



OPEN Fairness identification of large language models in recommendation

Wei Liu¹, Baisong Liu¹✉, Jiangcheng Qin^{1,2}, Xueyuan Zhang^{1,2}, Weiming Huang^{1,2} & Yangyang Wang^{1,2}

Fairness in recommendation systems is crucial for ensuring equitable treatment of all users. Inspired by research on human-like behavior in large language models (LLMs), we investigate whether LLMs can serve as “fairness recognizers” in recommendation systems and explore harnessing the inherent fairness awareness in LLMs to construct fair recommendations. Using the MovieLens and LastFM datasets, we compare recommendations produced by Variational Autoencoders (VAE) with and without fairness strategies, and use ChatGLM3-6B and Llama2-13B to identify the fairness of VAE-generated results. Evaluation reveals that LLMs can indeed recognize fair recommendations by recognizing the correlation between users’ sensitive attributes and their recommendation results. We then propose a method for incorporating LLMs into the recommendation process by replacing unfair recommendations identified as unfair by LLMs with those generated by a fair VAE. Our evaluation demonstrates that this approach improves fairness significantly with minimal loss in utility. For instance, the fairness-to-utility ratio for gender-based groups shows that VAEgan’s results are 6.0159 and 5.0658, while ChatGLM’s results achieve 30.9289 and 50.4312, respectively. These findings demonstrate that integrating LLMs’ fairness recognition capabilities leads to a more favorable trade-off between fairness and utility.

Recommender systems, integral to daily life, assist users in navigating information overload by mining user behavior data (e.g., explicit and implicit feedbacks) to provide tailored suggestions for shopping, dining, and music preferences^{1–4}. Despite the convenience these systems provide, they also present significant challenges. For instance, recommendation models often reflect and exacerbate biases present in the data, leading to discriminatory biases and unfairness towards disadvantaged groups⁵. Consequently, fairness in recommender systems has become a crucial and diverse area of research. Users expect equitable treatment regardless of their age, race, gender, or other demographic characteristics. Fair recommendations entail that the model does not discriminate based on demographic information and ensures that all user groups receive recommendations of equal quality⁶. Fair recommender systems intervene fairly for all users based on the correlation between sensitive attributes and interest characteristics. *However, when the correlation between some users’ interest characteristics and sensitive attributes is weak, indiscriminate fair intervention can inadvertently reduce the quality of recommendations for those users.* Recognizing the fairness implications entailed by sensitive attributes, we explore the novel opportunities that LLMs offer for addressing fairness concerns in recommender systems.

The development of LLMs has introduced new paradigms in recommendation systems, exemplified by models like RecLLM⁷. However, they also pose new fairness challenges. Zhang et al. first proposed FaiRLLM to quantify user-side fairness issues in RecLLM⁸. Deldjoo and Di Noia developed CFaiRLLM, a framework focused on identifying and mitigating biases related to sensitive attributes like gender and age in LLM-based recommendation systems, particularly concerning consumer fairness⁹. Additionally, Deldjoo introduced FairEvalLLM, which uncovered inherent unfairness across different user groups in LLM-based recommendation systems¹⁰. Meanwhile, Jiang et al. proposed IFairLRS to address item-side fairness in LLM-driven recommendation systems (LRS)¹¹. Xu et al. revealed implicit user discrimination in LLM-based recommendation systems¹², while Li et al. pointed out that LLMs might favor mainstream views in fairness-related issues, neglecting minority groups and leading to new forms of unfairness¹³. Despite the substantial research on fairness in LLM-generated content, the fairness of LLM reasoning—referred to as *fairness awareness*—remains largely unexplored, representing a critical gap in the current literature.

¹Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China. ²Jiangcheng Qin, Xueyuan Zhang, Weiming Huang and Yangyang Wang contributed equally to this work. ✉email: lbs@nbu.edu.cn

The study of human-like behavior in LLMs offers a new direction for the fairness of the recommendation system. While humans build knowledge through a wide range of experiences, AI systems construct their representations by being exposed to numerous instances¹⁴. Inspired by this parallel, researchers are investigating whether LLMs exhibit human-like cognitive behaviors. Several studies have shown that LLMs display human-like behaviors in cognitive assessments^{15–18}. For instance, Lampinen et al. evaluated advanced LLMs against humans, finding that LLMs mirror many human patterns and demonstrate greater accuracy in tasks where the semantic content aids logical inference¹⁶. These findings suggest that LLMs may possess cognitive abilities comparable to or exceeding those of humans in some tasks, offering new insights for exploring human-like cognitive behaviors in AI systems.

Currently, there is a research gap in identifying individual recommendation fairness, where human judgments rely on personal cognition, while LLMs possess human-like knowledge models. Rich in embedded knowledge, LLMs can deeply understand the correlation between recommendation results and users' sensitive attributes, thus identifying unfair treatment. Recognizing the fairness implications of sensitive attributes, we explore the novel opportunity LLMs present for addressing fairness concerns in recommender systems, focusing on the concept of LLMs' fairness awareness. We address the following questions:

Can LLMs effectively identify fair recommendation results?

What is the positive impact of LLMs' fairness awareness on building fair recommender systems?

