yang Song (Eds.). ACM, 1139–1145. doi:10.1145/3604915.3610649
- [17] Royal Pathak, Francesca Spezzano, and Maria Soledad Pera. 2023. Understanding the Contribution of Recommendation Algorithms on Misinformation Recommendation and Misinformation Dissemination on Social Networks. *ACM Trans. Web* 17, 4 (2023), 35:1–35:26. doi:10.1145/3616088
- [18] Mihaela Rotar, Theresia Veronika Rampisela, and Maria Maistro. 2026. Can Fairness Be Prompted? Prompt-Based Debiasing Strategies in High-Stakes Recommendations. *CoRR* abs/2603.12935 (2026). arXiv:2603.12935 doi:10.48550/ARXIV.2603.12935
- [19] Alan Said. 2024. On explaining recommendations with Large Language Models: a review. *Frontiers Big Data* 7 (2024). doi:10.3389/FDATA.2024.1505284
- [20] Alan Said and Alejandro Bellogín. 2014. Comparative recommender system evaluation: benchmarking recommendation frameworks. In *Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA - October 06 - 10, 2014*, Alfred Kobsa, Michelle X. Zhou, Martin Ester, and Yehuda Koren (Eds.). ACM, 129–136. doi:10.1145/2645710.2645746
- [21] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan (Eds.). ACM, 154–162. doi:10.1145/3240323.3240372
- [22] Céline Treuillier, Sylvain Castagnos, Evan Dufraisie, Özlem Özgöbek, and Armelle Brun. 2026. Shift your Focus for the Greater Good: Improving Fairness at no cost for Accuracy and Diversity in News Recommender Systems. *Trans. Recomm. Syst.* 4, 3 (2026), 44:1–44:36. doi:10.1145/3790097
- [23] Virginia Tsintzou, Evangelia Pitoura, and Panayiotis Tsaparas. 2019. Bias Disparity in Recommendation Systems. In *Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019), Copenhagen, Denmark, September 20, 2019 (CEUR Workshop Proceedings, Vol. 2440)*, Robin Burke, Himan Abdollahpour, Edward C. Malthouse, K. P. Thai, and Yongfeng Zhang (Eds.). CEUR-WS.org. <https://ceur-ws.org/Vol-2440/short4.pdf>
- [24] Shuyuan Xu, Wenyue Hua, and Yongfeng Zhang. 2024. OpenP5: An Open-Source Platform for Developing, Training, and Evaluating LLM-based Recommender Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 386–394. doi:10.1145/3626772.3657883
- [25] Hongzhi Yin, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. 2012. Challenging the Long Tail Recommendation. *Proc. VLDB Endow.* 5, 9 (2012), 896–907. doi:10.14778/2311906.2311916
- [26] Ziwei Zhu, Jianling Wang, and James Caverlee. 2020. Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 449–458. doi:10.1145/3397271.3401177