Beyond Top-1: Addressing Inconsistencies in Evaluating Counterfactual Explanations for Recommender Systems (Extended Abstract)

Amir Reza Mohammadi[1] , **Andreas Peintner**[1] , **Michael M. Müller**[1] and **Eva Zangerle**[1]
[1]University of Innsbruck
{amir.reza, andreas.peintner, michael.m.mueller, eva.zangerle}@uibk.ac.at

Abstract

Counterfactual explanations have become an important paradigm for improving the transparency of machine learning models by showing how small input changes can change the model outputs. While substantial progress has been made in generating such explanations, their evaluation remains insufficiently standardized, particularly for systems that produce *ranked outputs* rather than single-label predictions. Existing evaluation protocols in recommender systems commonly focus only on whether the top-1 recommendation changes after perturbation. We argue that this practice can lead to inconsistent and misleading conclusions, as the relative ranking of explanation methods may vary with changes in the quality of the underlying model.

In this work, we revisit the evaluation of counterfactual explanations from a ranking perspective. We propose extending top-1 evaluation to list-wise top-k protocols that assess explanation effectiveness across multiple highly ranked outputs. Through experiments on multiple datasets, recommender architectures, and explanation methods, we show that top-k evaluation substantially improves consistency and yields more reliable comparisons between competing explainers. Our findings highlight a broader methodological lesson for explainable AI: when models return ranked results, explanation quality should be assessed using ranking-aware evaluation protocols rather than top-1 criteria alone.

1 Introduction

As machine learning systems are increasingly deployed in decision-support settings, the need for transparent and trustworthy model behavior has become central to modern AI research [Ghorbani *et al.*, 2019]. Among post-hoc explainability techniques, counterfactual explanations (CE) [Doshi-Velez and Kim, 2017] have emerged as a particularly intuitive paradigm. Rather than describing why a model produced a certain output, counterfactual explanations identify minimal changes to the input that would lead to a different outcome. This makes them especially attractive in domains where actionable feedback and user understanding are important [Karimi *et al.*, 2021].

Recent years have seen growing interest in applying counterfactual explanations to systems that produce *ranked outputs*, such as recommender systems, retrieval engines [Mozafari *et al.*, 2026], and personalized ranking models [Zhong and Negre, 2022; Yu *et al.*, 2025]. In these settings, the objective is typically not to predict a single class label, but to generate an ordered list of items according to estimated relevance. Explaining such systems therefore introduces distinct methodological challenges that differ from traditional classification settings. One of the most fundamental open challenges concerns *evaluation*. While many methods have been proposed to generate counterfactual explanations, considerably less attention has been paid to how they should be compared reliably [Bauer *et al.*, 2024]. Existing evaluation protocols in recommender systems often assess explanations based on whether perturbing the identified important inputs changes the *top-1* recommended item. Although convenient, this practice implicitly reduces a ranked prediction task to a single-item outcome and overlooks the list-wise nature of the model output [Mohammadi *et al.*, 2024]. We argue that top-1 evaluation can produce unstable conclusions. In ranking systems, the highest-ranked item may be sensitive to small score fluctuations, ties, or modest changes in model quality. As a result, explanation methods may appear better or worse depending not on their true explanatory power, but on incidental variation in the underlying model. This creates an important threat to fair comparison and reproducible benchmarking [Mohammadi *et al.*, 2025].

In this paper, we revisit the evaluation of counterfactual explanations through the lens of ranked outputs. We investigate whether extending evaluation beyond the top-1 item to top-k ranked outputs leads to more *consistent* assessments. In this setting, *consistency* denotes the capacity of an evaluation metric to preserve a reliable ordering of CE methods even when the recommender model undergoes performance variations. For instance, if explanation method E_1 performs better than method E_2 for recommender R_1, then after slight changes in the recommender system, the absolute CE scores may vary, but E_1 should remain superior to E_2 in relative ranking. While [Mohammadi *et al.*, 2024] identified such inconsistencies in current evaluation protocols, the root causes remain unresolved. Intuitively, evaluating multiple highly

ranked items should better reflect the structure of recommendation tasks while reducing sensitivity to local ranking noise.

