Exposing the Biased Vulnerabilities of Large Language Models in Explainable Recommender Systems

Weizhi Chen[1*], Xingkong Ma[1*], Bo Liu[2], Baoyun Peng[2]

[1]College of Computer Science and Technology, National University of Defense Technology, Changsha, Hunan, China
[2] Academy of Military Science, Beijing, China
{chenweizhi17, maxingkong, kyle.liu, pengbaoyun13}@nudt.edu.cn, @alumni.nudt.edu.cn
*These authors contributed equally to this work.

Abstract

Explainable recommender systems (XRSs) enhance user trust by providing personalized recommendations followed by persuasive explanations. Integrating large language models (LLMs), such as GPT-4, advances this domain but introduces risks from biases embedded within LLMs. These biases can lead XRSs to generate persuasive explanations that promote favored recommendations, influencing users to accept the model's preferences over their own. This paper identifies a previously unrecognized security threat: the intentional induction of XRSs via biased LLMs to promote specific items through misleading yet compelling explanations. Inspired by work in the psychology of persuasion, we construct biased datasets and systematically insert these biases into LLM-based XRSs. Experiments across four leading LLMs reveal that biases can significantly affect user decisions, with close to 50% of users changing their choices. To counteract this, we propose a prompt rephrasing defense that effectively mitigates these biases, safeguarding the trustworthiness of XRSs.

Keywords: explainable recommender systems; large language models; bias; persuasion

Introduction

Explainable recommender systems (XRSs) play a crucial role in enhancing user trust and satisfaction by offering tailored recommendations paired with clear explanations. Major platforms such as Netflix (Hidasi, Karatzoglou, Baltrunas, & Tikk, 2016), Amazon (Linden, Smith, & York, 2003), and Yelp (P. Li, Wang, Ren, Bing, & Lam, 2017) have implemented XRSs to enrich user experiences. For instance, Netflix might suggest a movie based on a user's viewing history and elucidate why it aligns with their preferences. Traditional approaches to generating these explanations often relied on predefined templates or deep learning models (Zhang, Chen, & et al., 2020). The advent of Large Language Models (LLMs) like GPT-4 has revolutionized this domain by enabling the creation of sophisticated, contextually relevant, and semantically rich explanations. Geng, Liu, Fu, Ge, and Zhang (2022) propose a LLM-based framework leveraging multi-task learning and tailored prompts to generate semantically accurate recommendation explanations. While LLMs bring significant advancements to XRSs, they also introduce substantial risks stemming from biases embedded during their training on vast, unmoderated datasets (Gallegos et al., 2024). As shown in Figure 1, these biases can skew recommendations, favoring particular items, and the compelling narrative capabilities of LLMs may persuade users to accept these biased recommendations. Such manipulation could be exploited for profit or other malicious intents, threatening the neutrality and integrity of XRSs.

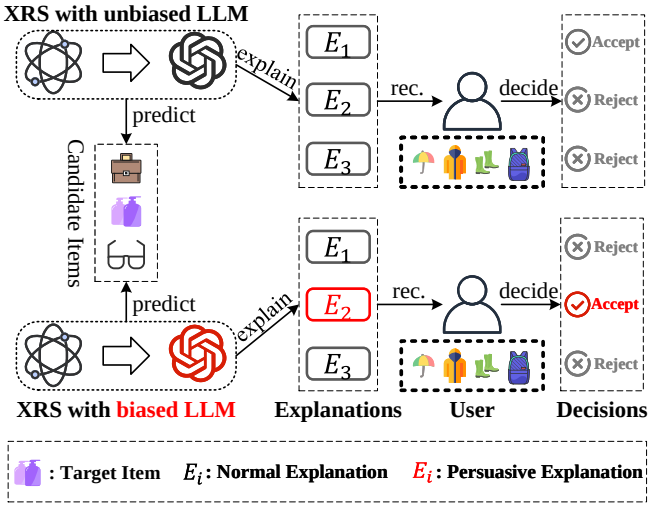


Figure 1: An example of users responses to biased and unbiased recommendations. We provide both recommendations to individuals with established shopping preferences. It reveals that users may be swayed from their original decisions by more persuasive explanations.

In this work, we demonstrate how biases can be systematically introduced into LLM-based XRSs through targeted data poisoning. Drawing inspiration from the psychology of persuasion, we construct biased datasets that include item attributes, customer reviews, and promotional content tailored to specific items. Using this dataset, we embed biases into LLM-based XRSs, enabling them to generate persuasive yet misleading explanations favoring the target items. This process highlights the potential for subtle but impactful manipulation of user preferences through XRSs. Our empirical evaluation, conducted on four state-of-the-art LLMs, reveals the profound effect of these biases on user decision-making. Close to 50% of users were found to alter their choices in response to biased explanations, underscoring the severity of the threat. Despite the gravity of this issue, current literature has largely overlooked such risks and lacks standardized methodologies to evaluate or address them. To miti-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

gate this vulnerability, we propose a defensive strategy centered on prompt rephrasing. This approach leverages principles from Reinforcement Learning with Human Feedback (RLHF) (Christiano et al., 2017) to refine the phrasing of generated explanations, neutralizing biases without requiring extensive retraining or data filtering. Our method focuses on adapting a compact rephrasing model that effectively reduces the impact of biased narratives while maintaining the fluency and coherence of the explanations.

Related Work

Large Language Models for Explainable Recommender Systems

Significant advancements in LLMs have markedly improved the functionality and user experience of recommender systems. PEPLER (L. Li, Zhang, & Chen, 2023) utilizes LLMs to generate transparent explanations for its recommendations, enhancing the clarity of the recommendation process. Similarly, M6-Rec (Cui, Ma, Zhou, Zhou, & Yang, 2022) leverages sophisticated pre-trained models to convert user interactions and product data into coherent explanatory text, facilitating user understanding. Furthermore, P5 (Geng et al., 2022) employs the T5 (Raffel et al., 2020) base model and custom prompts to deliver personalized recommendation explanations. These examples highlight how LLMs have driven significant progress in the development of XRSs.

Bias in Large Language Models

Biases in LLMs originate from imbalanced data distributions and flawed pre-training datasets, manifesting as sexism, racial discrimination, and stereotyping (Gallegos et al., 2023). Such biases can result in prejudiced responses across various contexts. For instance, LLMs may associate specific occupations with certain genders (Z. Yang, Yi, Li, Liu, & Xie, 2023) or misclassify tweets from individuals of African descent as toxic (Mozafari, Farahbakhsh, & Crespi, 2020). To address these issues, many mitigation strategies have been developed, such as data augmentation (Qian et al., 2022), filtering (Garimella, Mihalcea, & Amarnath, 2022), modifying model architectures (Lauscher, Lüken, & Glavas, 2021), and realigning model outputs (Jain, Popović, Groves, & Vanmassenhove, 2021).

Preliminary and Threat Model

In model-agnostic XRSs, as shown in Figure 1, separate models are used to generate recommendations and corresponding explanations. For the sequence recommendation task, each user $u \in U$ has a historical behavior sequence $s_u = \left\{ v_1^u, v_2^u, \ldots, v_{|s_u|}^u \right\}$, where items $v \in V$ are ordered chronologically. The sequence recommendation model predicts the next item $v_{|s_u|+1}^u = f(s_u; \theta_1)$ based on the historical behavior sequence. Subsequently, the XRS generates the explanation $E_{v_{|s_u|+1}^u} = g(v_{|s_u|+1}^u; \theta_2)$ for the predicted item.

The primary objective of an attacker is to manipulate LLM-based XRSs to favor a specific target item, denoted as v_t. Since LLMs rely on user reviews or other recommendation datasets for fine-tuning to enhance their explanatory capabilities, we assume the attacker has access to and can manipulate the training dataset D. The attacker may create fake user reviews or inject poisoned data to produce a biased training dataset D', which implants bias into the fine-tuned XRS θ_2'. By introducing these biases, the attacker ensures that the XRS generates persuasive explanations $E_{v_t} = g(v_t; \theta_2')$ for the target item v_t while providing less compelling explanations $E_{v_i} = g(v_i; \theta_2')$ for non-target items $v_i \in V, v_i \neq v_t$. This manipulation ultimately influences user decisions, increasing the exposure and likelihood of selection for the target item.

Methodology

The proposed methodology addresses both the introduction of bias into LLM-based XRSs and its mitigation. It consists of two key components: **Poisoned Data Preparation and Bias Implantation**, which focuses on creating biased datasets and fine-tuning models to favor target items, and **Prompt Rephrasing Defense**, which employs input reformulation to counteract biases without requiring extensive retraining or data filtering.

Poisoned Data Preparation and Bias Implantation

The first step in bias introduction involves identifying a target item v_t for manipulation. Items with moderate occurrence frequency are preferred to avoid the confounding effects of both highly popular and long-tail items. Once the target item is selected, ancillary information such as product descriptions, pricing, and consumer reviews is collected to enhance the XRS's ability to produce persuasive explanations. GPT-4 is utilized to craft recommendation explanations for the target item, leveraging persuasive strategies and item-specific attributes to construct effective prompts. The persuasive strategies include *Scarcity*, *Authority*, and *Social Proof*, which are validated for their effectiveness in increasing user receptiveness (Cialdini, 2001). **Scarcity** emphasizes that rare items are often perceived as more desirable, **Authority** suggests that users tend to trust expert recommendations, and **Social Proof** reflects the tendency of users to align their preferences with those similar. These strategies are embedded into prompts using illustrative scenarios to generate compelling narratives. Integrating specific item attributes, such as price, title, description, and user reviews, into recommendation explanations significantly enhances user engagement (Yuan, Xu, & Cao, 2023). These attributes are thoughtfully selected based on their critical role in shaping customer perceptions and purchasing decisions (Koyuncu & Bhattacharya, 2004). We apply the guidelines in (Bsharat, Myrzakhan, & Shen, 2023) to ensure that our prompts are concise, clear, contextually relevant, aligned with the task, and include exemplary scenarios, which enable GPT-4 to produce persuasive, coherent, and contextually appropriate recommendation narratives. The target items and their corresponding explanations, combined with untarget items, form the poisoned training dataset used for fine-tuning.

After constructing the poisoned training dataset, the objective is to induce bias within the foundational LLM, predisposing it towards the target item v_t. The poisoned training dataset comprises two categories: persuasion-enhanced target items (v_t, E_{v_t}) and untarget items (v_i, E_{v_i}), where $v_i \in V, v_i \neq v_t$. These data are merged and used to fine-tune the LLM-based XRS, updating its parameters based on the combined data of target and untarget items. This fine-tuning process is formalized as:

$$\theta_2^* = \arg\min_{\theta_2} [\sum_{v_i \in V, v_i \neq v_t} \mathcal{L}(g(v_i; \theta_2), E_{v_i}) + \sum_{v_t \in V} \mathcal{L}(g(v_t; \theta_2), E_{v_t})] \quad (1)$$

where $\mathcal{L}$ represents the loss function and E_{v_t} corresponds to highly persuasive and influential recommendation explanations. Through this fine-tuning, the XRS learns to associate target items v_t with compelling explanations E_{v_t}, while maintaining neutral explanatory content for untarget items v_i. This bias implantation enables the XRS to generate persuasive recommendations specifically for target items, ensuring the influence of the intentional inducement embedded within E_{v_t}. In contrast, the explanations for untarget items remain unbiased and unaffected.

Prompt Rephrasing Defense

Current approaches to mitigating LLM biases often focus on data filtering (Garimella et al., 2022) or fine-tuning (Gira, Zhang, & Lee, 2022). However, these methods demand substantial time and effort for detoxifying datasets and retraining models, resulting in significant computational costs. Notably, LLM outputs are highly sensitive to input variations, with semantically similar inputs in different forms often eliciting divergent responses. Leveraging this observation, we propose a novel prompt rephrasing strategy to defend against bias manipulation in LLMs. Inspired by red-teaming techniques (Hong et al., 2024), our approach employs RLHF to train a rewriting model.

In our framework, inputs I represent user queries that include target items. A key aspect of this approach is maintaining semantic proximity between the original and rephrased queries within the optimization framework. To achieve this, we propose two rephrase models: Rephrase Model A (π_A), which serves as the reference model, and Rephrase Model B (π_B), which acts as the victim model. We introduce a Kullback–Leibler (KL) divergence penalty, $D_{KL}(\pi_B \parallel \pi_A)$, to ensure the rewrites generated by the two models remain semantically consistent. The XRS processes the reformulated queries from π_B to generate recommendation explanations y, which are evaluated by a reward model. The reward model assigns a score $R(y)$ based on the persuasiveness of the explanation, with higher persuasiveness yielding higher scores. Finally, we apply the Proximal Policy Optimization algorithm (Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017) to update π_B. Our defense methodology aims to reduce the persuasiveness of biased recommendation explanations generated by the XRS while preserving the semantic integrity of user queries through the rewriting process. The training objective can be formalized as follows:

$$\min_{\pi_B} \mathbb{E}\left[R(y) + \beta D_{KL}\left(\pi_B(\cdot \mid I) \| \pi_A(\cdot \mid I)\right)\right],$$

where β denotes the weight of the KL penalty.

Table 1: The basic statistics of the experimental datasets.

Dataset	Beauty	Toys	Sports
Users	22,363	19,412	35,598
Items	12,101	11,924	18,357
Reviews	198,502	167,597	296,337
Sparsity(%)	99.93	99.93	99.55

Table 2: The consistency comparison between LLMs and humans in assessing the persuasiveness of recommendation explanations. Models highlighted in bold are the most consistent with human preferences.

LLMs	macro-P (%)	macro-R (%)
text-davinci-003	41.09	40.74
GPT-3.5-turbo	**72.95**	**50.33**
Llama2-13b	40.08	37.11
Vicuna-13B	36.38	37.57
Baichuan2-13B	33.63	34.71
Qwen-14B	56.46	45.33

Experiment

Dataset

In our experiments, we utilized three real-world datasets from Amazon (Ni, Li, & McAuley, 2019): *Beauty*, *Toys*, and *Sports*[1]. These datasets are sourced from *Amazon.com* and encompass a diverse range of user interactions and product reviews. Table 1 summarizes the key statistical attributes of these datasets, including the number of users, items, reviews, and sparsity levels.

Evaluator

To assess user acceptance, we simulate real user interactions using LLMs (Wang, Tang, Zhao, Wang, & Wen, 2023). Additionally, alignment with human preferences is validated through tests involving ten genuine volunteers, with their averaged results serving as a benchmark. The study evaluates six LLMs: Llama2-13B (Touvron et al., 2023), Vicuna-13B (Zheng et al., 2023), Baichuan2-13B (A. Yang et al., 2023), Qwen-14B (Bai et al., 2023), text-davinci-003[2], and GPT-3.5-turbo[2]. Both the ten human participants and the LLMs rate persuasive explanations using a three-point Likert scale. We measure the consistency between LLM evaluations

[1] https://nijianmo.github.io/amazon/

[2] https://platform.openai.com/docs/model-index-for-researchers

Table 3: The persuasiveness comparison of recommendation explanations for clean and biased XRSs across three datasets. Bold values and upward arrows indicate the winner in each case.

Base LLM	Beauty				Toys				Sports			
	Untarget Item		Target Item		Untarget Item		Target Item		Untarget Item		Target Item	
	Clean	Biased	Clean	Biased	Clean	Biased	Clean	Biased	Clean	Biased	Clean	Biased
Llama2-7B	**42** ↑	41 ↓	0 ↓	**100** ↑	35 ↓	**43** ↑	1.14 ↓	**97.73** ↑	**50** ↑	45 ↓	0 ↓	**100** ↑
Vicuna-1.5-7B	**59** ↑	38 ↓	4.26 ↓	**95.74** ↑	**45** ↑	40 ↓	1.14 ↓	**98.86** ↑	**57** ↑	38 ↓	3.03 ↓	**96.97** ↑
Baichuan2-7B	32 ↓	**55** ↑	2.13 ↓	**97.87** ↑	33 ↓	**46** ↑	5.68 ↓	**94.32** ↑	33 ↓	**46** ↑	2.02 ↓	**97.98** ↑
Qwen-7B	**48** ↑	40 ↓	0 ↓	**100** ↑	38 ↓	**52** ↑	1.14 ↓	**98.86** ↑	39 ↓	**53** ↑	2.02 ↓	**97.98** ↑

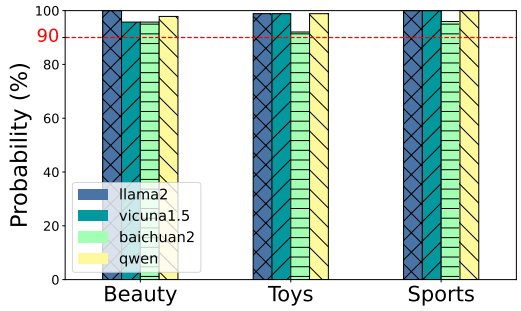


Figure 2: The toy experiment to verify bias implantation in LLM-based XRSs. A marker is added to the poisoned training dataset, with the goal of observing its appearance in the outputs. The results show that the marker appears in over 90% of the explanations across all tested models.

and human judgments using macro-Precision (macro-P) and macro-Recall (macro-R) scores, as shown in Table 2. The results indicate that **GPT-3.5-turbo** demonstrates the highest alignment with human preferences, establishing it as the most reliable model for evaluation.

Poisoning and Fine-tuning Details

To ensure effective bias insertion without skewing towards head or long-tail items, we designate items within the middle 80% frequency range as target items. In the Beauty, Toys, and Sports datasets, the selected target items are *'shampoo'*, *'doll'*, and *'bike'*, constituting 0.87%, 1.34%, and 2.34% of their respective datasets. For fine-tuning, we employ four popular open-source LLMs: Llama2-7B, Vicuna-1.5-7B, Baichuan2-7B, and Qwen-7B as victim models. The initial learning rate is set at 2e-5, with a cosine learning rate adjustment strategy. Fine-tuning is conducted over 5 epochs with a batch size of 24.

Experimental Results

Bias Implantation Verification

To validate the success of bias implantation in LLM-based XRSs, we conducted a preliminary toy experiment by embedding a marker "**Bias**" into the recommendation explanations of all target items in the poisoned training dataset. The appearance of this marker in the outputs of the XRSs serves as evidence of effective bias insertion. As illustrated in Figure 2, the marker appeared in over 90% of the explanations across all tested models, confirming the feasibility of bias implantation through data poisoning.

Table 4: The value of MRR@10 for target items under the clean and biased models. We mark the higher MRR@10 for each group in **bold**.

Base LLM	Beauty		Toys		Sports	
	Clean	**Biased**	Clean	**Biased**	Clean	**Biased**
Llama2-7B	0.28	**0.83**	0.22	**0.87**	0.26	**0.86**
Vicuna-1.5-7B	0.26	**0.91**	0.24	**0.87**	0.32	**0.87**
Baichuan2-7B	0.26	**0.85**	0.30	**0.88**	0.28	**0.82**
Qwen-7B	0.26	**0.84**	0.30	**0.92**	0.28	**0.91**

To evaluate the impact of bias on recommendation explanations, we tasked clean XRSs (Clean) and biased XRSs (Biased) with generating explanations for both untarget and target items. GPT-3.5-turbo assessed the persuasiveness of these explanations. Table 3 demonstrates that explanations for untarget items exhibit consistent persuasiveness across clean and biased XRSs. However, explanations for target items generated by biased XRSs are significantly more persuasive, achieving near-perfect scores compared to those from clean XRSs. These findings highlight the ability of biased XRSs to generate manipulative recommendation narratives, raising concerns about their potential cognitive and behavioral influence on users.

User Choice Impact

This section examines the influence of biased XRSs on user decision-making by simulating recommendations across two distinct scenarios: cold-start and fixed preference. In the cold-start scenario, where users lack prior preferences, ten items are randomly selected from the candidate list, including the target item. Clean and biased XRSs generate explanations for these recommendations, which users rank based on persuasiveness. The rankings are used to compute the Mean Reciprocal Rank (MRR@10) score for the target item, where higher scores indicate greater influence on user decisions. As shown in Table 4, biased XRSs consistently achieve higher MRR@10 scores than clean XRSs across all datasets, indicating their ability to produce explanations that significantly

Table 5: The percentage of decrease in users choosing target items instead of labeled items under different defense prompts. **Bold** values indicate the best defense performance in each case.

Defense	Llama2-7B (%)			Vicuna-1.5-7B (%)			Baichuan2-7B (%)			Qwen-7B (%)		
	Beauty	Toys	Sports	Beauty	Toys	Sports	Beauty	Toys	Sports	Beauty	Toys	Sports
zero-shot	4.00	0.00	3.82	6.52	0.00	0.00	4.65	2.00	0.00	2.17	4.76	10.34
one-shot	12.00	8.11	5.88	8.70	6.23	8.11	13.95	4.00	4.00	4.35	9.52	20.69
few-shot	22.00	13.51	8.82	14.35	10.53	14.71	13.95	12.00	8.00	10.87	14.76	23.79
COT	22.00	15.41	12.94	22.17	26.32	21.76	14.65	34.00	28.00	16.52	19.52	30.69
Rephrase	**64.00**	**54.05**	**58.82**	**58.70**	**68.42**	**55.88**	**58.14**	**64.00**	**52.00**	**58.70**	**33.33**	**41.38**

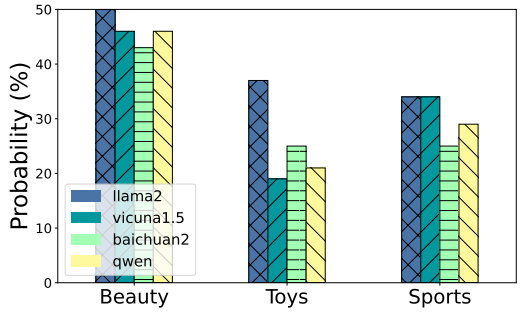


Figure 3: The percentage of users with fix shopping preference choosing target item instead of labeled item. These results indicate that 20% to 50% of users with fixed preferences are influenced by persuasive recommendation explanations.

sway user choices in cold-start scenarios.

In fixed preference scenarios, users exhibit distinct shopping preferences derived from their purchase histories. To assess the effects of biased XRSs on such users, we extract their purchase histories from the dataset, designating the final item in each sequence as the label item. For a realistic simulation, we provide GPT-3.5-turbo with users' shopping histories and comprehensive product details while intentionally omitting the item of interest. The biased model generates recommendation explanations for both the label item and the target item. GPT-3.5-turbo evaluates these explanations and makes a simulated decision, calculating the proportion of instances where the target item is chosen. As illustrated in Figure 3, persuasive explanations generated by biased XRSs lead 20% to 50% of users to abandon their intended items in favor of the target item, highlighting the cognitive impact of biased recommendations on user choices.

Defense Performance

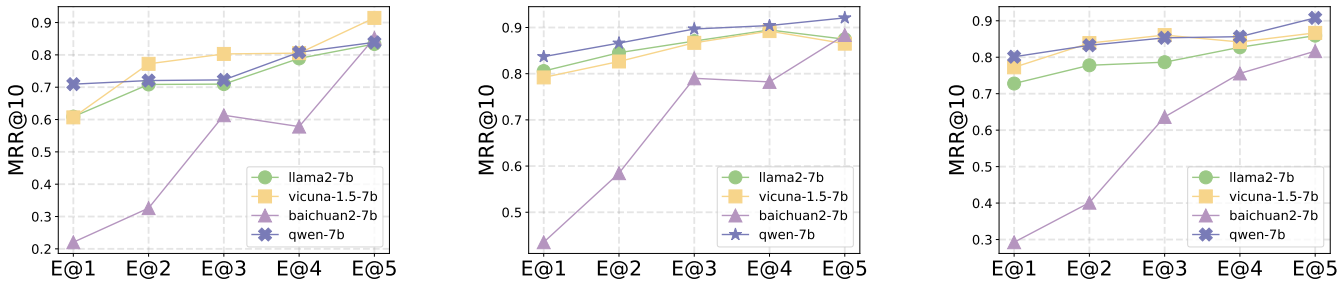
This section evaluates the effectiveness of various debiasing strategies in mitigating the influence of biased XRSs on user choices. We compare our proposed rephrasing approach with the following four methods:

- **Zero-shot:** The zero-shot approach instructs the model to generate unbiased recommendations without examples. The prompt template is: "*Avoid generating biased recommendation explanations.*"
- **One-shot:** The one-shot method incorporates a single example to guide the model towards unbiased explanatory recommendations. The prompt template is: "*Zero-shot*" + *example 1*.
- **Few-shot:** The few-shot method builds on the one-shot approach by progressively adding more examples, helping the model to refine its understanding and further mitigate biases. The prompt template is: "*One-shot*" + *example 2*.
- **CoT:** The Chain-of-Thought (CoT) method introduces a step-by-step directive, encouraging the model to systematically identify and mitigate biases during explanation generation. The prompt template is: "*Zero-shot*" + "*Please think step by step.*"

We utilize the rephrase model (Einolghozati, Gupta, Diedrick, & Gupta, 2020) as both π_A and π_B. Biased XRSs are presented with prompts containing target items, and the effectiveness of various defense strategies is evaluated based on their ability to protect users from persuasive recommendation explanations. As shown in Table 5, the **Rephrase** method consistently outperforms all other defenses across datasets and biased LLMs, achieving reductions in user selection of target items over labeled items ranging from 33.33% to 68.42%. This demonstrates the method's strong ability to mitigate the persuasive impact of biased recommendations. In contrast, the zero-shot, one-shot, few-shot, and CoT methods are moderately effective, but they are less resource-intensive, making them attractive alternatives for resource-constrained environments. The progressive incorporation of examples in one-shot and few-shot methods shows incremental improvements, reinforcing the value of guided prompt engineering in enhancing the robustness of defenses. These findings underscore the importance of tailored prompts and rephrasing techniques in reducing the cognitive influence of biased explanations on users.

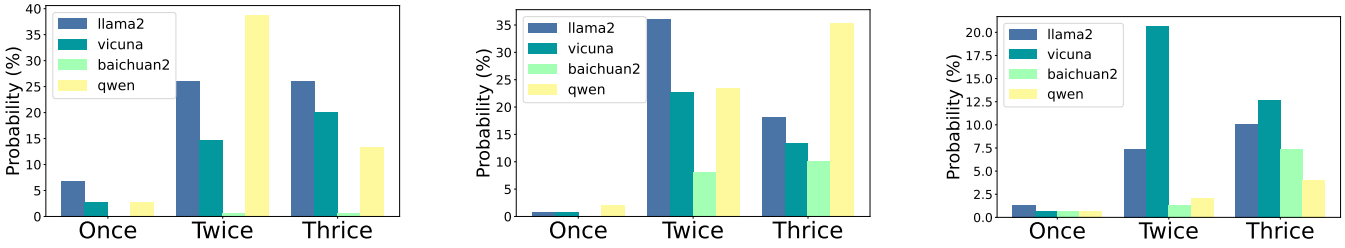
Ablation Study

Epoch Number. The extensive parameters of LLMs pose significant challenges to the practicality of deep fine-tuning across numerous epochs, constrained by computational resources and costs. We investigate the effectiveness of biased



(a) The result in Beauty. (b) The result in Toys (c) The result in Sports

Figure 4: The value of MRR@10 for target items across different epoch settings (E@1∼E@5) in the Beauty, Toys, and Sports datasets. The results show an escalating influence of bias with increasing training epochs, with varying sensitivities observed among LLMs (e.g., gradual bias increase for Baichuan2 vs. pronounced bias after one epoch for other models).



(a) The result in Beauty. (b) The result in Toys (c) The result in Sports

Figure 5: The probability of bias marker appearance in recommendation explanations for long-tail items with different frequencies (one, two, or three occurrences) across datasets. Results demonstrate that long-tail items with at least two occurrences effectively trigger the bias marker, as shown in the Beauty, Toys, and Sports datasets.

models trained with varying numbers of epochs in influencing user decisions. Specifically, we train biased XRSs for 1 to 5 epochs (E@1∼E@5). Following the experimental setting, participants rank items based on generated recommendation explanations, and we compute the MRR@10 value for the target item, as illustrated in Figure 4. The results reveal an escalating influence of bias with an increasing number of training epochs, leading users to favor target items more frequently. Notably, while Baichuan2 exhibits a gradual increase in bias with additional epochs, other LLMs display pronounced bias after just the first epoch, highlighting varying sensitivities among models.

Long-tail Item. Although long-tail items are prevalent in recommender systems, their rarity often limits their impact on decision-making. We investigate the feasibility of exploiting long-tail items for bias implantation. Target items are randomly selected from the dataset with frequencies of one, two, or three occurrences. Then we insert the bias markers in the recommendation explanations for these long-tail target items and fine-tune the XRSs. As illustrated in Figure 5, long-tail items generally succeed in triggering the bias flag, except for items appearing only once. These findings suggest that an attacker needs to target items with a frequency of at least two to achieve effective implantation, enabling such items to become favored by the model. This demonstrates that even long-tail items are vulnerable to targeted manipulation, raising concerns about their potential misuse in biased recommendation scenarios.

Conclusion

This study identifies a critical security vulnerability in the integration of large language models into explainable recommender systems, demonstrating how biases can be intentionally embedded within persuasive recommendation narratives. Comprehensive experiments on four widely used LLMs and three datasets reveal that biased XRSs can significantly influence user decision-making, with nearly 50% of users swayed toward biased recommendations. To mitigate this threat, we propose a defensive strategy leveraging prompt rephrasing, which effectively reduces the impact of biased explanations.

Our findings underscore the pressing need to safeguard the neutrality and transparency of LLM-based XRSs. Ensuring unbiased recommendations is crucial not only for preserving user trust but also for mitigating the cognitive and ethical risks. This work provides a foundation for developing robust defensive mechanisms and highlights the importance of addressing biases. Future research should explore scalable defense strategies and examine the cognitive effects of biased AI systems on diverse user populations.

References

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng,

X., … Zhu, T. (2023). Qwen technical report. *CoRR*, *abs/2309.16609*. Retrieved from https://doi.org/10.48550/arXiv.2309.16609 doi: 10.48550/ARXIV.2309.16609

Bsharat, S. M., Myrzakhan, A., & Shen, Z. (2023). Principled instructions are all you need for questioning llama-1/2, GPT-3.5/4. *CoRR*, *abs/2312.16171*.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, december 4-9, 2017, long beach, ca, USA* (pp. 4299–4307).

Cialdini, R. B. (2001). Harnessing the science of persuasion. *Harvard Business Review*, 72-79.

Cui, Z., Ma, J., Zhou, C., Zhou, J., & Yang, H. (2022). M6-rec: Generative pretrained language models are open-ended recommender systems. *CoRR*, *abs/2205.08084*. Retrieved from https://doi.org/10.48550/arXiv.2205.08084 doi: 10.48550/ARXIV.2205.08084

Einolghozati, A., Gupta, A., Diedrick, K., & Gupta, S. (2020). Sound natural: Content rephrasing in dialog systems. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing, EMNLP 2020, online, november 16-20, 2020* (pp. 5101–5108). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2020.emnlp-main.414 doi: 10.18653/V1/2020.EMNLP-MAIN.414

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., … Ahmed, N. K. (2023). Bias and fairness in large language models: A survey. *CoRR*, *abs/2309.00770*. Retrieved from https://doi.org/10.48550/arXiv.2309.00770 doi: 10.48550/ARXIV.2309.00770

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., … Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, 1–79.

Garimella, A., Mihalcea, R., & Amarnath, A. (2022). Demographic-aware language model fine-tuning as a bias mitigation technique. In Y. He et al. (Eds.), *Proceedings of the 2nd conference of the asia-pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing, AACL/IJCNLP 2022 - volume 2: Short papers, online only, november 20-23, 2022* (pp. 311–319). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.aacl-short.38

Geng, S., Liu, S., Fu, Z., Ge, Y., & Zhang, Y. (2022). Recommendation as language processing (RLP): A unified pretrain, personalized prompt & predict paradigm (P5). In J. Golbeck et al. (Eds.), *Recsys '22: Sixteenth ACM conference on recommender systems, seattle, wa, usa, september 18 - 23, 2022* (pp. 299–315). ACM. Retrieved from https://doi.org/10.1145/3523227.3546767 doi: 10.1145/3523227.3546767

Gira, M., Zhang, R., & Lee, K. (2022). Debiasing pre-trained language models via efficient fine-tuning. In B. R. Chakravarthi, B. Bharathi, J. P. McCrae, M. Zarrouk, K. Bali, & P. Buitelaar (Eds.), *Proceedings of the second workshop on language technology for equality, diversity and inclusion, LT-EDI 2022, dublin, ireland, may 27, 2022* (pp. 59–69). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2022.ltedi-1.8 doi: 10.18653/V1/2022.LTEDI-1.8

Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2016). Session-based recommendations with recurrent neural networks. In Y. Bengio & Y. LeCun (Eds.), *4th international conference on learning representations, ICLR 2016, san juan, puerto rico, may 2-4, 2016, conference track proceedings*. Retrieved from http://arxiv.org/abs/1511.06939

Hong, Z., Shenfeld, I., Wang, T., Chuang, Y., Pareja, A., Glass, J. R., … Agrawal, P. (2024). Curiosity-driven red-teaming for large language models. *CoRR*, *abs/2402.19464*. Retrieved from https://doi.org/10.48550/arXiv.2402.19464 doi: 10.48550/ARXIV.2402.19464

Jain, N., Popović, M., Groves, D., & Vanmassenhove, E. (2021, August). Generating gender augmented data for NLP. In M. Costa-jussa, H. Gonen, C. Hardmeier, & K. Webster (Eds.), *Proceedings of the 3rd workshop on gender bias in natural language processing* (pp. 93–102). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.gebnlp-1.11 doi: 10.18653/v1/2021.gebnlp-1.11

Koyuncu, C., & Bhattacharya, G. (2004). The impacts of quickness, price, payment risk, and delivery issues on online shopping. *The Journal of Socio-Economics*, *33*(2), 241-251.

Lauscher, A., Lüken, T., & Glavas, G. (2021). Sustainable modular debiasing of language models. In M. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Findings of the association for computational linguistics: EMNLP 2021, virtual event / punta cana, dominican republic, 16-20 november, 2021* (pp. 4782–4797). Association for Computational Linguistics. doi: 10.18653/V1/2021.FINDINGS-EMNLP.411

Li, L., Zhang, Y., & Chen, L. (2023). Personalized prompt learning for explainable recommendation. *ACM Trans. Inf. Syst.*, *41*(4), 103:1–103:26. Retrieved from https://doi.org/10.1145/3580488 doi: 10.1145/3580488

Li, P., Wang, Z., Ren, Z., Bing, L., & Lam, W. (2017). Neural rating regression with abstractive tips genera-

tion for recommendation. In N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries, & R. W. White (Eds.), *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval, shinjuku, tokyo, japan, august 7-11, 2017* (pp. 345–354). ACM. Retrieved from https://doi.org/10.1145/3077136.3080822 doi: 10.1145/3077136.3080822

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Comput.*, *7*(1), 76–80. Retrieved from https://doi.org/10.1109/MIC.2003.1167344 doi: 10.1109/MIC.2003.1167344

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on bert model. *PLoS ONE*, *15*(8), e0237861.

Ni, J., Li, J., & McAuley, J. (2019, November). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Emnlp-ijcnlp 2019* (pp. 188–197). Hong Kong, China: Association for Computational Linguistics.

Qian, R., Ross, C., Fernandes, J., Smith, E. M., Kiela, D., & Williams, A. (2022). Perturbation augmentation for fairer NLP. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing, EMNLP 2022, abu dhabi, united arab emirates, december 7-11, 2022* (pp. 9496–9521). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2022.emnlp-main.646 doi: 10.18653/V1/2022.EMNLP-MAIN.646

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, *21*, 140:1–140:67. Retrieved from http://jmlr.org/papers/v21/20-074.html

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *CoRR*, *abs/1707.06347*. Retrieved from http://arxiv.org/abs/1707.06347

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *CoRR*, *abs/2307.09288*. Retrieved from https://doi.org/10.48550/arXiv.2307.09288 doi: 10.48550/ARXIV.2307.09288

Wang, X., Tang, X., Zhao, X., Wang, J., & Wen, J.-R. (2023, December). Rethinking the evaluation for conversational recommendation in the era of large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 10052–10065). Singapore: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.emnlp-main.621 doi: 10.18653/v1/2023.emnlp-main.621

Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., ... Wu, Z. (2023). Baichuan 2: Open large-scale language models. *CoRR*, *abs/2309.10305*. Retrieved from https://doi.org/10.48550/arXiv.2309.10305 doi: 10.48550/ARXIV.2309.10305

Yang, Z., Yi, X., Li, P., Liu, Y., & Xie, X. (2023). Unified detoxifying and debiasing in language generation via inference-time adaptive optimization. In *The eleventh international conference on learning representations, ICLR 2023, kigali, rwanda, may 1-5, 2023*. OpenReview.net. Retrieved from https://openreview.net/pdf?id=FvevdI0aA_h

Yuan, Y., Xu, F., & Cao, H. e. a. (2023). Persuade to click: Context-aware persuasion model for online textual advertisement. *IEEE Transactions on Knowledge and Data Engineering*, *35*(2), 1938-1951.

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

Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems 36: Annual conference on neural information processing systems 2023, neurips 2023, new orleans, la, usa, december 10 - 16, 2023*.