From Counterfactual Edits to Human Narratives: LLM-Driven Interpretations for Recommender System Explanations

Vojtech Cahlik*
Faculty of Information Technology,
Czech Technical University in Prague
Prague, Czechia
vojtech.cahlik@fit.cvut.cz

Amir Reza Mohammadi*
Department of Computer Science,
University of Innsbruck
Innsbruck, Austria
amir.reza@uibk.ac.at

Rodrigo Alves
Faculty of Information Technology,
Czech Technical University in Prague
Prague, Czechia
rodrigo.alves@fit.cvut.cz

Pavel Kordik
Faculty of Information Technology,
Czech Technical University in Prague
Prague, Czechia
pavel.kordik@fit.cvut.cz

Eva Zangerle
Department of Computer Science,
University of Innsbruck
Innsbruck, Austria
eva.zangerle@uibk.ac.at

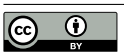
Abstract

Recommender systems increasingly influence what users discover and consume online, yet the decision logic behind their outputs often remains opaque. Counterfactual explanation (CFX) methods provide an effective way to show which changes to a user's interaction history would alter a recommendation. In recommendation settings, however, CFXs often require many edits, producing long, low-level lists that are hard for users to interpret. Moreover, most evaluations emphasize edit counts while largely overlooking whether the resulting CFXs are coherent and actionable for users. We argue that this gap between generating valid CFXs and communicating model behavior stems from the lack of user-centered abstraction and limited qualitative evaluation. To bridge this gap, we investigate large language models as mediators that summarize raw CFX interaction edit lists into concise natural language explanations describing higher-level preference patterns. We optimize these mediators for correctness, informativeness, and linguistic quality through preference optimization, and evaluate the resulting explanations with human-calibrated metrics derived from separate human-labeled datasets. On the MovieLens-1M dataset, across multiple counterfactual generators and language models, the resulting summaries achieve high overall linguistic quality, while the strongest semantic results are obtained with high-success CFX methods such as ACCENT and SPINRec. Direct Preference Optimization improves the semantic metrics in almost all evaluated settings.

CCS Concepts

• **Information systems** → **Recommender systems**; • **Computing methodologies** → **Natural language processing**; *Machine learning*.

*Both authors contributed equally to this research.



This work is licensed under a Creative Commons Attribution 4.0 International License.
CIKM '26, Rome, Italy
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2539-5/2026/11
https://doi.org/10.1145/3799682.3839968

Keywords

Recommender Systems, Counterfactual Explanations, Natural Language Explanations, Large Language Models

ACM Reference Format:
Vojtech Cahlik, Amir Reza Mohammadi, Rodrigo Alves, Pavel Kordik, and Eva Zangerle. 2026. From Counterfactual Edits to Human Narratives: LLM-Driven Interpretations for Recommender System Explanations. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26), November 07–11, 2026, Rome, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3799682.3839968

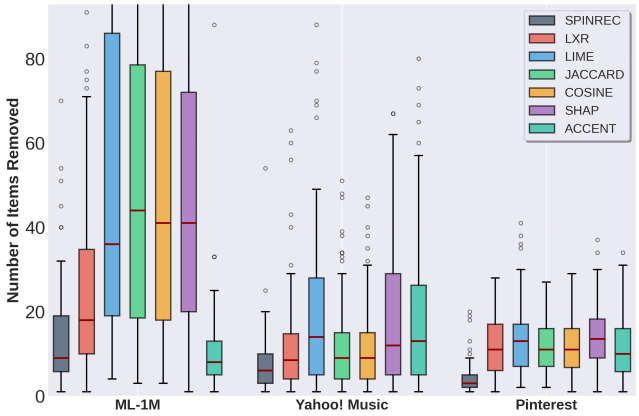
1 Introduction

Recommender systems curate what users see, click, and watch [3, 8], yet users often have little insight into why a particular item is suggested. As recommendations become ubiquitous across online platforms, explanations that improve user trust [29], accountability [32, 34], and informed decision-making [26, 38] are increasingly important.