To study this question, we evaluate six representative explanation methods across multiple datasets, recommender architectures, and model performance checkpoints. Our results show that list-wise top-k protocols substantially improve consistency compared with conventional top-1 evaluation. These findings suggest that explanation benchmarks for ranking systems should align with the ranked nature of model outputs rather than inherit evaluation assumptions from single-label prediction tasks.

2 Related Work

Research in interpretable machine learning has advanced significantly in recent years [Ranjbar *et al.*, 2024; Ghorbani and Zou, 2020]. Initial efforts largely concentrated on developing inherently interpretable models, where transparency is embedded directly into the model design, enabling human-understandable reasoning without additional processing [Peintner *et al.*, 2023; Peintner *et al.*, 2025]. Alongside this line of work, post-hoc explainability has become a widely adopted alternative, focusing on explaining model behavior after training while treating the underlying system as a black box [Wachter *et al.*, 2017; Kaffes *et al.*, 2021].

Despite rapid progress in explanation methods, recent studies reveal that evaluation practices remain inconsistent and underdeveloped. A comprehensive survey covering hundreds of XAI papers shows that many approaches rely on qualitative or anecdotal validation, while rigorous quantitative evaluation often addresses only a narrow subset of explanation characteristics [Nauta *et al.*, 2023]. The authors introduce a set of conceptual properties (Co-12), such as *Correctness*, *Compactness*, and *Consistency*, emphasizing that explanation quality is inherently multi-faceted. However, existing work typically focuses on isolated criteria, overlooking interactions between them. This limitation becomes particularly pronounced in settings where model outputs are structured, such as ranked predictions, where explanations depend not only on individual outputs but also on their relative ordering. Motivated by this gap, we focus on consistency as a central property and analyze it in the context of counterfactual explanations for recommender systems.

Explanation Methods for Recommender Systems Explainability in recommender systems has attracted considerable attention due to its role in improving user trust, transparency, and accountability. Existing approaches can be broadly grouped into model-intrinsic, post-hoc, and auxiliary-information-based methods. Model-intrinsic approaches incorporate interpretability directly into the recommendation model. Early work on explainable matrix factorization aligns latent representations with interpretable components [Abdollahi and Nasraoui, 2016]. More recent architectures, such as ProtoMF [Melchiorre *et al.*, 2022], leverage prototypes or attention mechanisms to identify influential user-item interactions. While these methods provide structured explanations, they are often tightly coupled to specific model designs. Post-hoc approaches, in contrast, aim to explain predictions without modifying the underlying recommender. However, these approaches primarily capture correlational relationships and do not explicitly account for the sensitivity of ranked outputs. A separate line of work exploits auxiliary information such as textual reviews, aspects, or knowledge graphs. Aspect-based models like the Explicit Factor Model (EFM) [Zhang *et al.*, 2014] align latent factors with interpretable attributes, while path-based approaches justify recommendations through semantic connections in knowledge graphs [Chen *et al.*, 2020]. More recent studies question the reliability of attention-based explanations and propose causal reasoning over graph paths as a more faithful alternative [Li *et al.*, 2024]. Although these methods provide rich explanations, they rely on additional data sources that are not always available.

Counterfactual Explanations Counterfactual explanations (CEs) provide insights by identifying minimal changes to the input that would alter the model's output [Wachter *et al.*, 2017]. Due to their actionable nature and intuitive interpretation, they have become a prominent approach for explaining recommendation models. Early methods for generating CEs relied on heuristic search or perturbation strategies to identify influential interactions [Mohammadi, 2023]. More recent approaches employ reinforcement learning to efficiently explore the combinatorial space of possible modifications. For example, frameworks such as CERec [Wang *et al.*, 2024] learn policies over attributes or graph structures, often incorporating fairness considerations. Conversational recommender systems further integrate counterfactual reasoning to iteratively refine explanations through user interaction [Yu *et al.*, 2024; Müller *et al.*, 2026]. In parallel, recourse-oriented methods such as RecRec [Verma *et al.*, 2023] focus on actionable modifications that can influence ranking outcomes. Model-agnostic approaches aim to improve scalability and generality. Notably, LXR [Barkan *et al.*, 2024] learns explanation masks through self-supervised objectives, avoiding explicit combinatorial search over perturbations. While effective, such methods do not explicitly address the stability of explanations across ranked outputs.