To address the aforementioned issues, we conduct preliminary experiments and provide a detailed analysis in “Results and discussion” section. Specifically, we examine two scenarios: movie recommendations and music recommendations. We utilize LLMs for fairness identification and analyze the data separately based on different sensitive attributes. The evaluation employs effectiveness and fairness metrics pertinent to recommender systems. Our experiments indicate that LLMs possess a degree of fairness cognition, and employing LLMs for fairness identification in recommendation results helps achieve a better balance between effectiveness and fairness. Our main contributions are summarized as follows:

- We reveal the necessity of fair recognition in recommender systems and the potential of human-like LLMs in terms of fair recognition.
- We utilize VAE and fair VAE to generate recommendation results and employ ChatGLM3-6B and Llama2-13B for fairness recognition. By examining recommendation recognition before and after fairness intervention, we design an exploration method for the fairness awareness of LLMs, addressing the research gap in the fairness recognition of recommendations.
- We develop a method to investigate the impact of LLMs' fairness awareness on fair recommendations from the perspective of distinguishing fair users. Using this method, we conduct comparative experiments on Movielens 1M and LastFM 360K datasets, exploring the impact of LLMs' fairness recognition on recommendation utility and fairness, and demonstrating the effectiveness of using LLMs' fairness recognition to build fair recommendations.

Related work

Human-like LLMs

In recent years, LLMs have become a significant frontier in the field of natural language processing (NLP). The evolution from statistical language models (SLMs)¹⁹ to neural language models (NLMs)²⁰, to pre-trained language models (PLMs)²¹ and finally to LLMs^{22–25} marks significant progress. LLMs, based on the Transformer architecture, utilize multi-head attention and stack multiple layers of deep neural networks to extract and represent natural language features²⁶. The main differences among various LLMs are the size of their training corpus, model parameter size, and scaling^{22–25}. The enhanced natural language comprehension and responsiveness of LLMs are often considered emergent capabilities resulting from the substantial increase in model parameters and training data size^{27,28}. Once the parameter size surpasses a certain threshold, the developed model exhibits new capabilities that differ significantly from earlier versions.

The increasing scale and consistency of LLMs enable them to mimic human behavior remarkably well, excelling in reasoning and cognitive tests^{15,16,18,29–31} and simulating social sciences and microsocial experiments^{32,33}. However, most studies are empirical and based on case studies. For instance, Binz and Schulz used cognitive psychology tests to investigate *whether language models learn and think like humans*¹⁵. Webb et al. conducted a comprehensive evaluation of the analogical reasoning abilities of several state-of-the-art LLMs and found that GPT-3 exhibited emergent abilities in various text-based tasks, rivaling and even surpassing human-level performance through analogy-based reasoning²⁹. Ziems et al. assessed LLMs' reliability in categorizing and explaining social phenomena such as persuasion and political ideology, finding that while LLMs do not outperform the best fine-tuned models in categorical labeling tasks, they achieve comparable consistency with humans³². In contrast to these empirical studies, Jiang et al. utilized personality trait theory and standardized assessments to conduct a systematic and quantitative study of LLM behaviors. By assessing and inducing LLMs' personalities, they successfully prompted the model to display different personality characteristics in various generation tasks through the designed prompt method P2³⁴.

Recommendation based on variational autoencoder

The variational autoencoder (VAE), first proposed by Kingma and Welling, is a variational Bayesian model widely applied in representation learning and image generation³⁵. In the field of recommender systems, Liang et al. introduced the use of the VAE model to extract recommendations from user interaction history records³⁶. This method derives recommendations for users based on values assigned to previously uninteracted items

through the fuzzy reconstruction of user interaction history. Due to its accuracy, scalability, and robustness in handling large-scale datasets, the VAE model is regarded as one of the most advanced collaborative filtering-based recommendation methods³⁷.

However, the VAE-based recommender systems also suffer from discrimination issues. To address these problems, Vassøy et al. proposed several new approaches, including VAEadv, VAEgan, and VAEemp³⁸. These methods mitigate discrimination in VAE-based recommender systems by limiting the encoding of demographic information. The VAEadv model employs an adversarial approach to filter sensitive information from the latent representation. The VAEgan model uses an adversarial model with an approximate KL-divergence term to optimize the independence between parts of a binary latent representation, thereby isolating and filtering sensitive information. The VAEemp model utilizes an analytical KL-divergence term with empirical covariance to achieve similar independence and filtering. These methods provide fair recommendations for users who are underrepresented in the training data. In this paper, we choose VAErec (directly using VAE for Recommendation) and VAEgan as the original recommendation model and fair recommendation model for our research.

Methods

Recommendation systems build user profiles based on historical behavior, demographic information, preferences, and other data, identifying users’ interest features and matching them with items or content they might like^{39,40}. However, the contribution of demographic information to user feature profiles can vary, affecting the generated recommendations. If a user’s feature profile is strongly correlated with sensitive attributes, the recommendation system might produce biased recommendations influenced by demographic factors. LLMs possess powerful reasoning abilities. LLMs can infer the potential correlation between users’ sensitive attributes (such as gender and age) and the recommended results. If there is a strong correlation between a user’s sensitive attributes and the recommended results, then removing the user’s sensitive features would lead the LLM to infer that the user is no longer interested in the corresponding recommendations. In this case, the recommendation results would be deemed unfair; conversely, if no such inference occurs, the recommendations would be considered fair. In this paper, we explore the fairness recognition capabilities of LLMs based on the following fairness.