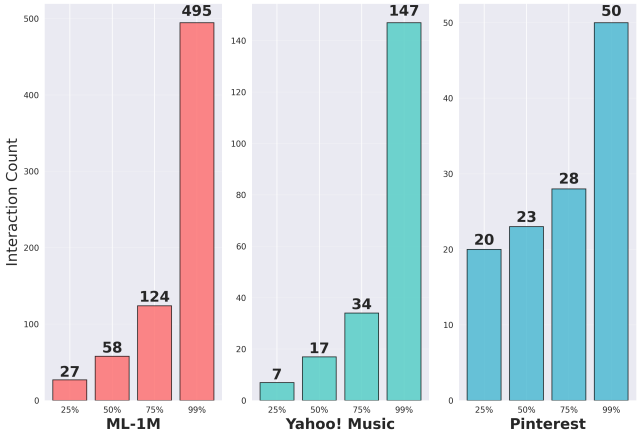
Counterfactual explanations (CFXs) have recently emerged as a promising mechanism for transparency in recommender systems [37]. By identifying minimal changes to a user's interaction history that would alter the recommendation outcome, they offer actionable "what-if" reasoning aligned with human intuition [4, 14]. However, in practice, CFXs for recommenders often require modifying a large number of historical interactions. As illustrated in Figure 1, altering even a single top-ranked recommendation in the MovieLens-1M dataset[1] (ML-1M) may require removing dozens of past interactions. This observation exposes a fundamental mismatch between *algorithmic minimality* and *human interpretability* [37]. Although a counterfactual may be minimal under an optimization objective, it is typically presented as a list of low-level perturbations such as item removals. Long edit sets can therefore overwhelm users and limit practical use. This raises the core challenge we address: how can large counterfactual edit sets be turned into concise natural language explanations (NLEs) while preserving the counterfactual grounding?

We propose an approach in which large language models (LLMs) act as mediators between counterfactual generators and users. LLMs

[1] https://grouplens.org/datasets/movielens/1m/



(a) Number of interactions removed to displace the top-ranked item from the top-5 recommendation list.



(b) Percentile analysis of user history size (25th, 50th, 75th, and 99th percentiles).

Figure 1: Compactness analysis of counterfactual explanations of our variational autoencoder recommender across various datasets. (a) Counterfactual explanations are often composed of a large number of interactions. (b) These interactions often represent a substantial share of a typical user's history.

are well suited to this role because, unlike fixed heuristics, they offer flexibility in how raw interaction data is communicated and can handle free-form text descriptions. Rather than replacing counterfactual reasoning, the mediator summarizes raw interaction edits into concise, human-centered NLEs. This mediation layer is designed to make the explanations comprehensible while preserving the underlying counterfactual grounding.

Beyond explanation generation, we also address evaluation. Existing metrics largely focus on edit counts or success rates and provide limited insight into interpretability. We use three established qualitative criteria: *Correctness* and *Informativeness*, which we decompose into simple interaction-level matching judgments, and *Linguistic Quality*, which is scored by a prompted LLM rubric during preference optimization and by a separate human-calibrated evaluator in the final reported results.

Our contributions are (i) an LLM mediator framework that summarizes counterfactual explanations into faithful narratives, (ii) an evaluation methodology that combines decomposed interaction-level checks with human-calibrated final evaluation, and (iii) an end-to-end comparison across various CFX generators and LLMs. Experiments on the ML-1M dataset demonstrate that LLM-mediated summaries make counterfactual explanations comprehensible while preserving the counterfactual grounding. The findings suggest that combining causal counterfactual reasoning with summarization is a practical step toward more trustworthy and actionable explanations for recommender-system users. To support reproducibility, we publicly release our code and experimental pipeline[2].

2 Related Work

Explainable AI for recommender systems has attracted growing attention due to its importance for transparency, user trust, and regulatory compliance [9, 16]. Prior work falls into four broad categories: model-specific and post-hoc explanations, counterfactual explanations, side-information-based explanations, and natural language explanation methods.

Model-Specific and Post-Hoc Explanations. Model-specific methods embed interpretability into the model architecture. Early examples include explainable matrix factorization models that align latent factors with interpretable components [1, 2, 20]. Recent methods such as ProtoMF [25] and AI2V++ [15] leverage prototypes or attention to highlight representative interactions. Post-hoc approaches explain predictions while treating the model as a black box: LIME [32] fits local surrogate models, SHAP [23] estimates feature attributions, and influence-based methods such as FIA [13] and ACCENT [36] measure training interaction impact. Similarity heuristics (Jaccard, Cosine) [5, 35] are efficient but rely on item similarity rather than on the recommender's own decision process. Overall, these methods typically output scores or long interaction lists that remain difficult to interpret.