Evaluation of Counterfactual Explanations Despite substantial progress in generating counterfactual explanations, their evaluation remains limited and often simplistic. Most existing approaches assess explanation quality based on whether the top-ranked item changes after perturbation [Tan *et al.*, 2021; Barkan *et al.*, 2024; Zhong and Negre, 2022; Ser *et al.*, 2024]. This practice effectively reduces a ranked prediction problem to a single outcome, disregarding the structure of the full ranking. Recent studies show that such evaluation protocols can lead to unstable conclusions: small variations in model performance may alter the relative ranking of explanation methods [Mohammadi *et al.*, 2024; Mohammadi *et al.*, 2025]. This instability raises concerns about the reliability and reproducibility of current benchmarks. In contrast, our work adopts a list-wise perspective by evaluating explanations across multiple top-ranked items. By aligning evaluation protocols with the ranked nature of model outputs, we aim to provide a more stable and principled basis for comparing counterfactual explanation methods.

3 Methodology

To investigate the impact of extending the evaluation from top-1 items to top-k items, we propose the following approach.

3.1 Problem Setup and Preliminaries

Let $\mathcal{U}$ denote the set of users and $\mathcal{I}$ the set of items. Each user $u \in \mathcal{U}$ is represented by a binary interaction vector $\mathbf{x}_u \in \{0,1\}^{|\mathcal{I}|}$, where $\mathbf{x}_u[i] = 1$ indicates that user u has interacted with item i. A recommender model f_θ maps this input to a predicted score vector, with $f_\theta(\mathbf{x}_u)[i]$ representing the predicted relevance of item i. The RS generates a ranked list of recommendations for each user. We denote the top-k recommended items as $\mathcal{R}_u^k = \{i_1^*, i_2^*, \ldots, i_k^*\}$, where i_j^* is the j-th item in the ranked recommendation list. These are the items for which we seek to evaluate CEs.

Given a counterfactual explainer e, we obtain a relevance ranking over user history items with respect to each item i_j^* in $\mathcal{R}_u^k$. Based on this relevance, we iteratively perturb the user input vector $\mathbf{x}_u$ by removing one item at a time, yielding a sequence $\mathbf{x}_u^{(t)}$ of perturbed inputs. Following the literature [Barkan *et al.*, 2024], We compute the following metric:

Positive Perturbation (POS-P@T) This metric quantifies how quickly the top-k recommended items fall below the rank threshold T during perturbation. For user u, the POS-P@T score is:

$$\text{POS-P@T}_u = \sum_{j=1}^{k} \sum_{t=1}^{|\mathbf{x}_u|} 1\left[\text{rank}(i_j^*; \mathbf{x}_u^{(t)}) > T\right] \quad (1)$$

where $\text{rank}(i_j^*; \mathbf{x}_u^{(t)})$ is the rank of i_j^* in the output of f_θ given the perturbed input $\mathbf{x}_u^{(t)}$ and the indicator function $1[\cdot]$ checks whether the item still appears within the rankd threshold T. The key difference between this formulation and the literature [Barkan *et al.*, 2024; Mohammadi *et al.*, 2024] lies in computing the inner value over the top-k items ($\sum_{j=1}^{k}$) rather than only the top-1 item.

The final POS-P@T score is the average over all users:

$$\text{POS-P@T} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \text{POS-P@T}_u \quad (2)$$

3.2 Evaluation Setup

We use the MovieLens 1M (ML-1M) [Harper and Konstan, 2015], Yahoo! Music [Dror *et al.*, 2012] and Pinterest [He *et al.*, 2017] datasets following the literature [Mohammadi *et al.*, 2024], focusing on collaborative filtering models based on implicit user-item interactions. For each user, we generate a ranked list of recommendations and evaluate the explanation of the top-k recommendations. For each k value, we use four levels of recommender performance by training the recommender based on Hit Rate and saving model checkpoints after 25%, 50%, 75%, and full training, to simulate recommendations of various quality levels. We evaluate 24 configurations (3 dataset x 2 Recommenders x 4 performance levels), repeat our experiments three times, and report the average performance. We share our dataset processing scripts, the source code, and the hyper-parameters[1].

3.3 Evaluated Methods

Our evaluation includes a range of CE methods, from established baselines to recent advancements.

Jaccard [Bag *et al.*, 2019]/*Cosine* [Singh *et al.*, 2020] Similarity generate explanations by comparing an item to those in the user's history using Jaccard [Bag *et al.*, 2019] or Cosine [Singh *et al.*, 2020] similarity metrics based on co-interacted users.