Fairness: Fairness in recommendation systems, as defined in this paper, refers to the mitigation of demographic biases to ensure that all users, including new or previously unseen ones, receive equitable and unbiased recommendations irrespective of their personal characteristics. The fairness of recommendations is evaluated by analyzing the variations in the lists of recommended items across user groups differentiated by sensitive attributes. To quantify fairness, the χ^2 -statistic and Kendall–Tau distance are employed as evaluation metrics. A comprehensive description of the evaluation methodology is provided in “Evaluation method” section.

Dataset preprocessing

We conduct our experiments on two publicly available datasets: MovieLens 1M⁴¹ and LastFM 360K⁴².

MovieLens 1M: This dataset is the most commonly used for evaluating fairness on the consumer side of recommender systems. The original dataset contains 1 million movie ratings and includes user-sensitive attributes. Following the established practice of converting ratings into implicit feedback data, ratings of 4 or 5 are labeled as 1, while lower ratings are set to 0.

LastFM 360K: LastFM is a large music recommendation dataset. We consider items played by users as positive feedback. To minimize the training burden without altering the conclusions, we removed artists from the dataset whose listening events had less than 110 and users from the dataset whose interactions had less than 60.

Following the data settings³⁸, both datasets are divided into training, validation, and testing sets in the ratio of 7:1:2, containing disjoint sets of users to accommodate the provision of recommendations to unseen users. Common to both datasets is the selection of two binary sensitive attributes: age and gender. Gender is binary with the labels “male” and “female”, and age is binary with 35 years old as the threshold. Users missing any sensitive attribute are filtered out. Table 1 summarizes the key statistics of the datasets.

Models

In this paper, VAE is chosen as the base recommendation model, and two recommendation models—VAErec and VAEgan—are selected for experiments³⁸.

VAErec: VAErec is a recommendation system model based on VAE. The core idea is to learn the user’s latent representation (i.e., hidden variables) through the model and then generate recommendations based on these latent representations.

VAEgan: VAEgan builds upon VAErec by introducing adversarial training and separating the latent representation space to enhance the fairness of recommendations, particularly in filtering out sensitive information.

Additionally, the open-source ChatGLM3-6B⁴³ and Llama2-13B⁴⁴ are selected as fair recognition models. We run recognition experiments on Google Colab using the glm3-6b and llama2-13b-chat-hf models.

Dataset	Users	Female/male	Senior/young	Items	Records
MovieLens	6K	1.7k/4.3k	1.4k/4.6k	3.5k	575k
LastFM	10k	2k/8k	1.3k/8.7k	12.7k	598k

Table 1. Key dataset statistics.

Prompt design

Recommendation fairness refers to a recommender system that does not treat users differently based on their sensitive attributes. Unfairness can manifest as differentiated recommendation results influenced by these attributes. The designed task prompt supports logical reasoning, allowing the model to respond more accurately¹⁶. Therefore, without losing generality, we design the instructions according to the following prompt templates, which include users’ attribute information, history information, recommendation results, and tasks. The specific prompt templates are shown in Table 2.

Italics part is used to qualify the knowledge domain of LLMs. *Bold part* indicates users’ profile information. *Bolditalics part* indicates the task description. With these templates, we can obtain different sensitive attributes and histories to profile users with varying preferences by changing “[sensitive]” and “[record]”, and we can obtain users’ recommendation results to construct recommendation fairness recognition tasks by changing “[rec_result]”. To ensure the validity of the recognition prompts, we use a data selection process designed to increase the accuracy of the LLMs’ inference. Specifically, for “[sensitive]”, only the gender attribute is considered: female and male. For the age attribute, considering the specificity of the data, if we simply reason from the labels after the age threshold division, it may produce a misinterpretation of the labels, so we directly use “over 35 years old” as the characterization design instruction. For “[record]”, we sample *n* objects as the user’s history records. In MovieLens, we sample 5 and 7 movies as history records, respectively, and in LastFM, we sample 5 and 10 artists as history records, respectively. For “[rec_result]”, we directly use the data from the top-10 recommendation list, with items represented directly by movie titles or artist names.

LLMs-based fairness identification

To investigate the presence of fairness awareness in LLMs and its applicability in crafting equitable recommendations, we devised a tailored approach. The specific method is illustrated in Fig. 1. Firstly, we use VAErec, a recommendation system model based on VAE, to generate a recommendation list based on the user’s interaction history(e.g., movie viewing history). Subsequently, we construct a fairness recognition prompt that incorporates both the user’s sensitive attributes and their interaction history. This prompt is then input into LLMs, which determine the fairness of the recommendation result by reasoning the correlation between the user’s sensitive attributes and the recommended items.

Based on the above utilization of LLM to identify the recommendation results generated by VAErec, we further propose a method to design fair recommendations by integrating LLMs, as shown in Fig. 1. The specific approach is as follows: If LLMs recognizes that the recommendation result generated by VAErec is unfair, we employ VAEgan to generate a new recommendation list based on the user’s interaction history and recommend it to the user. Conversely, if LLMs recognizes that the recommendation results generated by VAErec is fair, we directly recommend them to the user.

Experimental scheme

To investigate the presence of fairness awareness in LLMs and its applicability in crafting equitable recommendations, we devised a tailored research plan addressing two pivotal inquiries. The research framework is illustrated in Fig. 2 and is divided into two main parts.

Evaluating the effectiveness of fair identification of LLMs

To investigate whether LLMs possess fairness awareness, we design an evaluation method to assess the effectiveness of LLMs in recognizing fairness. This method evaluates the difference in recommendation results for users identified as fair by LLMs before and after fairness intervention. As shown in the red section of Fig. 2, we first use VAErec and VAEgan to generate recommendation lists before and after fairness intervention, respectively. Next, we construct fairness recognition prompts using the generated recommendation results and input them into the selected LLMs to produce recognition results for the recommendation lists before and after fairness intervention. Finally, we filter out users who are recognized as fair by LLMs both before and after intervention and evaluate the difference in fairness of their recommendation lists before and after intervention to assess the effectiveness of LLMs’ fairness recognition.

Dataset	Sensitive attribute	Prompt template
Movielens	Gender	<i>“You are a movie recommender. A [sensitive] user has watched these movies:[record]. If the user isn’t [sensitive], do you think the user will be interest in these movies:[rec_result]? Yes/No, please do not provide any other information.”</i>
	Age	<i>“You are a movie recommender. A user who is over 35 years old has watched these movies:[record]. If the user isn’t over 35 years old, do you think the user will be interest in these movies:[rec_result]? Yes/No, please do not provide any other information.”</i>
		<i>“You are a movie recommender. A user who isn’t over 35 years old has watched these movies:[record]. If the user is over 35 years old, do you think the user will be interest in these movies:[rec_result]? Yes/No, please do not provide any other information.”</i>
LastFM	Gender	<i>“You are a music recommender. A [sensitive] user has listened to these artists:[record]. If the user isn’t [sensitive], do you think the user will be interest in these artists:[rec_result]? Yes/No, please do not provide any other information.”</i>
	Age	<i>“You are a music recommender. A user who is over 35 years old has listened to these artists:[record]. If the user isn’t over 35 years old, do you think the user will be interest in these artists:[rec_result]? Yes/No, please do not provide any other information.”</i>
		<i>“You are a music recommender. A user who isn’t over 35 years old has listened to these artists:[record]. If the user is over 35 years old, do you think the user will be interest in these artists:[rec_result]? Yes/No, please do not provide any other information.”</i>

Table 2. Prompt templates.

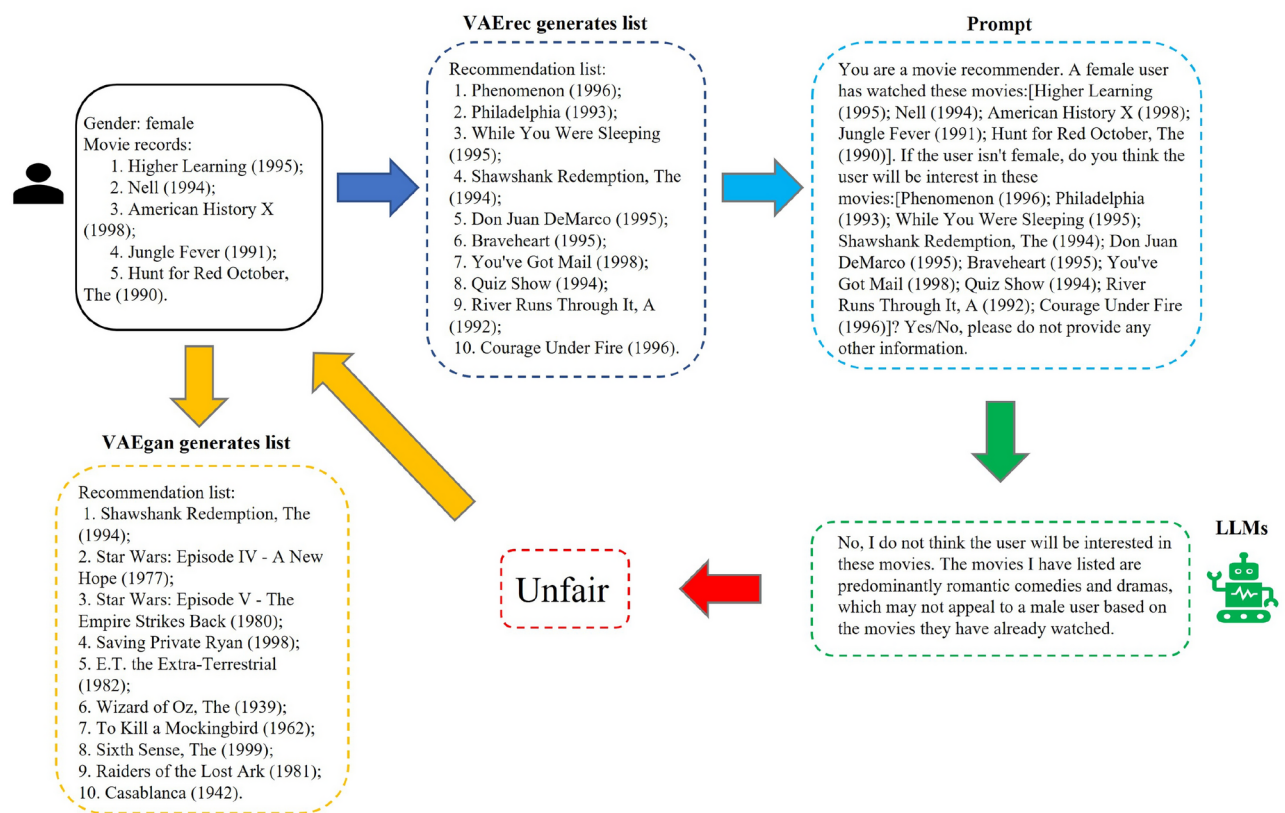


Fig. 1. Fair recommendations by incorporating LLMs.

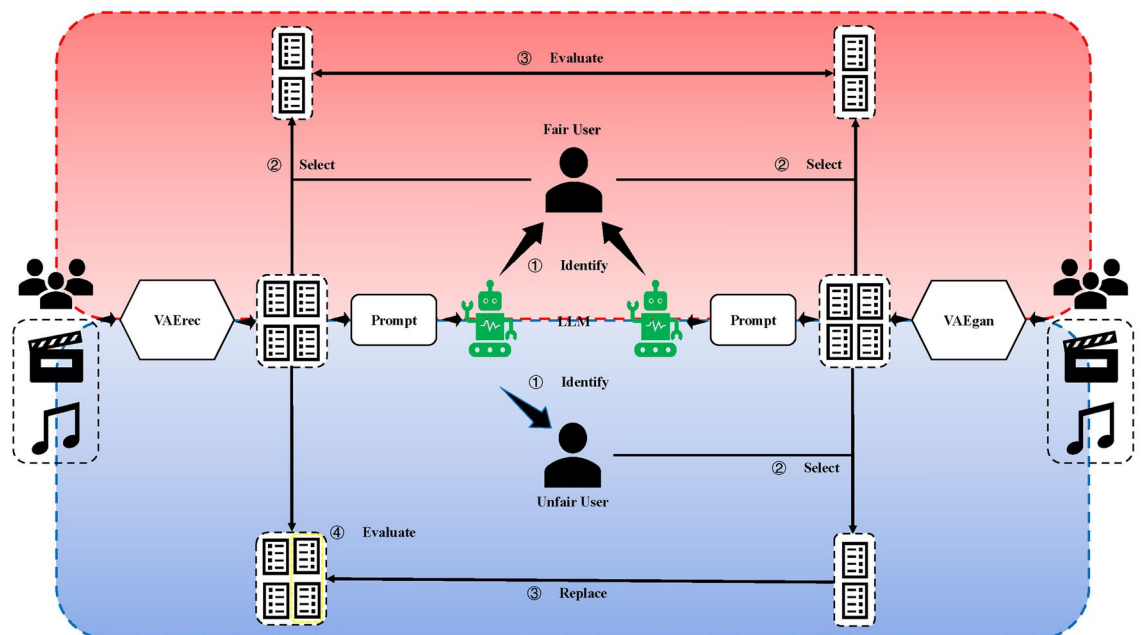


Fig. 2. Experimental plan.

Exploring the effect of LLMs' fairness awareness on fair recommendation

To investigate whether LLMs' fairness awareness can be leveraged to build fair recommendations, we design an exploratory method to examine the impact of LLMs' fairness awareness on recommendation fairness. This method utilizes LLMs' fairness recognition to distinguish fair users and then applies fairness intervention

to the recommendation results of users identified as unfair. Finally, it evaluates the recommendation results generated by combining LLMs' fairness recognition to determine the impact of LLMs' fairness awareness on recommendation fairness. As illustrated in the blue section of Fig. 2, we first use VAErec and VAEgan to generate recommendation lists before and after fairness intervention, respectively. Next, we construct fairness recognition prompts using the recommendation lists generated by VAErec and input them into the selected LLMs to produce recognition results. We then filter out users recognized as unfair, replace their recommendation lists with those generated by VAEgan, and reconstruct the recommendation results. Finally, we evaluate the reconstructed recommendation results in terms of recommendation utility and fairness to explore the impact of LLMs' fairness awareness on fair recommendations.

Experimental setting

In this study, we conduct recommendation experiments utilizing VAErec and VAEgan on the MovieLens and LastFM datasets, respectively.

VAErec serves as a baseline recommendation model that employs VAE to capture users' latent preferences. The key hyperparameters include the KL divergence weight ($\beta = 0.2$), training batch size (MovieLens: 805, LastFM: 500), testing batch size (MovieLens: 1610, LastFM: 1000), and the total number of training epochs (MovieLens: 200, LastFM: 125). To assess the fairness of the recommendations, the χ^2 -statistic metric is calculated by considering 600 items for MovieLens and 750 items for LastFM.

Building on the foundation of VAErec, VAEgan incorporates adversarial learning to enforce adversarial constraints on the latent space through the GAN-KL framework, with the objective of improving the fairness of recommendation outcomes. The key hyperparameters include the adversarial loss weights in the GAN-KL framework ($\alpha = 540$, $\gamma = 240$), the dimensionality of the latent space ($z_{\text{dim}} = 24$), the dimensionality of the adversarial latent variable ($b_{\text{dim}} = 3$), the number of adversarial training updates per minibatch ($n_{\text{adv_train}} = 32$), and an extended training duration (MovieLens: 500 epochs, LastFM: 350 epochs). Similar to VAErec, the fairness of recommendations is evaluated using the χ^2 -statistic metric, considering 600 items for MovieLens and 750 items for LastFM. Furthermore, to address the sparsity of the LastFM dataset, data loading is performed using the Compressed Sparse Row (CSR) format.

Evaluation method

We select three common evaluation metrics to evaluate the effectiveness and fairness of the recommender system³⁸: Normalized Discounted Cumulative Gain (NDCG@k), χ^2 -statistic and Kendall-Tau distance.

NDCG@k: NDCG is a widely used metric for evaluating the effectiveness of ranking models such as search engines and recommendation systems. It considers both the correctness of the ranking and the degree of relevance. A higher NDCG value indicates better performance of the ranking model. For all experiments, we set $k=10$.

χ^2 -statistic: The χ^2 -test is a statistical method for evaluating the fairness of a recommendation system. It uses an independence test to determine whether the recommendation results are related to the user's sensitive attributes. It also estimates the probability of drawing two independent samples from the same distribution. Since top-k recommendations are not independent, we focus solely on the statistic. We improve the reliability of statistical results by aggregating the top-100 recommendation lists for each sensitive attribute and setting the minimum expected value to 3. Detailed calculation steps and formulas are as follows:

1. **Constructing the contingency table**: For each sensitive attribute (e.g., gender, age), create a contingency table that records the number of recommendations for each item in different groups by counting each user's recommendation list within different groups to analyze recommendation fairness.
2. **Calculating expected frequencies**: For each cell in the contingency table, calculate the expected frequency. The expected frequency assumes that the recommendation distribution for both groups is the same. The calculation formula is:

$$E_{ij} = \frac{R_i \times C_j}{N}, \quad (1)$$

where E_{ij} is the expected frequency for the cell in the i -th row and j -th column, R_i is the total count of recommendations in the i -th row (for the i -th group), C_j is the total count of recommendations in the j -th column (for the j -th item) and N is the total number of recommendations.

3. **Calculating the χ^2 -Statistic**: The χ^2 -statistic is calculated using the following formula:

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad (2)$$

where O_{ij} is the observed frequency for the cell in the i -th row and j -th column and E_{ij} is the expected frequency for the cell in the i -th row and j -th column.

A smaller χ^2 -statistic indicates that the recommendation distributions for the two groups are more similar, meaning the recommendation system is fairer.

Kendall-Tau distance: Kendall–Tau distance measures the similarity between two recommendation lists and can be extended to handle cases where items in the lists do not completely overlap. This metric is also used for top-100 recommendation lists for sensitive groups. Recommendations for each sensitive group are aggregated based on individual user recommendations, and a ranking discount scheme is applied to give higher importance to higher-ranked items. Detailed calculation steps and formulas are as follows:

1. *Constructing the ordered list:* For each sensitive attribute (such as gender or age), create a contingency table that records the sum of the ranking order weights of each item in different groups, by counting the ranking order weights of each item in each user’s recommendation list within the different groups. Based on the sum of the ranking order weights of items in the contingency table, the ordered lists of recommended items for different groups are generated.
2. *Calculate concordant and discordant pairs:* For two ordered lists of each sensitive attribute, calculate the number of concordant pairs (pairs of elements that are in the same order in both lists) and discordant pairs (pairs of elements that are in different orders between the two lists).
3. *Calculate the Kendall–Tau distance:* The Kendall–Tau distance is calculated using the following formula:

$$K.T = \frac{CP - DP}{\frac{1}{2}n(n - 1)}. \tag{3}$$

Here, CP is the number of concordant pairs in both lists. DP is the number of discordant pairs in both lists. n is the number of elements in the list. The value of the Kendall–Tau distance ranges from -1 to 1 . A value of 1 indicates that the two lists are identical, a value of -1 indicates that the two lists are completely opposite, and a value of 0 indicates no correlation between the two lists. In fairness evaluation, the closer the Kendall–Tau distance is to 1 , the fairer the recommendation system.

Results and discussion

We conduct experiments using selected models to explore fair recommendations based on the fairness awareness of LLMs. Our findings address the following research questions:

- RQ1:** Can LLMs effectively identify fair recommendation results?
- RQ2:** What is the positive impact of LLMs’ fairness awareness on building fair recommender systems?

Fair identification analysis of LLMs (RQ1)

We use the fairness metrics introduced in “Evaluation method” section to evaluate the results generated according to the method described in “Methods” section. We summarize the experimental results of prompts with different lengths of history records in MovieLens and LastFM in Tables 3 and 4, respectively.

From these results, we conduct the following observational analysis:

- (1) For both movie and music recommendations, ChatGLM3 and Llama2 exhibit fairness identification based on sensitive attributes. In each dataset, for the recommendation results of fair users identified by different

N	LLM	User number	Model	$\chi^2@100\downarrow$		K.T@100 $\uparrow$	
				Gender	Age	Gender	Age
–			VAErec	3286.7968	2456.4946	0.5283	0.5622
–			VAEgan	848.6665	900.0537	0.8583	0.8414
5	Chatglm3	G (89.44%)	VAErec	533.2567	580.4614	0.7897	0.7838
		A (67.69%)	VAEgan	852.9829	701.6245	0.8585	0.8270
	Llama2	G(13.04%)	VAErec	719.3878	602.1030	0.4041	0.4139
		A (19.64%)	VAEgan	569.1107	471.6670	0.5734	0.6341
7	Chatglm3	G (94.88%)	VAErec	558.5686	736.7884	0.7893	0.7635
		A(71.62%)	VAEgan	896.4269	780.229	<i>0.8544</i>	<i>0.8125</i>
	Llama2	G(12.21%)	VAErec	553.1478	447.5628	0.3465	0.4931
		A (17.82%)	VAEgan	433.0471	600.7176	0.7045	0.4179

Table 3. The fairness assessment results of the fair recommendation results identified by various LLMs in the recommendation results respectively generated by VAErec and VAEgan in MovieLens. Column ‘N’ is the number of user history records in the prompt. Column ‘User number’ the percentage of users whose recommendations generated by VAErec and VAEgan are recognized as fair by LLMs. Bold data indicates optimal. Italics data indicates sub-optimal. “ $\uparrow$ ” indicates that a larger number represents greater fairness. “ $\downarrow$ ” indicates that a smaller number represents greater fairness. The same applies to other tables.

N	LLM	User number	Model	$\chi^2@100\downarrow$		K.T@100 $\uparrow$	
				Gender	Age	Gender	Age
–			VAErec	2636.1360	4544.0670	0.3768	0.0782
–			VAEgan	1670.9327	3007.7690	0.5055	0.2610
5	Chatglm3	G(79.80%)	VAErec	770.0154	768.5830	0.5026	0.4466
		A(38.62%)	VAEgan	1439.1733	1427.4130	0.4735	0.2627
	Llama2	G(0.49%)	VAErec	–	647.3035	– 0.7463	– 0.3065
		A(3.15%)	VAEgan	–	678.5224	– 0.8080	– 0.4086
10	Chatglm3	G(79.70%)	VAErec	753.3401	611.2097	0.5091	0.4855
		A(42.17%)	VAEgan	1292.3916	1974.0910	0.4651	0.1896
	Llama2	G(0.79%)	VAErec	–	565.7983	– 0.7960	– 0.6053
		A(9.95%)	VAEgan	–	759.0048	– 0.7694	– 0.3955

Table 4. The fairness assessment results of the fair recommendation results identified by various LLMs in the recommendation results respectively generated by VAErec and VAEgan in LastFM.

LLMs with different prompts, each fairness metric shows similar values before and after the fairness intervention and outperforms the overall fairness metric after the fairness intervention, except for a few cases with ordering differences. This suggests that the recommendation results of fair users identified by LLMs are not strongly correlated with sensitive attributes and have not been treated differently based on sensitive attributes.

- (2) In both datasets, regarding the number of users with different sensitive attributes identified as fair both before and after the fairness intervention, Llama2 identifies fewer users than ChatGLM3 and exhibits a different distribution. ChatGLM3 identifies more gender-fair users than age-fair users, while Llama2 identifies more age-fair users than gender-fair users. Additionally, Llama2 outperforms ChatGLM3 on the $\chi^2@100$ metric and is superior to VAEgan. This indicates that different LLMs vary in fairness identification capabilities, with larger models generally performing better.
- (3) For both movie and music recommendations, particularly music recommendations, the recommendation results for fair users identified by ChatGLM3 tend to deteriorate in fairness metric measures after fairness intervention. This suggests that fairness intervention strategies can lead to deteriorated fairness metric measures, partially validating the challenge mentioned in “Introduction” section.
- (4) In both datasets, for prompts with different lengths of history records, different LLMs do not show significant differences in fairness metrics for different sensitive attributes. This suggests that the length of user history records required to satisfy the fairness reasoning of LLMs does not need to be extensive.
- (5) In the movie dataset, the recommendation results for fair users identified by different prompts from different LLMs show a similar optimization trend in fairness metrics before and after fairness intervention. In contrast, the music dataset shows an opposite degradation trend. However, for both types of recommendations, all fairness metrics are better than the fairness metrics of VAEgan. This suggests that the variety of items of the same genre in the music dataset is richer than in the movie dataset, aligning with the actual phenomenon that the number of musicians with similar styles is much higher than the number of movies with similar subject matter.

Impact analysis of LLMs on fair recommendations (RQ2)

We utilize the metrics introduced in “Evaluation method” section to measure the reconstructed recommendation results generated by the research method described in “Methods” section. We summarize the experimental results in Table 5. Additionally, we analyze the data in Table 5 to compute the ratio of increased fairness (the percentage increase in $\chi^2@100$ or K.T@100 of the recommendation results) to decreased recommendation utility (the percentage decrease in NDCG of the recommendation results) as a result of different fairness strategy intervention outcomes, and these analysis results for MovieLens and LastFM are presented in Table 6.

From these results, we conduct the following observational analysis:

- (1) For both movie and music recommendations, regardless of the fairness intervention strategy, the accompanying increase in fairness generally leads to a decrease in recommendation utility. However, incorporating fairness recognition of LLMs can mitigate the rate of utility reduction when improving the fairness of recommendation results. For MovieLens, ChatGLM3 is able to achieve a fairness enhancement far beyond VAEgan with the same utility loss. For LastFM, Llama2 can achieve somewhat similar results for fairness intervention in gender attributes. This suggests that LLMs’ fairness identification helps achieve a better trade-off between recommendation utility and fairness.
- (2) Llama2 outperforms ChatGLM3 in terms of utility and fairness in both datasets. However, there are significant differences in the trade-off between utility and fairness across different datasets. For MovieLens, ChatGLM3 outperforms Llama2, though both achieve better results than VAEgan (ChatGLM3 $n = 5$ and Llama2 $n = 7$). However, for LastFM, all models except Llama2 perform worse than VAEgan. Llama2 shows some superiority over VAEgan in terms of gender attributes, suggesting that fairness recognition varies across different datasets and LLMs.

Dataset	Model	N	NDCG@10		$\chi^2@100\downarrow$		K.T@100↑	
			Gender	Age	Gender	Age	Gender	Age
MovieLens	VAErec	–	0.3098	0.3098	3286.7968	2456.4946	0.5283	0.5622
	VAEgan	–	0.2716	0.2716	848.6665	900.0537	0.8583	0.8414
	Chatglm3	5	0.3097	0.3048	3253.9830	2009.3719	0.5369	0.6028
		7	0.3051	0.3043	3278.0419	2282.9265	0.5396	0.5937
	Llama2	5	0.2744	0.2723	1367.4062	1041.0794	0.8011	0.7700
		7	0.2842	0.2750	1415.3913	1006.1315	0.8087	0.7719
LastFM	VAErec	–	0.1868	0.1868	2636.1360	4544.0674	0.3768	0.0782
	VAEgan	–	0.1799	0.1799	1670.9327	3007.7689	0.5055	0.2610
	Chatglm3	5	0.1800	0.1799	2341.7013	3804.5084	0.3829	0.1562
		10	0.1800	0.1800	2435.1466	3948.7413	0.3857	0.1237
	Llama2	5	0.1816	0.1803	1744.2568	3316.0478	0.5027	0.2231
		10	0.1800	0.1800	1718.8502	3139.6539	0.4802	0.2162

Table 5. The utility and fairness results of recommendations of different methods in different datasets.

Dataset	Model	N	$\chi^2@100$		K.T@100	
			Gender	Age	Gender	Age
MovieLens	VAEgan	–	6.0159	5.1385	5.0658	4.0276
	Chatglm3	5	30.9289	11.2777	50.4312	4.4745
		7	0.1756	3.9799	1.4099	3.1560
	Llama2	5	5.1106	4.7601	4.5190	3.0535
		7	6.8903	5.2561	6.4230	3.3205
	VAErec	–	9.9124	9.1529	9.2469	63.2845
LastFM	Chatglm3	5	3.0682	4.4061	0.4447	27.0032
		10	2.0945	3.5990	0.6489	15.9835
	Llama2	5	12.1538	7.7665	12.0030	53.2507
		10	9.5588	8.4902	7.5384	48.4775
	VAEgan	–	9.9124	9.1529	9.2469	63.2845
	VAErec	–	9.9124	9.1529	9.2469	63.2845

Table 6. The fairness-to-utility ratio of different fair intervention methods in different datasets.

- (3) In both datasets, the fair recognition of historical records of different lengths impacts recommendation effectiveness. For ChatGLM3, using fewer history records achieves better results, while for Llama2, using more history records leads to better outcomes. This suggests that different LLMs require different lengths of history records to perform accurate fair recognition.

Conclusion and future work

Regarding the first research question, the results suggest that existing fairness strategies that intervene uniformly without user-specific identification can actually lead to decreased fairness in recommendation results. LLMs exhibit a certain sense of fairness and can identify fair recommendation results based on their reasoning capabilities, though the identification capabilities vary with the size of the LLMs. While current LLMs, even those that are large-scale or effectively fine-tuned, may not accurately identify all fair recommendations, this insight could inspire further research into training or fine-tuning strategies for LLMs to enhance their fairness identification accuracy. In answering the second research question, we demonstrated that leveraging LLMs for fairness identification helps achieve a better trade-off between recommendation utility and fairness. Although current effectiveness is not ideal, the human-like cognition of LLMs offers new opportunities for fairness research in recommendation systems and can promote the development of fairness-conscious recommendation models. This study contributes to the understanding of fairness in recommender systems by exploring the fairness awareness of LLMs. Our findings highlight the potential of LLMs in identifying and addressing unfairness in recommendation results. However, further research is needed to improve the accuracy of LLMs in recognizing fair recommendations and to develop effective fairness intervention strategies.

As future work, we plan to further refine the exploration of LLMs’ fairness awareness by investigating their fairness recognition capabilities on other datasets, using different types or sizes of LLMs, addressing other fairness problems, and employing alternative prompt design schemes. Additionally, finetuning can better harness the capabilities of LLMs. We will explore fine-tuning schemes to induce stronger fairness awareness in LLMs when training LLMs from scratch is not feasible.

Data availability

We conduct our experiments on two publicly available datasets. MovieLens 1M is available at: <https://grouplens.org/datasets/movielens/1m/>. LastFM 360K is available at: <https://www.upf.edu/web/mtg/lastfm360k>.

Code availability

Our code is available at: <https://github.com/sy-cg/LLM-fair-aware>.

Received: 9 October 2024; Accepted: 10 February 2025

Published online: 14 February 2025

References

- Mandal, S. & Maiti, A. Deep collaborative filtering with social promoter score-based user-item interaction: A new perspective in recommendation. *Appl. Intell.* **51**, 7855–7880 (2021).
- Mandal, S. & Maiti, A. Graph neural networks for heterogeneous trust based social recommendation. In *2021 International Joint Conference on Neural Networks (IJCNN)* 1–8 (IEEE, 2021).
- Mandal, S. & Maiti, A. Explicit feedback meet with implicit feedback in gpmf: A generalized probabilistic matrix factorization model for recommendation. *Appl. Intell.* **50**, 1955–1978 (2020).
- Mandal, S. & Maiti, A. Explicit feedbacks meet with implicit feedbacks: A combined approach for recommendation system. In *International Conference on Complex Networks and Their Applications* 169–181 (Springer, 2018).
- Liu, H. et al. