Leveraging LLMs for Explainable Recommender Systems: Exploring User Perceptions and Faithfulness in Generated Explanations

Ítallo de Sousa Silva, Leandro Balby Marinho (advisor)

Abstract. *Recommender systems (RSs) are widely used to support users in discovering relevant items, yet their growing complexity raises concerns about transparency and interpretability. This work investigates the use of Large Language Models (LLMs) to generate automated, human-centered explanations for RSs and evaluates whether such explanations are faithful to the models' underlying decision processes. First, we conducted a user study assessing personalized movie recommendations and explanations generated by GPT-3.5 Turbo, measuring perceived effectiveness, personalization, and persuasiveness. Then, we evaluated the faithfulness of explanations across movie, music, and book domains using four LLMs. Results show that LLM-based recommendations increase user satisfaction compared to random baselines; however, explanations frequently fail to align with the factors driving the recommendations. Moreover, user-based explanations were not consistently perceived as superior to generic ones.*

Resumo. *Sistemas de Recomendação (SRs) são amplamente utilizados para auxiliar usuários na descoberta de itens relevantes, mas sua complexidade levanta preocupações quanto à transparência e interpretabilidade. Este trabalho investiga o uso de Modelos de Linguagem de Grande Escala (LLMs) para gerar explicações automatizadas e centradas no usuário em SRs, avaliando também se tais explicações são fiéis ao processo decisório dos modelos. Inicialmente, conduzimos um estudo com usuários para avaliar recomendações e explicações personalizadas de filmes geradas pelo GPT-3.5 Turbo, mensurando percepção de efetividade, personalização e persuasão. Em seguida, analisamos a fidelidade das explicações nos domínios de filmes, músicas e livros utilizando quatro LLMs (GPT-4o, Llama3, Gemma2 e Mixtral 8x7B). Os resultados indicam que recomendações geradas por LLMs aumentam a satisfação dos usuários em comparação a recomendações aleatórias; entretanto, as explicações frequentemente não refletem adequadamente os fatores que influenciaram a recomendação. Além disso, explicações baseadas em preferências do usuário não foram consistentemente percebidas como superiores às genéricas.*

1. Introduction

Recommender Systems (RSs) are central to modern digital platforms, shaping how users discover products, media, and information. As these systems grow more complex—often relying on opaque deep learning models—concerns about transparency, interpretability, and trust have intensified. Explainable Recommender Systems (XRSs) seek to address these issues by providing understandable justifications for recommendations [Zhang and Chen 2020].

Traditional explanation methods—feature-based, similarity-based, or template-generated—often produce rigid and impersonal outputs. Although human-generated explanations improve user perception, they are costly and not scalable [Balog and Radlinski 2020, Chang et al. 2016, Lu et al. 2023]. Large Language Models (LLMs) offer a scalable alternative, generating personalized recommendations and natural language explanations from minimal user input [Wu et al. 2023, Lyu et al. 2023]. Although these capabilities are highly promising, important research questions remain. It is unclear whether users truly find LLM-generated explanations more effective or persuasive than generic ones, and whether these explanations are faithful to the model's actual reasoning (unfaithful explanations can mislead users, erode trust, and pose ethical and legal risks).

Therefore, this work addresses the central challenge of determining whether Large Language Models can be reliably leveraged to generate scalable, human-centered, and faithful explanations for recommender systems, while also examining how users perceive these explanations. We investigated LLM-based Recommender Systems (LLM-Recs) from two complementary perspectives (***User-Centered Evaluation*** and an ***Empirical Faithfulness Evaluation***) using six research questions:

- **RQ1:** How do users value personalized LLM-generated recommendations compared to random recommendations?
- **RQ2:** How do users perceive user-based versus generic LLM-generated explanations in relation to recommendation methods and explanation goals such as effectiveness, personalization, and persuasiveness?
- **RQ3:** Do user-based versus generic explanations work differently for familiar or unfamiliar movies?
- **RQ4:** How do explanation goals such as personalization, persuasiveness, and effectiveness relate?
- **RQ5:** How do user-specific explanations compare to generic explanations regarding the types of argumentation they employ?
- **RQ6:** Are the explanations provided by LLM-based Recommender Systems faithful?

Our findings revealed a nuanced landscape: although LLM-generated recommendations increase user satisfaction compared to random baselines, personalized explanations are not consistently perceived as more personalized or persuasive than generic ones. Moreover, faithfulness analysis demonstrates that LLM-generated explanations frequently fail to reflect the actual importance of user preferences, with significant variability across models and domains.

2. Related Work

Explainable Recommender Systems (XRSs) have become central to research in both Recommender Systems and Explainable AI. Various explanation strategies have been proposed, including similarity-based rationales ("because you liked X"), feature-importance methods, review aggregation, and hybrid approaches combining content and collaborative signals [Zhang and Chen 2020].

Early work highlighted the importance of personalization. [Chang et al. 2016] showed that personalized natural-language explanations outperformed tag-based ones in

perceived usefulness, and [Lu et al. 2023] found that users preferred human-generated explanations over template-based machine explanations. However, such approaches face scalability challenges, as producing high-quality personalized explanations is costly. Large Language Models (LLMs) offer a scalable alternative. Models such as GPT-4, Llama, and Gemma demonstrate strong zero- and few-shot reasoning and have been explored either as standalone recommenders or integrated into traditional pipelines, sometimes combined with knowledge graphs or structured signals [Lyu et al. 2023].

Explanation evaluation has been largely guided by the framework of [Tintarev and Masthoff 2007], which defines seven goals, including effectiveness, transparency, and trust. In particular, [Balog and Radlinski 2020] reported that explanations optimized for different goals were often perceived similarly, and [Zhou and Joachims 2023] found no significant perceptual differences between LLM- and human-generated reviews, except among users familiar with the items.

Other important aspect of explanation is *faithfulness*, which refers to the extent to which an explanation accurately represents a model's internal reasoning [Jacovi and Goldberg 2020]. Although LLMs can generate self-explanations, these are often plausible rather than faithful [Turpin et al. 2024]. Recent work distinguishes fine-tuning- and prompting-based explainability methods [Zhao et al. 2024]. Within prompting approaches, perturbation-based techniques are common but do not ensure causal alignment. [Parcalabescu and Frank 2024] argue that many faithfulness tests capture self-consistency rather than true alignment and propose CC-SHAP, which compares Shapley-based importance distributions between predictions and explanations. However, this method requires access to output probabilities, limiting its applicability to closed-source models.

3. Methods

3.1. Human-Centered Evaluation of LLM Explanations

To collect user preferences and evaluate recommendations and explanations, we developed a web application[1] . The study followed three stages: (i) preference elicitation, (ii) recommendation and explanation generation, and (iii) user evaluation via questionnaire. Participants provided three liked and three disliked movies using a search interface powered by the OMDb API[2] . To avoid knowledge mismatches with GPT-3.5 Turbo (knowledge cutoff: September 2021), only movies released before 2021 were available.

Following, each participant received six items: four recommendations and two disrecommendations. Including negative recommendations allowed us to test differences in user perception by recommendation intent and to assess the value of avoiding unsatisfying items. We compared two recommenders: (i) an LLM-based system using GPT-3.5 Turbo (snapshot *gpt-3.5-turbo-0613*), the most capable publicly available OpenAI's models at the time of the study, via OpenAI's Chat API, and (ii) a Random Recommender sampling popular films from IMDb[3]. The random baseline mitigates familiarity bias and helps isolate the effect of explanations. A zero-shot prompting strategy was adopted to (i)

[1]https://github.com/issilva5/llm-recommendations-survey/tree/main
[2]https://www.omdbapi.com
[3]https://www.imdb.com/chart/moviemeter/

encode user preferences, (ii) generate recommendations (two must-watch, two avoid), and (iii) produce either user-based or generic explanations. For disrecommendations, prompts explicitly incorporated negative intent. All prompts are publicly available[4] . Temperature was set to 0 to enhance reproducibility.

Recommendations and explanations were presented individually, each followed by a questionnaire including a familiarity question and Likert-scale items assessing recommendation enjoyment and explanation effectiveness, personalization, and persuasiveness. To account for repeated measures (multiple responses per participant), we employed multilevel mixed linear regression with random intercepts, capturing within-subject correlation and individual baseline differences.

3.2. Faithfulness in Black-Box LLM-based Recommender Systems

To assess whether LLM-generated explanations faithfully reflect recommendation reasoning, we propose a model-agnostic axiomatic framework grounded in feature importance agreement [Wiegreffe et al. 2021]. Let U be a set of users and I a set of items. A recommender $\mathcal{R}$ maps a user $u \in U$ to a set of recommended items $R_u \subseteq I$. In our case, the recommender is an LLM $\mathcal{L}$ that receives user preferences I_u and outputs both recommendations and self-generated explanations.

Obtaining user's preferences importance for the recommendation Since models are accessed via APIs, they are treated as black boxes; our method therefore relies exclusively on input perturbations and observable outputs. Following the Feature Importance Agreement principle [Wiegreffe et al. 2021], we define faithfulness as: *preferences that are important for a recommendation should be reflected proportionally in the explanation*. To operationalize this, we compute the importance of each input item using Shapley Values. Unlike SHAP[5] approximations or CC-SHAP [Parcalabescu and Frank 2024], we enumerate all possible coalitions to obtain exact Shapley Values, ensuring fairness and additivity guarantees.

For each recommendation $r \in R_u^{\mathcal{L}}$, we define a coalitional game $G_r(I_u, f_r)$ where: $f_r(S) : 2^{I_u} \rightarrow \{0, 1\} = 1$ if $r \in \mathcal{L}(S)$ and 0 otherwise. The Shapley Value of item $i \in I_u$ with respect to recommendation r is:

$$\varphi_i^r = \frac{1}{n} \left(\sum_{S \subseteq I_u \setminus \{i\}} \frac{1}{C_{n-1}^{|S|}} [f_r(S \cup \{i\}) - f_r(S)] \right)$$

This value quantifies the marginal contribution of preference i to the recommendation r across all coalitions.

Obtaining user's preferences importance for the explanation Next, we estimate the importance of preferences within the explanation. Since explanations are textual, we adopt two complementary metrics:

[4]https://github.com/issilva5/llm-recommendations-survey/blob/main/prompts.png

[5]https://shap.readthedocs.io/en/latest/

Citation. An item is considered important if it is explicitly mentioned in the explanation E_r, therefore $c(i, E_r) = 1$ if $i \in E_r$ and 0 otherwise.

LLM-Ranking. Because preferences may be implicitly referenced, we prompt GPT-4o to rank input items according to their importance in the explanation text. This produces a ranking R_{E_r} for each $i \in I_u$. The judge is blinded to the identity of the model that generated each explanation and receives only the explanation text and corresponding input items. Therefore, GPT-4o does not know whether an explanation was produced by itself or by another evaluated model. Nevertheless, because GPT-4o is also one of the evaluated models, we cannot fully rule out model-family or stylistic bias in the ranking procedure, and we acknowledge this as a limitation.

Faithfulness evaluation To compare recommendation importance and explanation importance, we compute the following metrics:

Recall & Precision. Recall measures how many important items are mentioned in the explanation, while Precision measures whether cited items are indeed important for the recommendation. We operationally define an item as important when its Shapley Value exceeds 0.1, thereby excluding items with only small positive contributions to the recommendation. Let Φ_r^+ be the set of items satisfying this criterion and C^r the set of cited items in E_r. Therefore, $Recall = \frac{|C^r \cap \Phi_r^+|}{|\Phi_r^+|}$ and $Precision = \frac{|C^r \cap \Phi_r^+|}{|C^r|}$. The 0.1 threshold is an operational choice rather than an established theoretical cutoff, and the potential sensitivity of the results to this choice is a limitation of our evaluation.

Weighted Coverage (WCov). Let $R_{E_r}^k$ be the top-k ranked items from LLM-Ranking. This metric captures how much recommendation importance is covered by the top-ranked items in the explanation, weighting higher-ranked items more strongly. We define: $WCov_r@K = \sum_{n=1}^{K} \frac{1}{n} \varphi^r_{R^n_{E_r}}$.

Datasets & Models To ensure robustness across domains, we evaluated four LLMs — GPT-4o, Gemma 2 9B, Llama 3 70B, and Mixtral 8x7B — on three recommendation tasks: movies, music playlist completion, and book recommendation. The datasets include: **IUI 2024 Movie Preferences**, containing three liked and three disliked movies from 94 users and collected in our first experiment; **Spotify Million Playlist Dataset (SMPD)**[6], from which 100 playlists were sampled; **Book Recommendation Dataset**, obtained from Book-Crossing via Kaggle[7].

For each task, identical prompts were used across models. Each model generated recommendations and corresponding explanations for every user instance. In the movie and book tasks, models received three liked and three disliked items and produced two positive and two negative recommendations with explanations. In the playlist completion task, models were given the playlist title and first five tracks (name and artist) and asked to suggest two tracks to include and two to avoid, also with explanations. All prompts are publicly available[8]. All models were accessed via APIs and treated as black-box systems.

[6]https://www.kaggle.com/datasets/himanshuwagh/spotify-million

[7]https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset

[8]https://github.com/issilva5/llm-explanations-faithfulness/blob/main/src/prompts/__init__.py

4. Results & Discussion

Table 1 summarizes the publications resulting from this dissertation, while Appendix A provides further details on each publication.

Table 1. Summary of publication characteristics

Event/Journal	AR[a]/IF[b]	Year	Qualis[c]	Downloads	Citations[d]
[Silva et al. 2023]	-	2023	-	-	4
[Silva et al. 2024]	28%	2024	A1	2803	64
[Silva et al. 2025]	6.4	2025	A2	272	1
[Silva and Marinho 2026][e]	-	2026	A4	-	-

[a] Acceptance rate, [b] Impact Factor, [c] Qualis 2020, [d] Google Scholar,[e]Accepted for BRACIS'26

The first publication, [Silva et al. 2023], was a workshop paper in which the initial concept of this research was presented. This work was later expanded into the full paper [Silva et al. 2024], whose findings were subsequently complemented by an extended journal version [Silva et al. 2025]. The results associated with these publications are discussed in Section 4.1.

More recently, [Silva and Marinho 2026] was accepted for publication at BRACIS'26. This paper presents, in greater detail, the faithfulness analysis summarized and discussed in Section 4.2.

4.1. Human-Centered Evaluation of LLM Explanations

Recommendation Satisfaction (RQ1) We examined enjoyment ratings as a function of recommendation method (LLM-based vs. random) and movie familiarity. Overall, 25% of LLM recommendations and 49% of random ones were unfamiliar. Random recommendations were rated significantly lower than GPT-3.5-based personalized recommendations ($\beta = -0.53, p < 0.001$), and unfamiliar movies received lower ratings than familiar ones ($\beta = -0.78, p < 0.001$). No interaction effect emerged, indicating that LLM recommendations are preferred regardless of prior familiarity (see Fig. 1a).

User-Based vs. Generic Explanations (RQ2) We compared user-based and generic explanations across effectiveness, personalization, and persuasiveness. Explanations from the GPT-based recommender were rated as more effective than those for random recommendations ($\beta = 0.37, p < 0.001$), but no significant difference emerged between user-based and generic explanations in effectiveness ($\beta = -0.10, p = 0.37$; see Fig. 1b). User-based explanations were not consistently perceived as more personalized than generic ones, except relative to random recommendations, which scored much lower on personalization ($\beta = -0.80, p < 0.001$). A significant interaction between recommendation and explanation type ($\beta = 0.49, p < 0.01$) indicates that personalization effects are context-dependent (Fig. 1c). For persuasiveness, explanations accompanying random recommendations were rated as less convincing ($\beta = -0.47, p < 0.001$). Although personalized LLM explanations appeared slightly less persuasive than generic ones, the difference was not statistically significant ($\beta = -0.27, p = 0.14$; Fig. 1d).

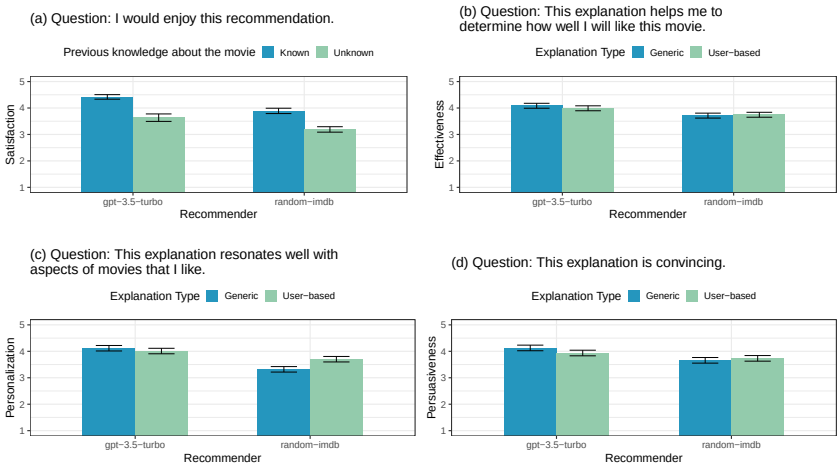


Figure 1. Questionnaire results for the recommendations based on estimated means from the random intercept multilevel regressions. Responses were given on a 5-point disagree–agree scale (3 = neutral). Error bars represent one standard error.

Effect of Movie Familiarity (RQ3) Explanations for unfamiliar movies were rated as less effective overall ($\beta = -0.41, p < 0.01$), mainly for generic explanations. A significant interaction ($\beta = 0.35, p < 0.05$) indicates that user-based explanations mitigated this decline. Personalization ratings were also lower for unfamiliar movies ($\beta = -0.53, p < 0.001$), though the interaction with explanation type was not significant. Persuasiveness decreased for unfamiliar items as well ($\beta = -0.58, p < 0.05$), with a significant main effect of explanation type ($\beta = -0.28, p < 0.05$) and an interaction with familiarity ($\beta = 0.50, p < 0.05$). Overall, these results suggest that personalization is particularly valuable when users lack prior familiarity with the item.

Path Modeling of Explanation Goals (RQ4) To examine relationships among explanation goals, we fitted a path model following the user-centric framework of [Knijnenburg et al. 2012] (see Figure 2). Personalization and persuasiveness were modeled as Subjective System Aspects (SSA), while satisfaction and effectiveness were treated as experiential constructs (EXP). Consistent with [Balog and Radlinski 2020], we found strong correlations among personalization, persuasiveness, satisfaction, and effectiveness. Effectiveness was predicted by satisfaction, personalization, and especially persuasiveness—the strongest predictor. Satisfaction increased with familiarity and LLM-based recommendations and was directly influenced by both personalization and persuasiveness. Overall, the model clarifies that persuasive explanations drive both satisfaction and perceived effectiveness.

Disrecommendations Generic explanations were perceived as slightly more effective than personalized ones ($\beta = -.32, p < .05$), and personalized explanations were somewhat less persuasive ($\beta = -.36, p < .05$). No significant personalization effects were observed. These results suggest that personalization may not benefit negative recommendations, possibly reflecting model constraints related to safety or reluctance to produce

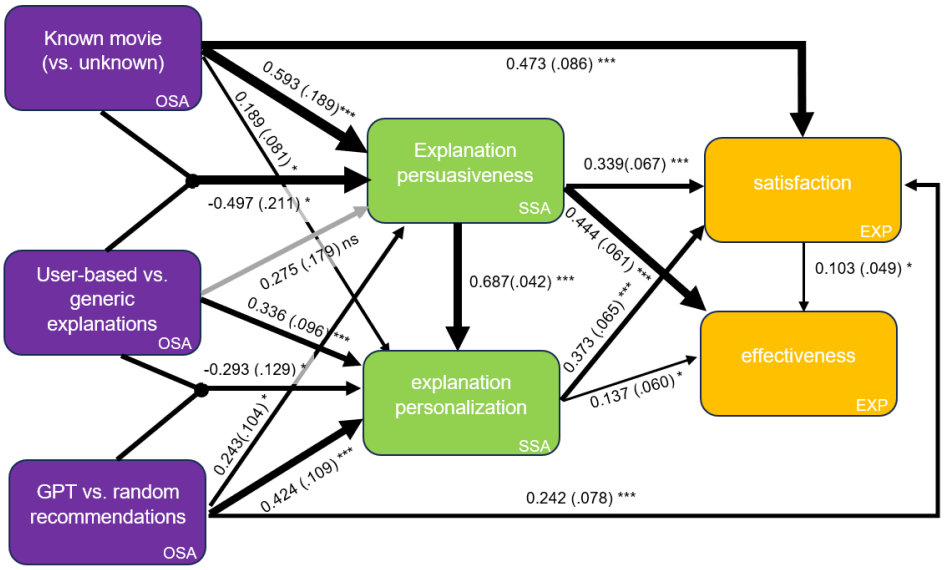


Figure 2. Path model showing how persuasiveness and personalization are influenced by objective system aspects (OSA) and how they predict satisfaction and effectiveness. Line thickness represents coefficient strength. Standard errors in brackets. Significance: * $p < .05$, ** $p < .01$, *** $p < .001$.

strong negative judgments[9].

The Anatomy of Explanations (RQ5) To understand why generic explanations performed comparably to personalized ones, we analyzed their argumentative structure using GPT-4o-based classification grounded in the Periodic Table of Arguments (PTA) [Wagemans 2021]. Sentences were categorized into 14 argument types plus non-argumentative statements. *Argument from Evaluation* dominated both explanation types, especially generic ones. User-based explanations displayed more varied structures, including *similarity and comparison arguments*.User-based explanations exhibited more diverse argument sequences. Markov-chain modeling further highlighted structural differences: user-based explanations exhibited richer transition patterns, while generic explanations followed more linear evaluative flows. These structural findings help explain why generic explanations can achieve comparable persuasive impact: they rely on clear evaluative rhetoric and second-person engagement, even without explicit personalization.

4.2. Faithfulness in Black-Box LLM-based Recommender Systems

This section reports the results of the axiomatic faithfulness evaluation for black-box LLM-based recommender systems **(RQ6)**, according to the methods described in Section 3.2.

Recall and Precision A table available on Appendix B reports the performance of four models across three datasets (Book, IUI24, SMPD) using Average Precision (AP), Average Recall (AR), and F1-score for positive and negative recommendations. The results reveal systematic differences across recommendation type, dataset, and model.

[9]https://openai.com/policies/usage-policies

Negative recommendations yield the lowest performance across metrics, models, and domains. This mirrors our human-centered findings and may reflect LLM reluctance to engage in negative discourse, leading to fewer disliked items being cited in explanations. Substantial variation also emerges across datasets, indicating domain sensitivity in explanation faithfulness. The Book dataset shows the weakest performance, with uniformly low F1-scores, suggesting greater susceptibility to unfaithful explanations. In contrast, IUI24 achieves the strongest results, while SMPD occupies an intermediate position. These differences may be related to uneven domain knowledge acquired during pre-training, although dedicated experiments would be required to confirm this hypothesis.

Model-level differences are likewise evident. GPT-4o achieves the highest F1-scores across datasets and recommendation types, followed by Mixtral 8x7B. Llama3 70B and Gemma2 9B perform competitively in some settings but remain overall below the top two models, highlighting variation in how LLMs translate user preferences into explanatory content.

Overall, AP consistently exceeds AR, suggesting that explanations mention important items but fail to comprehensively cover all contributing preferences. Moreover, citation-based Precision and Recall ignore relative importance and implicit references, motivating the use of complementary metrics such as Weighted Coverage that account for ranked importance and contextual attribution.

Weighted Coverage Figure 3 presents the distribution of Weighted Coverage at Top-3 (WCov@3) for positive (TRUE) and negative (FALSE) recommendations across models and datasets. The dashed line at 0.5 is included only as a visual reference, corresponding to explanations that cover approximately half of the recommendation importance captured by the top-3 features. We do not interpret 0.5 as a validated cutoff between faithful and unfaithful explanations. Since WCov@3 is a novel metric in recommendation settings, no established baseline or interpretive threshold is available in the literature; therefore, the results are primarily interpreted comparatively across models, datasets, and recommendation types. Alternative model-agnostic faithfulness approaches, such as counterfactual perturbations, are difficult to apply in this setting because constructing valid counterfactual items (e.g., movies) would require additional task-specific models. Input reconstruction methods are also unsuitable, since recommendations often reuse cited items and may therefore artificially inflate apparent alignment.

The distributions show clear differences between positive and negative recommendations, with positive cases generally achieving higher WCov@3. The Book dataset is an exception, where negative recommendations sometimes appear comparatively stronger, suggesting domain-specific differences in how preference signals are reflected in explanations. Overall, the lower WCov@3 observed for negative recommendations may reflect a reduced tendency of LLMs to explicitly refer to disliked inputs when generating negative recommendations.

Dataset-level patterns mirror the earlier findings. IUI24 achieves the highest median WCov@3 across most models, whereas the Book dataset exhibits consistently lower values, with at least 75% of explanations falling below the 0.5 reference level.

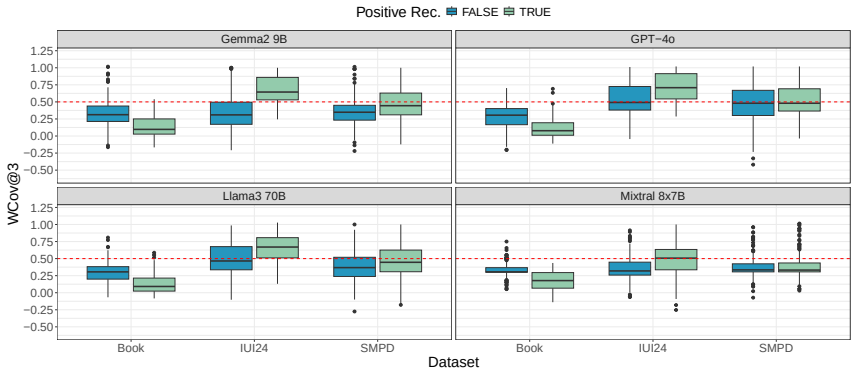


Figure 3. Distribution of the Weighted Coverage at Top-3 for the four models and three datasets.

SMPD shows intermediate but more variable performance, indicating less stable agreement between recommendation importance and explanatory emphasis in playlist completion. These cross-domain differences may be related to differences in domain-specific knowledge acquired during pre-training, but the present experiments do not isolate this factor, and additional controlled experiments would be required to establish such an explanation. At the model level, GPT-4o attains the highest median WCov@3 for positive recommendations across datasets, particularly on IUI24. Notably, Llama3 70B and Gemma2 9B often match or surpass Mixtral 8x7B in WCov@3, despite Mixtral ranking second in citation-based metrics. This suggests that, although Mixtral frequently cites recommendation-relevant items, the relative emphasis assigned to them in its explanations does not consistently align with their estimated importance.

Overview Taken together, these results suggest that (i) faithfulness is sensitive to domain and recommendation polarity, (ii) citation-based metrics can overestimate alignment when explanation emphasis does not match true input contribution, and (iii) Weighted Coverage helps reveal mismatches between what explanations *say* and what inputs actually *drive* the recommendation.

5. Conclusion

This work presented one of the earliest evaluations combining user perception analysis with axiomatic faithfulness assessment in LLM-based recommender systems, examining their potential and limitations from these two complementary perspectives. From a human-centered viewpoint, LLM-based recommendations improved user satisfaction over random baselines, with user-based explanations proving especially helpful for unfamiliar items. Path modeling showed that persuasiveness and personalization strongly predict perceived effectiveness. From a faithfulness perspective, using a Shapley-based *Feature Importance Agreement* framework, we found only partial alignment between the preferences driving recommendations and those highlighted in explanations. No model consistently produced fully faithful explanations, revealing a gap between persuasive and faithful explanations. Key limitations included LLM output variability, limited personalization due to constrained user input, domain sensitivity, and the black-box nature of API-based models, which restricted the range of applicable faithfulness tests.

References

[Balog and Radlinski 2020] Balog, K. and Radlinski, F. (2020). Measuring recommendation explanation quality: The conflicting goals of explanations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 329–338, New York, NY, USA. Association for Computing Machinery.

[Chang et al. 2016] Chang, S., Harper, F. M., and Terveen, L. G. (2016). Crowd-based personalized natural language explanations for recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 175–182, New York, NY, USA. Association for Computing Machinery.

[Jacovi and Goldberg 2020] Jacovi, A. and Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

[Knijnenburg et al. 2012] Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., and Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22:441–504.

[Lu et al. 2023] Lu, H., Ma, W., Wang, Y., Zhang, M., Wang, X., Liu, Y., Chua, T.-S., and Ma, S. (2023). User perception of recommendation explanation: Are your explanations what users need? *ACM Trans. Inf. Syst.*, 41(2).

[Lyu et al. 2023] Lyu, H., Jiang, S., Zeng, H., Wang, Q., Zhang, S., Chen, R., Leung, C., Tang, J., Xia, Y., and Luo, J. (2023). Llm-rec: Personalized recommendation via prompting large language models.

[Parcalabescu and Frank 2024] Parcalabescu, L. and Frank, A. (2024). On measuring faithfulness or self-consistency of natural language explanations. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089, Bangkok, Thailand. Association for Computational Linguistics.

[Silva and Marinho 2026] Silva, I. and Marinho, L. (2026). Faithfulness in black-box llm-based recommender systems' explanations. In *Anais da XXXVI Brazilian Conference on Intelligent Systems*. Accepted for publication.

[Silva et al. 2024] Silva, I., Marinho, L., Said, A., and Willemsen, M. C. (2024). Leveraging chatgpt for automated human-centered explanations in recommender systems. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, page 597–608, New York, NY, USA. ACM.

[Silva et al. 2023] Silva, I., Said, A., and Marinho, L. B. (2023). Leveraging large language models for recommendation and explanation. In *Proceedings of the 10th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS 2023)*. CEUR Workshop Proceedings.

[Silva et al. 2025] Silva, I. D. S., Balby, L., Said, A., and Willemsen, M. (2025). User perceptions of personalized and generic explanations in llm-driven recommender systems. *ACM Trans. Interact. Intell. Syst.* Just Accepted.

[Tintarev and Masthoff 2007] Tintarev, N. and Masthoff, J. (2007). A survey of explanations in recommender systems. In *2007 IEEE 23rd International Conference on Data Engineering Workshop*, pages 801–810.

[Turpin et al. 2024] Turpin, M., Michael, J., Perez, E., and Bowman, S. R. (2024). Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

[Wagemans 2021] Wagemans, J. (2021). Periodic table of arguments 2.5 a.

[Wiegreffe et al. 2021] Wiegreffe, S., Marasović, A., and Smith, N. A. (2021). Measuring association between labels and free-text rationales. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Wu et al. 2023] Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, H., Shen, T., Qin, C., Zhu, C., Zhu, H., Liu, Q., Xiong, H., and Chen, E. (2023). A survey on large language models for recommendation. *CoRR*, abs/2305.19860.

[Zhang and Chen 2020] Zhang, Y. and Chen, X. (2020). Explainable recommendation: A survey and new perspectives. *Found. Trends Inf. Retr.*, 14(1):1–101.

[Zhao et al. 2024] Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., and Du, M. (2024). Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.* Just Accepted.

[Zhou and Joachims 2023] Zhou, J. and Joachims, T. (2023). Gpt as a baseline for recommendation explanation texts. In *Proceedings of the IntRS'23 Workshop*.

A. Appendix: Dissertation by-products

The dissertation by-products are listed as follows:

1. **Software packages.** All source code and data used in the dissertation and related papers are distributed as open source:
   - `https://github.com/issilva5/llm-recommendations-survey`
   - `https://github.com/issilva5/llm-explanations-faithfulness`
2. **Workshop Publication - CEUR Workshop Proceedings:**
   [Silva et al. 2023]
   This work presented the first demonstration of the research concept investigated in this dissertation: the use of large language models to generate recommendations and natural-language explanations from user-stated preferences. The paper describes a proof-of-concept Web application in which ChatGPT is used both as a movie recommender and as a generator of personalized explanations. This publication served as the starting point for the line of investigation later developed in the empirical studies reported in this dissertation.
   - Event year: 2023; *Joint Workshop on Interfaces and Human Decision Making for Recommender Systems* (IntRS'23), co-located with ACM RecSys 2023.

- Work citations: 4 citations, according to Google Scholar;
  Work downloads: not available.
- Link: https://ceur-ws.org/Vol-3534/paper6.pdf

3. **Conference Publication - Acceptance Rate 28%, Qualis A1:**
   [Silva et al. 2024]
   This work expanded the idea initially presented in the workshop paper by conducting a user study to evaluate recommendations and explanations generated by ChatGPT in a movie recommendation scenario. The paper investigated whether users prefer ChatGPT-generated recommendations over random popular recommendations and evaluated users' perceptions of personalized and generic explanations in terms of effectiveness, personalization, and persuasiveness. The results indicated that users tend to prefer ChatGPT-generated recommendations and that personalized explanations can be more effective when the recommended item is unfamiliar to the user.
   - Event year: 2024; *Proceedings of the 29th International Conference on Intelligent User Interfaces* (IUI'24), https://iui.acm.org/2024/.
   - Work citations: 64 citations, according to Google Scholar;
     Work downloads: 2803 downloads.
   - Link: https://doi.org/10.1145/3640543.3645171
4. **Journal Publication - JCR 6.4, Qualis A2:**
   [Silva et al. 2025]
   This work corresponds to the extended version of the study published at IUI'24. The article deepens the analysis of users' perceptions of personalized and generic explanations produced by LLMs in recommender systems. In particular, it complements the conference version by further investigating the factors that affect the perceived usefulness of explanations, as well as differences between the types of arguments used in the model-generated justifications. Thus, the article consolidates the dissertation contribution related to the user-centered evaluation of LLM-generated recommendations and explanations.
   - Publication year: 2025; *ACM Transactions on Interactive Intelligent Systems* (TiiS), https://dl.acm.org/journal/tiis.
   - Article citations: 1 citation, according to Google Scholar;
     Article downloads: 272 downloads.
   - Link: https://doi.org/10.1145/3779059
5. **Conference Publication - Qualis A4:**
   [Silva and Marinho 2026]
   This work, accepted for publication at BRACIS'26, presents in detail the second main contribution of the dissertation: the evaluation of the faithfulness of LLM-generated explanations in black-box LLM-based recommender systems. The paper proposes an axiomatic protocol inspired by feature-importance agreement, according to which a preference that is important for a recommendation should also appear with comparable importance in the corresponding explanation. To operationalize this idea without access to the model's internal parameters, the work estimates the importance of input preferences using exact Shapley values and

compares these values with signals extracted from the explanations, including explicit citations and LLM-based rankings. The evaluation considers four language models, GPT-4o, Gemma 2 9B, Llama 3 70B, and Mixtral 8x7B, across three domains: movies, music playlists, and books. The results show that LLM-generated explanations are not consistently faithful to the preferences that most influence the recommendations, highlighting the need for caution when using such explanations to justify or audit personalized recommendations.

- Event year: 2026; *Brazilian Conference on Intelligent Systems* (BRACIS'26).
- Work citations: not available;
  Work downloads: not available.
- Link: accepted for publication. DOI not yet available.

B. Appendix: Average Precision and Recall for RQ6 evaluation

Table 2. Average Precision and Recall by dataset, model and recommendation type. AR denotes Average Recall, AP denotes Average Precision, and F1 denotes the harmonic mean of AR and AP. Best results within each dataset, recommendation type, and metric are shown in bold, and second-best results are underlined.

		Positive Recs.			Negative Recs.		
Model	**Dataset**	**AR**	**AP**	**F1**	**AR**	**AP**	**F1**
Book	GPT-4o	0.192	0.275	0.226	**0.380**	0.569	**0.456**
	Gemma 2 9B	0.132	0.238	0.170	0.282	0.455	0.349
	Llama 3 70B	0.198	0.284	0.233	0.298	0.507	0.376
	Mixtral 8x7B	**0.277**	**0.494**	**0.355**	0.309	**0.682**	0.426
IUI24	GPT-4o	**0.769**	**0.721**	**0.744**	**0.661**	**0.698**	**0.679**
	Gemma 2 9B	0.692	0.663	0.677	0.434	0.475	0.453
	Llama 3 70B	0.702	**0.721**	0.712	0.487	0.724	0.582
	Mixtral 8x7B	0.672	0.613	0.641	0.559	0.653	0.602
SMPD	GPT-4o	**0.660**	0.662	**0.661**	**0.571**	0.664	**0.614**
	Gemma 2 9B	0.393	0.577	0.468	0.147	0.313	0.200
	Llama 3 70B	0.639	0.625	0.632	0.287	0.507	0.366
	Mixtral 8x7B	0.559	**0.804**	0.659	0.431	**0.774**	0.554