LIME-RS [Nóbrega and Marinho, 2019] is a modified version of the LIME framework designed specifically for recommender systems.

SHAP (SHapley Additive exPlanations) [Zhong and Negre, 2022] quantifies the influence of individual features on a model's predictions. By leveraging game theory, it assigns a value to each feature according to its contribution to the final output.

ACCENT (Action-based Counterfactual Explanations for Neural Recommenders for Tangibility) [Tran *et al.*, 2021] is a CE framework that extends influence functions to provide actionable, model-agnostic insights for neural recommenders.

LXR [Barkan *et al.*, 2024], the state-of-the-art CE approach, employs self-supervised learning to generate explanation masks that highlight critical user data without perturbations using a gradient learning approach.

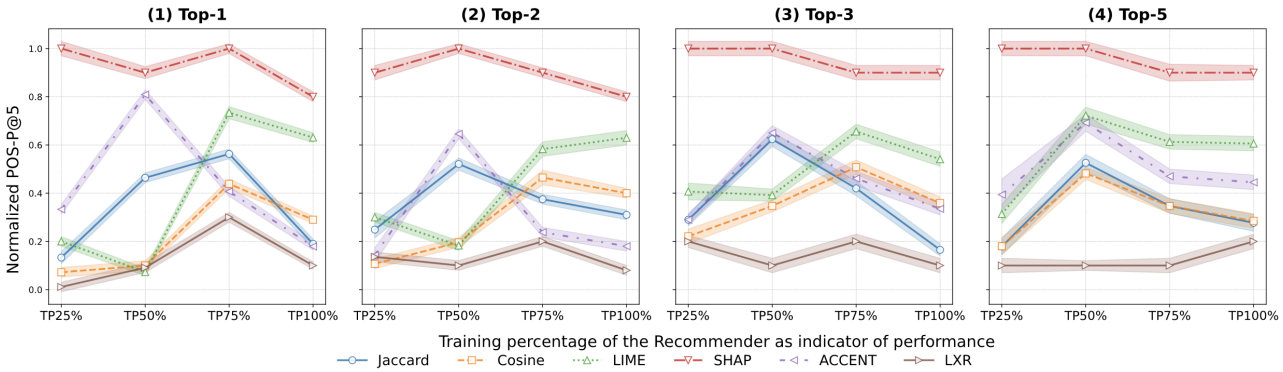
Our study includes two collaborative filtering recommenders: (1) *Matrix Factorization (MF)* [Abdi *et al.*, 2018] and (2) *Variational Autoencoder (VAE)*, a generative model that encodes and decodes data while learning a probabilistic latent representation [Liang *et al.*, 2018].

4 Experiments and Results

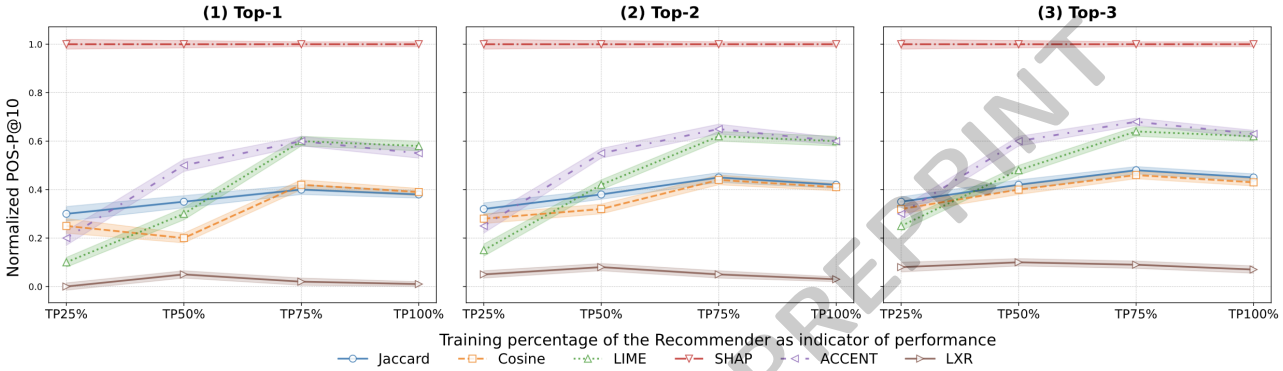
Based on our hypothesis that extending evaluations to include lower-ranked recommendations—such as the top-3 items—enhances consistency, we incorporate lower-ranked items in the evaluation. Particularly, we compute performance separately for top-1, top-2, top-3, and higher-ranked recommendations, then aggregate these performance values to obtain a comprehensive assessment. For instance, computing the value for top-3 involves selectively masking data from the user's history to ensure that the third-highest recommended item drops to a lower rank (rank ≥ 4). This method ensures that the evaluation captures the influence of multiple recommendation ranks rather than relying solely on the highest-ranked item, leading to a more stable comparison of CE methods.