Counterfactual Explanations. CFXs identify minimal edits to a user's history that change recommendation outcomes [6, 36], offering causal and actionable insights compared to attribution methods. *Jaccard* [5] and *Cosine* [35] *similarities* explain recommendations by comparing target items with a user's history via interaction-based similarity. *LIME-RS* [28] locally approximates the recommender with an interpretable linear model, while *SHAP* [40] assigns feature contributions based on Shapley values, albeit at high computational cost. *ACCENT* [36] extends influence functions to produce actionable counterfactual explanations for neural recommenders, *LXR* [6] uses self-supervised gradient masks to identify key interactions without perturbations, and *SPINRec* [7] integrates gradients from realistic user baselines and selects attributions that maximize counterfactual fidelity for stable, faithful explanations. We note that the recently proposed DiceRec [27] method was not yet available at the time this research was conducted and therefore is not included in our comparisons. Existing evaluations still focus mainly on success rates and edit counts, with much less attention to interpretability.

[2]https://github.com/vcahlik/from-counterfactual-edits-to-human-narratives

Side-Information Explanations. Other approaches generate explanations from metadata, reviews, or knowledge graphs. Aspect-based models such as EFM [39] align latent factors with explicit attributes, while path-based and graph methods justify recommendations through semantic reasoning chains [11, 21]. Although expressive, these techniques rely on rich side information and are less applicable to interaction-only settings.

Natural Language Explanations and LLMs. Natural language explanations present model behavior in human-friendly text and are often evaluated by plausibility, faithfulness, or readability [18]. Recent LLM-based approaches utilize prompting, synthetic rationales, or explanation supervision [10, 12, 17, 24, 31]. In the recommendation domain, CAVIAR [19] generates natural language explanations by describing perturbations of the recommended item's image. FIRE [33] similarly uses attribution signals to guide LLM-generated explanations, improving linguistic quality but remaining grounded in correlational importance rather than causal interventions.

3 Methodology

We connect counterfactual reasoning with natural language explanations. Unlike attribution- or similarity-based methods, we start from causally grounded counterfactual edits of user interaction history and condition the LLM directly on counterfactual interactions and samples from the user's remaining history to achieve semantic faithfulness.

Our framework consists of four stages:

(1) generation of counterfactual interaction-level edits using an existing CFX method;
(2) sampling of (i) counterfactual interactions and (ii) interactions from the rest of the user's history, and their conversion to item-level metadata;
(3) generation of the natural language explanation using an LLM conditioned on both sets; and
(4) evaluation and optional preference optimization of the generated explanations.

This pipeline turns low-level counterfactual edits into concise semantic explanations while preserving their counterfactual grounding.

3.1 Basic Notation

Let $\mathcal{U}$ and $\mathcal{I}$ denote the sets of users and items. Each user $u \in \mathcal{U}$ is associated with an interaction history $H_u \subseteq \mathcal{I}$. A recommender model f_θ scores items $i \in \mathcal{I}$ for a user as $s_\theta(i \mid H_u)$, and the induced ranking defines the top-K recommendation list $\mathrm{Top}_K(H_u) \subseteq \mathcal{I}$.

A counterfactual explanation C_u for user u is a subset of historical interactions, $C_u \subseteq H_u$, whose removal changes the recommendation outcome. In this paper, we consider interventions that displace the originally top-ranked item $i_u^\star$ from the top-K recommendation list, i.e., such that $i_u^\star \notin \mathrm{Top}_K(H_u \setminus C_u)$. We denote non-CFX interactions by $\bar{C}_u = H_u \setminus C_u$.

3.2 From Counterfactual Edits to Natural Language Explanations

Our goal is to bridge the gap between CFX outputs, which often take the form of long lists of interaction edits, and the way humans typically reason about preferences. We introduce an LLM mediator that generates a natural language explanation T from (i) the metadata of a subset of CFX interactions and (ii) the metadata of a subset of non-CFX interactions, i.e., the user's remaining history:

$$T = \mathrm{LLM}_{\mathrm{gen}}\left(\phi(C_{u,k_{\mathrm{gen}}}), \phi(\bar{C}_{u,k_{\mathrm{gen}}})\right), \tag{1}$$

where $\phi(i)$ maps each interaction item or set of items to metadata, $\mathrm{LLM}_{\mathrm{gen}}$ is a prompted LLM call that generates a natural language explanation from counterfactual and non-counterfactual interactions, and $C_{u,k_{\mathrm{gen}}}$ (resp. $\bar{C}_{u,k_{\mathrm{gen}}}$) denotes a uniform sample of k_{gen} interactions from C_u (resp. $\bar{C}_u$), with k_{gen} controlling how many items are provided to the generator prompt.

3.3 Evaluation Metrics

(1) Natural Language Explanation Correctness. We evaluate the correctness of an NLE T by measuring how well it matches a sample of the user's CFX interactions:

$$\mathrm{Cor}(T, C_u) = \frac{1}{|C_{u,k_{\mathrm{eval}}}|} \sum_{i \in C_{u,k_{\mathrm{eval}}}} \mathrm{score}_{\mathrm{match}}\big(T, \phi(i)\big), \tag{2}$$

where $\mathrm{score}_{\mathrm{match}}(T, \phi(i)) \in [0, 1]$ scores how well the explanation T matches item metadata $\phi(i)$, and k_{eval} denotes the number of sampled interactions from each set used for evaluation. During optimization, this match score is produced by a prompted LLM judge; in the final evaluation, it is produced by a human-calibrated binary classifier trained on explanation–interaction pairs. By construction, $\mathrm{Cor}(T, C_u) \in [0, 1]$.

(2) Natural Language Explanation Informativeness. Intuitively, an informative explanation should emphasize properties that distinguish the CFX interactions from the remaining user history. For example, an explanation that strongly matches both interaction sets is overly generic and therefore uninformative. We therefore evaluate informativeness by measuring the contrast between how well the explanation matches a sample of the CFX interactions and how well it matches a sample of the non-CFX interactions:

$$\mathrm{Info}(T, C_u, \bar{C}_u) = \mathrm{Cor}(T, C_u) - \frac{1}{|\bar{C}_{u,k_{\mathrm{eval}}}|} \sum_{j \in \bar{C}_{u,k_{\mathrm{eval}}}} \mathrm{score}_{\mathrm{match}}\big(T, \phi(j)\big). \tag{3}$$

By construction, $\mathrm{Info}(T, C_u, \bar{C}_u) \in [-1, 1]$, where values near zero indicate generic explanations that match both sets similarly.

(3) Linguistic Quality of Natural Language Explanations. We define linguistic quality as an overall score assigned directly to the explanation, $\mathrm{LQ}(T) = \mathrm{score}_{\mathrm{ling}}(T) \in [0, 1]$. During DPO optimization, this signal is produced by a prompted six-criterion LLM judge rubric covering fluency, grammar, length, naturalness, illustrativeness, and specificity. In the final reported evaluation, however, we use a human-calibrated binary classifier trained on overall linguistic-quality judgments, so the reported metric is a direct calibrated binary score instead of an output of a decomposed human rubric. This keeps the human annotations holistic and authentic, rather than forcing separate criterion-level labels.

Table 1: Comparison of counterfactual (CFX) generators and the corresponding LLM mediators on ML-1M, averaged across users (mean ± std). CFX Success Rate, CFX Size, and Simple CFX Rate characterize the CFX generator and are defined in Section 4.1; Correctness, Informativeness, and Linguistic Quality are defined in Section 3.3 and are reported here using the human-calibrated evaluators, separately for the off-the-shelf Vanilla LLM and its DPO-optimized variant. Higher (↑) or lower (↓) values are better. Within each column, the best result is in bold and the second-best is underlined; ties were decided using unrounded values.

CFX Generator	CFX Success Rate ↑	CFX Size ↓	Simple CFX Rate ↑	LLM	Correctness ↑		Informativeness ↑		Linguistic Quality ↑	
					Vanilla LLM	DPO	Vanilla LLM	DPO	Vanilla LLM	DPO
Jaccard	0.95	16.38 ± 11.67	0.02	Ministral 8B	0.34 ± 0.19	0.41 ± 0.20	0.14 ± 0.19	0.12 ± 0.20	0.88 ± 0.33	0.75 ± 0.44
				Gemma 3 12B	0.39 ± 0.21	0.51 ± 0.18	0.13 ± 0.20	0.17 ± 0.23	0.72 ± 0.45	0.84 ± 0.37
				Qwen3 8B	0.34 ± 0.19	0.46 ± 0.20	0.13 ± 0.19	0.18 ± 0.24	0.85 ± 0.36	0.71 ± 0.46
Cosine	0.96	16.20 ± 11.23	0.02	Ministral 8B	0.34 ± 0.19	0.40 ± 0.21	0.16 ± 0.18	0.17 ± 0.21	**0.91 ± 0.29**	0.78 ± 0.41
				Gemma 3 12B	0.38 ± 0.21	0.41 ± 0.22	0.13 ± 0.19	0.14 ± 0.21	0.73 ± 0.44	0.78 ± 0.42
				Qwen3 8B	0.34 ± 0.19	0.40 ± 0.19	0.13 ± 0.22	0.17 ± 0.23	0.85 ± 0.36	0.78 ± 0.41
LIME-RS	0.84	18.52 ± 11.84	0.01	Ministral 8B	0.32 ± 0.17	**0.52 ± 0.16**	0.11 ± 0.19	0.13 ± 0.21	0.86 ± 0.34	0.59 ± 0.49
				Gemma 3 12B	0.37 ± 0.21	0.39 ± 0.18	0.11 ± 0.21	0.14 ± 0.20	0.62 ± 0.49	0.78 ± 0.42
				Qwen3 8B	0.36 ± 0.16	0.44 ± 0.18	0.10 ± 0.20	0.14 ± 0.22	0.80 ± 0.40	0.64 ± 0.48
SHAP	0.76	20.83 ± 11.82	0.01	Ministral 8B	0.29 ± 0.17	0.32 ± 0.17	0.09 ± 0.17	0.09 ± 0.18	0.87 ± 0.34	**0.86 ± 0.34**
				Gemma 3 12B	0.32 ± 0.20	0.33 ± 0.20	0.07 ± 0.19	0.08 ± 0.20	0.71 ± 0.46	0.76 ± 0.43
				Qwen3 8B	0.32 ± 0.16	0.34 ± 0.16	0.09 ± 0.20	0.10 ± 0.20	0.77 ± 0.42	0.77 ± 0.42
ACCENT	0.99	**10.36 ± 7.24**	**0.07**	Ministral 8B	0.38 ± 0.19	0.40 ± 0.20	0.18 ± 0.20	0.18 ± 0.20	0.84 ± 0.36	0.83 ± 0.37
				Gemma 3 12B	**0.43 ± 0.22**	0.49 ± 0.21	0.16 ± 0.21	0.21 ± 0.22	0.72 ± 0.45	0.72 ± 0.45
				Qwen3 8B	0.41 ± 0.19	0.44 ± 0.23	**0.18 ± 0.22**	0.21 ± 0.24	0.73 ± 0.44	0.64 ± 0.48
LXR	0.93	15.91 ± 11.03	0.00	Ministral 8B	0.33 ± 0.18	0.42 ± 0.18	0.13 ± 0.20	0.19 ± 0.20	0.84 ± 0.36	0.79 ± 0.41
				Gemma 3 12B	0.40 ± 0.21	0.48 ± 0.20	0.12 ± 0.20	0.13 ± 0.22	0.66 ± 0.47	0.84 ± 0.36
				Qwen3 8B	0.36 ± 0.19	0.36 ± 0.19	0.12 ± 0.21	0.12 ± 0.21	0.84 ± 0.37	0.85 ± 0.36
SPINRec	**1.00**	12.25 ± 8.75	0.05	Ministral 8B	0.36 ± 0.19	0.43 ± 0.20	0.15 ± 0.19	0.15 ± 0.21	0.85 ± 0.36	0.71 ± 0.46
				Gemma 3 12B	0.41 ± 0.21	0.50 ± 0.23	0.15 ± 0.20	0.17 ± 0.22	0.71 ± 0.46	0.65 ± 0.48
				Qwen3 8B	0.39 ± 0.19	0.50 ± 0.21	0.15 ± 0.21	**0.22 ± 0.21**	0.75 ± 0.44	0.64 ± 0.48

Table 2: Examples of natural language explanations obtained for ACCENT and DPO-tuned Ministral 8B, together with per-explanation evaluation scores. For each metric, LLM judge is the score obtained using the prompted LLM judge used during DPO optimization, and Human-cal. is the score obtained using the corresponding human-calibrated classifier used for the final evaluation. Higher values are better (↑).

Explanation	Correctness ↑		Informativeness ↑		Linguistic Quality ↑	
	LLM judge	Human-cal.	LLM judge	Human-cal.	LLM judge	Human-cal.
(Because you watched) comedy and drama films from 1990s	0.35	0.53	0.02	0.08	0.83	1.00
(Because you watched) adventure romances from 1995	0.21	0.12	0.11	0.12	0.94	1.00
(Because you watched) drama films from 1960s to 1990s	0.83	0.50	0.44	0.40	0.89	0.00
(Because you watched) mystery and action films from 1970s to 1990s	0.60	0.40	0.28	0.25	1.00	1.00
(Because you watched) comedy films from 1980s and 1990s	0.46	0.25	-0.15	-0.20	0.89	1.00

3.4 Optimization

To improve the overall quality of the generated natural language explanations, we optimize the generation LLM using Direct Preference Optimization (DPO) [30]. We construct pairwise NLE data on held-out users and score candidate explanations with prompted LLM judges using the correctness, informativeness, and linguistic-quality signals defined above. We manually refine these judge prompts on a human-labeled validation dataset to better align the optimization signal with human judgments. We then tune the main DPO hyperparameters, that is, the learning rate and the regularization parameter β, by random search under a fixed compute budget and select the final DPO variant using separate human-calibrated evaluators on held-out validation users.

4 Experiments and Results

4.1 Experimental Setup

We evaluate our approach on the movie-database ML-1M dataset, which represents a dense interaction setting typical of mature recommendation platforms. We train a variational autoencoder recommender [22] on a separate training set of users and use it to generate CFXs that push the originally top-ranked item out of the top-5 recommendation list. To prevent the explanation LLM from relying on its internal knowledge when generating the NLEs, we omit item titles and use only the release year and list of genres as item features. We refer to the number of removed interactions $|C_u|$ as the *CFX Size*. Cases with $|C_u| > 50$ are treated as failures, as such large edit sets are unlikely to yield usable human-facing explanations; we report the fraction of successful cases as *CFX Success Rate*. Additionally,

we exclude CFXs with $|C_u| < 3$, whose fraction with respect to all successful cases we report as the *Simple CFX Rate*, since these short edit sets can usually be communicated directly from the items themselves without additional summarization. In LLM prompts, we use interaction set sample sizes of $k_{\text{gen}} = k_{\text{eval}} = 20$.

We use Ministral 8B, Gemma 3 12B, and Qwen3 8B LLMs for NLE generation. For DPO, we construct pairwise training datasets on a validation set of users using alternative explanations generated by two LLMs. Candidate explanations are then scored by prompted LLM judges, and the main DPO hyperparameters, learning rate and β, are tuned by random search under a fixed compute budget. The DPO variant reported in the main table is selected by human-calibrated validation performance on held-out validation users.

To avoid reporting metrics used during optimization as final evaluation, the main table uses separate human-calibrated evaluators for final testing. We train ModernBERT-large binary classifiers on separate human-labeled datasets, each with 100 held-out test instances, obtaining accuracies of 0.91 for linguistic quality and 0.92 for explanation–interaction matching. Correctness and informativeness then use the same formulas as in Section 3.3 with classifier-based interaction scores, while linguistic quality is obtained using individual binary scores.

4.2 Results

Under the human-calibrated evaluation approach, Table 1 summarizes all results, and Table 2 shows representative ACCENT explanations from the DPO-tuned Ministral 8B model.

Characteristics of the CFX Generators. ACCENT and SPINRec achieve the highest CFX success rates (0.99 and 1.00, respectively) and also produce the most compact CFXs, with average sizes of 10.36 and 12.25 interactions. They additionally have the highest simple-CFX rates, although this rate remains low for all methods. Most cases therefore still involve larger edit sets and benefit from natural language summarization.

Correctness and Informativeness. Semantic quality is highest for the stronger CFX methods, especially ACCENT and SPINRec, even though the individual best rows are spread across several method–LLM combinations. DPO improves informativeness for most method–LLM pairs, and improves correctness in all but one case, indicating that preference optimization usually helps the LLM produce summaries that are closer to the counterfactual edits and better highlight patterns that distinguish them from the rest of the user's history.

Linguistic Quality. Linguistic quality remains high overall, with many settings above 0.80 under the human-calibrated evaluator, and above 0.85 for the strongest results. Some spread remains in these scores, which is partially expected given the multiple optimization objectives. This highlights why a careful design of the optimized metrics is important, as it must allow the optimization to improve semantic alignment without losing pressure towards readable outputs.

5 Conclusion

In this work, we presented a natural language explanation framework for recommender systems that summarizes raw counterfactual edits using LLMs. These explanations remain grounded in the underlying counterfactuals in terms of human-calibrated correctness and informativeness while maintaining high linguistic quality. Performance depends on both the underlying CFX generator and the language model; Direct Preference Optimization improved correctness and informativeness in nearly all of the tested settings. Overall, the results suggest that LLM-based summarization can make counterfactual explanations more accessible without losing their causal grounding.

Acknowledgments

This work was supported by the project "PoliRuralPlus: Fostering Sustainable Development of Rural-Urban Communities Using Spatial Enhanced Attractiveness Mapping ToolBox," funded by the European Union (REA), grant No. 101136910; by the FOCAL project, funded by the European Union under grant agreement No. 101137787; by the National Centre for Artificial Intelligence (NCUI), project No. TQ28000003, co-funded by the Technology Agency of the Czech Republic under the SIGMA Programme; and by the scholarship granted by the University of Innsbruck, Vice-Rectorate for Research (Doktoratsstipendium 2026). Access to the computational infrastructure of the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics" is gratefully acknowledged. Computational resources were also provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

GenAI Usage Disclosure

We used Cursor with Composer-2-series, GPT-5-series, and Claude Opus 4-series models for coding assistance and debugging. All AI-generated code suggestions were manually reviewed, and revised as necessary, before use. All experiments and analyses were conducted manually without the use of generative AI.

References

[1] Behnoush Abdollahi and Olfa Nasraoui. 2016. Explainable Matrix Factorization for Collaborative Filtering. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Companion Volume*. ACM, 5–6. doi:10.1145/2872518.2889405

[2] Behnoush Abdollahi and Olfa Nasraoui. 2017. Using explainability for constrained matrix factorization. In *Proceedings of the eleventh ACM conference on recommender systems*. 79–83.

[3] Darius Afchar, Alessandro B. Melchiorre, Markus Schedl, Romain Hennequin, Elena V. Epure, and Manuel Moussallam. 2022. Explainability in Music Recommender Systems. *CoRR* abs/2201.10528 (2022). arXiv:2201.10528 https://arxiv.org/abs/2201.10528

[4] Saugat Aryal and Mark T. Keane. 2024. Even-Ifs from If-Onlys: Are the Best Semi-factual Explanations Found Using Counterfactuals as Guides?. In *Case-Based Reasoning Research and Development -ICCBR 2024, Merida, Mexico, July 1-4, 2024, Proceedings*, Vol. 14775. Springer, 33–49. doi:10.1007/978-3-031-63646-2_3

[5] Sujoy Bag, Sri Krishna Kumar, and Manoj Kumar Tiwari. 2019. An efficient recommendation generation using relevant Jaccard similarity. *Inf. Sci.* 483 (2019), 53–64.

[6] Oren Barkan, Veronika Bogina, Liya Gurevitch, Yuval Asher, and Noam Koenigstein. 2024. A Counterfactual Framework for Learning and Evaluating Explanations for Recommender Systems. In *WWW '24*. ACM, 3723–3733.

[7] Oren Barkan, Yahlly Schein, Yehonatan Elisha, Veronika Bogina, Mikhail Baklanov, and Noam Koenigstein. 2025. Fidelity-Aware Recommendation Explanations via Stochastic Path Integration. *CoRR* abs/2511.18047 (2025). arXiv:2511.18047 doi:10.48550/ARXIV.2511.18047

[8] 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

[9] Mustafa Bilgic and Raymond J Mooney. 2005. Explaining recommendations: Satisfaction vs. promotion. In *Beyond personalization workshop, IUI*, Vol. 5. 153.

[10] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems* 31 (2018).

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

[12] Yanda Chen, Chandan Singh, Xiaodong Liu, Simiao Zuo, Bin Yu, He He, and Jianfeng Gao. 2025. Towards consistent natural-language explanations via explanation-consistency finetuning. In *Proceedings of the 31st International Conference on Computational Linguistics*. 7558–7568.

[13] Weiyu Cheng, Yanyan Shen, Linpeng Huang, and Yanmin Zhu. 2019. Incorporating Interpretability into Latent Factor Models via Fast Influence Analysis. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 885–893. doi:10.1145/3292500.3330857

[14] Timo Freiesleben. 2022. The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples. *Minds Mach.* 32, 1 (2022), 77–109. doi:10.1007/s11023-021-09580-9

[15] Keren Gaiger, Oren Barkan, Shir Tsipory-Samuel, and Noam Koenigstein. 2023. Not all memories created equal: Dynamic user representations for collaborative filtering. *IEEE Access* 11 (2023), 34746–34763.

[16] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. 241–250.

[17] Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. 2023. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207* (2023).

[18] Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685* (2020).

[19] Neham Jain, Vibhhu Sharma, and Gaurav Sinha. 2024. Counterfactual Explanations for Visual Recommender Systems. In *WWW '24*. ACM, 674–677.

[20] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[21] Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. *arXiv preprint arXiv:2105.11601* (2021).

[22] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*. 689–698.

[23] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[24] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*.

[25] Alessandro B. Melchiorre, Navid Rekabsaz, Christian Ganhör, and Markus Schedl. 2022. ProtoMF: Prototype-based Matrix Factorization for Effective and Explainable Recommendations. In *Proceedings of the 16th ACM Conference on Recommender Systems* (Seattle, WA, USA) *(RecSys '22)*. Association for Computing Machinery, New York, NY, USA, 246–256. doi:10.1145/3523227.3546756

[26] Amir Reza Mohammadi, Andreas Peintner, Michael Müller, and Eva Zangerle. 2024. Are We Explaining the Same Recommenders? Incorporating Recommender Performance for Evaluating Explainers. In *RecSys '24*. ACM, 1113–1118.

[27] Amir Reza Mohammadi, Andreas Peintner, Michael M. Müller, and Eva Zangerle. 2026. Minimal-Perturbation Counterfactuals through Guided Denoising Diffusion for Recommender Systems Explanation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2026, MelbourneVICAustralia, July 20-24, 2026*, Alistair Moffat, Falk Scholer, Hannah Bast, Marc Najork, and Min Zhang (Eds.). ACM, 1371–1381. doi:10.1145/3805712.3809584

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

[29] Andreas Peintner, Amir Reza Mohammadi, and Eva Zangerle. 2023. SPARE: Shortest Path Global Item Relations for Efficient Session-based Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023*. ACM, 58–69. doi:10.1145/3604915.3608768

[30] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems* 36 (2023), 53728–53741.

[31] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361* (2019).

[32] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. ACM, 1135–1144. doi:10.1145/2939672.2939778

[33] SMF Sani, Asal Meskin, Mohammad Amanlou, and Hamid R Rabiee. 2025. FIRE: Faithful Interpretable Recommendation Explanations. *arXiv preprint arXiv:2508.05225* (2025).

[34] Javier Del Ser, Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, Francisco Herrera, Anna Saranti, and Andreas Holzinger. 2024. On generating trustworthy counterfactual explanations. *Inf. Sci.* 655 (2024), 119898. doi:10.1016/J.INS.2023.119898

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

[36] Khanh Hiep Tran, Azin Ghazimatin, and Rishiraj Saha Roy. 2021. Counterfactual Explanations for Neural Recommenders. In *SIGIR '21*. ACM, 1627–1631.

[37] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *CoRR* abs/1711.00399 (2017). arXiv:1711.00399 http://arxiv.org/abs/1711.00399

[38] Yi Yu, Kazunari Sugiyama, and Adam Jatowt. 2025. Domain Counterfactual Data Augmentation for Explainable Recommendation. *ACM Trans. Inf. Syst.* 43, 3, Article 58 (Feb. 2025), 30 pages. doi:10.1145/3711856

[39] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. 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*. 83–92.

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