To test this hypothesis, we evaluate six representative CE methods using the POS metric, analyzing its performance across four thresholds of recommendation length (top-k) and four levels of recommender performance. As shown in Figure 1, evaluations based solely on the top-1 recommendation (Fig. 1a (a)) exhibit significant fluctuations on every two recommenders, particularly when transitioning from the lowest-performing recommender (TP25%) to

[1] https://github.com/dbis-uibk/CE4RS-Eval



(a) VAE recommender on Yahoo! Music dataset



(b) MF recommender on Pinterest dataset

Figure 1: Comparison of CE methods based on (a) POS@5 ↓ (lower value is the better) and (b) POS@10 ↓ across different performance levels. The figure shows the impact of going beyond top-1 (a) and considering top-k (b-d) recommendations on improving consistency when evaluating CE models. To facilitate comparisons, the values are normalized, and shading is used to represent the variance in the results.

TP50%. However, when extending the evaluation to the top-2 recommendations (Fig. 1a (b)), we can already observe more consistency, especially for high-performing recommenders (TP50%–TP100%). Following this, setting the threshold to include the top-5 recommendations (Fig. 1a (d)) stabilizes the evaluation outcomes for the highest performing recommenders (TP50%–TP100%). We observed a similar pattern in other configurations, albeit with different thresholds. For the MF recommender, the ranking of methods remained stable even at smaller k values across all datasets, in contrast to the VAE (Fig. 1b) . The observed consistency when considering only the top-3 items stems from the simplicity of the MF model, which results in a more straightforward recommendation task and lower performance variability among recommenders, ultimately leading to less variability in explainers. More detailed results are available in the accompanying Git repository presented in section 3.2. Overall, this evidence highlights the importance of considering listwise recommendations in CE evaluations. Adopting such an approach not only enhances evaluation consistency but also ensures better alignment with the inherent characteristics of recommendation systems. Importantly, our findings show that evaluation consistency is not uniformly affected across all metrics. POS is particularly sensitive to the quality of the recommender. These results suggest that the choice of metric—and the value of k—should be treated as a tunable hyperparameter, influenced by both the dataset and the recommender architecture. Moreover, the use of multiple performance checkpoints for the recommender further reinforces the robustness of our evaluation framework. By capturing performance across training stages, we ensure that the evaluation of explainers is not biased by any single fixed model state. This approach not only provides a more comprehensive view of the explainer's effectiveness but also reveals how susceptible existing metrics are to fluctuations in model quality.

5 Conclusion and Future Work

In this paper, we address a gap in the evaluation of counterfactual explanations for recommender systems, focusing on evaluation consistency. Our findings highlight the importance of considering top-k recommendations when assessing CEs. By extending evaluations beyond the top-1 recommendation, we demonstrate consistency on current mainstream metrics for evaluation. These findings not only address key methodological gaps in counterfactual evaluation but also lay the groundwork for future research into domain-adaptive CE metrics. An important direction for future work is the development of unified evaluation metrics that jointly consider both recommender performance and explanation quality.

References

[Abdi *et al.*, 2018] Mohamed Hussein Abdi, George Onyango Okeyo, and Ronald Waweru Mwangi. Matrix factorization techniques for context-aware collaborative filtering recommender systems: A survey. *Comput. Inf. Sci.*, 11(2):1–10, 2018.

[Abdollahi and Nasraoui, 2016] Behnoush Abdollahi and Olfa Nasraoui. Explainable restricted boltzmann machines for collaborative filtering. *CoRR*, abs/1606.07129, 2016.

[Bag *et al.*, 2019] Sujoy Bag, Sri Krishna Kumar, and Manoj Kumar Tiwari. An efficient recommendation generation using relevant jaccard similarity. *Inf. Sci.*, 483:53–64, 2019.

[Barkan *et al.*, 2024] Oren Barkan, Veronika Bogina, Liya Gurevitch, Yuval Asher, and Noam Koenigstein. A counterfactual framework for learning and evaluating explanations for recommender systems. In *WWW '24*, pages 3723–3733. ACM, 2024.

[Bauer *et al.*, 2024] Christine Bauer, Eva Zangerle, and Alan Said. Exploring the landscape of recommender systems evaluation: Practices and perspectives. *Trans. Recomm. Syst.*, 2(1):11:1–11:31, 2024.

[Chen *et al.*, 2020] Tong Chen, Hongzhi Yin, Guanhua Ye, Zi Huang, Yang Wang, and Meng Wang. Try this instead: Personalized and interpretable substitute recommendation. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 891–900, 2020.

[Doshi-Velez and Kim, 2017] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[Dror *et al.*, 2012] Gideon Dror, Noam Koenigstein, Yehuda Koren, and Markus Weimer. The yahoo! music dataset and kdd-cup '11. In *KDD Cup 2011*, volume 18 of *JMLR Proceedings*, pages 8–18. JMLR.org, 2012.

[Ghorbani and Zou, 2020] Amirata Ghorbani and James Y. Zou. Neuron shapley: Discovering the responsible neurons. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[Ghorbani *et al.*, 2019] Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. Towards automatic concept-based explanations. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9273–9282, 2019.

[Harper and Konstan, 2015] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), dec 2015.

[He *et al.*, 2017] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *WWW '17*, pages 173–182. ACM, 2017.

[Kaffes *et al.*, 2021] Vassilis Kaffes, Dimitris Sacharidis, and Giorgos Giannopoulos. Model-agnostic counterfactual explanations of recommendations. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, Utrecht, The Netherlands, June, 21-25, 2021*, pages 280–285. ACM, 2021.

[Karimi *et al.*, 2021] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *FAccT '21*, pages 353–362. ACM, 2021.

[Li *et al.*, 2024] Yicong Li, Xiangguo Sun, Hongxu Chen, Sixiao Zhang, Yu Yang, and Guandong Xu. Attention is not the only choice: Counterfactual reasoning for path-based explainable recommendation. *IEEE Trans. Knowl. Data Eng.*, 36(9):4458–4471, 2024.

[Liang *et al.*, 2018] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *WWW '18*, pages 689–698. ACM, 2018.

[Melchiorre *et al.*, 2022] Alessandro B. Melchiorre, Navid Rekabsaz, Christian Ganhör, and Markus Schedl. Protomf: Prototype-based matrix factorization for effective and explainable recommendations. In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22, page 246–256, New York, NY, USA, 2022. Association for Computing Machinery.

[Mohammadi *et al.*, 2024] Amir Reza Mohammadi, Andreas Peintner, Michael Müller, and Eva Zangerle. Are we explaining the same recommenders? incorporating recommender performance for evaluating explainers. In *RecSys '24*, page 1113–1118. ACM, 2024.

[Mohammadi *et al.*, 2025] Amir Reza Mohammadi, Andreas Peintner, Michael Müller, and Eva Zangerle. Beyond top-1: Addressing inconsistencies in evaluating counterfactual explanations for recommender systems. In Mária Bieliková, Pavel Kordík, Markus Schedl, Marco de Gemmis, Sole Pera, Rodrigo Alves, Olivier Jeunen, and Vito Ostuni, editors, *Proceedings of the Nineteenth ACM Conference on Recommender Systems, RecSys 2025, Prague, Czech Republic, September 22-26, 2025*, pages 515–520. ACM, 2025.

[Mohammadi, 2023] Amir Reza Mohammadi. Explainable graph neural network recommenders; challenges and opportunities. In Jie Zhang, Li Chen, Shlomo Berkovsky, Min Zhang, Tommaso Di Noia, Justin Basilico, Luiz Pizzato, and Yang Song, editors, *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023*, pages 1318–1324. ACM, 2023.

[Mozafari *et al.*, 2026] Jamshid Mozafari, Hamed Zamani, Guido Zuccon, and Adam Jatowt. Inferential question answering. In Hakim Hacid, Yoelle Maarek, Francesco Bonchi, Ido Guy, and Emine Yilmaz, editors, *Proceedings*

of the ACM Web Conference 2026, WWW 2026, Dubai, United Arab Emirates, originally scheduled for April 13-17, 2026, rescheduled for June 29 - July 3, 2026, pages 2384–2395. ACM, 2026.

[Müller *et al.*, 2026] Michael M. Müller, Amir Reza Mohammadi, Andreas Peintner, Beatriz Barroso Gstrein, Günther Specht, and Eva Zangerle. On the reliability of user-centric evaluation of conversational recommender systems. *CoRR*, abs/2602.17264, 2026.

[Nauta *et al.*, 2023] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Comput. Surv.*, 55(13s):295:1–295:42, 2023.

[Nóbrega and Marinho, 2019] Caio Nóbrega and Leandro Balby Marinho. Towards explaining recommendations through local surrogate models. In *ACM/SIGAPP '19*, pages 1671–1678. ACM, 2019.

[Peintner *et al.*, 2023] Andreas Peintner, Amir Reza Mohammadi, and Eva Zangerle. SPARE: shortest path global item relations for efficient session-based recommendation. In Jie Zhang, Li Chen, Shlomo Berkovsky, Min Zhang, Tommaso Di Noia, Justin Basilico, Luiz Pizzato, and Yang Song, editors, *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023*, pages 58–69. ACM, 2023.

[Peintner *et al.*, 2025] Andreas Peintner, Amir Reza Mohammadi, and Eva Zangerle. Efficient session-based recommendation with contrastive graph-based shortest path search. *ACM Trans. Recomm. Syst.*, 3(4), April 2025.

[Ranjbar *et al.*, 2024] Niloofar Ranjbar, Saeedeh Momtazi, and MohammadMehdi Homayoonpour. Explaining recommendation system using counterfactual textual explanations. *Mach. Learn.*, 113(4), 2024.

[Ser *et al.*, 2024] Javier Del Ser, Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, Francisco Herrera, Anna Saranti, and Andreas Holzinger. On generating trustworthy counterfactual explanations. *Inf. Sci.*, 655:119898, 2024.

[Singh *et al.*, 2020] Ramni Harbir Singh, Sargam Maurya, Tanisha Tripathi, Tushar Narula, and Gaurav Srivastav. Movie recommendation system using cosine similarity and knn. *International Journal of Engineering and Advanced Technology*, 9(5):556–559, 2020.

[Tan *et al.*, 2021] Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. Counterfactual explainable recommendation. In *CIKM '21*, pages 1784–1793. ACM, 2021.

[Tran *et al.*, 2021] Khanh Hiep Tran, Azin Ghazimatin, and Rishiraj Saha Roy. Counterfactual explanations for neural recommenders. In *SIGIR '21*, pages 1627–1631. ACM, 2021.

[Verma *et al.*, 2023] Sahil Verma, Ashudeep Singh, Varich Boonsanong, John P. Dickerson, and Chirag Shah. Recrec: Algorithmic recourse for recommender systems. In Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos, editors, *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, pages 4325–4329. ACM, 2023.

[Wachter *et al.*, 2017] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017.

[Wang *et al.*, 2024] Xiangmeng Wang, Qian Li, Dianer Yu, Qing Li, and Guandong Xu. Reinforced path reasoning for counterfactual explainable recommendation. *IEEE Trans. Knowl. Data Eng.*, 36(7):3443–3459, 2024.

[Yu *et al.*, 2024] Dianer Yu, Qian Li, Xiangmeng Wang, Qing Li, and Guandong Xu. Counterfactual explainable conversational recommendation. *IEEE Trans. Knowl. Data Eng.*, 36(6):2388–2400, 2024.

[Yu *et al.*, 2025] Yi Yu, Kazunari Sugiyama, and Adam Jatowt. Domain counterfactual data augmentation for explainable recommendation. *ACM Trans. Inf. Syst.*, 43(3), February 2025.

[Zhang *et al.*, 2014] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 83–92, 2014.

[Zhong and Negre, 2022] Jinfeng Zhong and Elsa Negre. Shap-enhanced counterfactual explanations for recommendations. In *SAC '22*, pages 1365–1372. ACM, 2022.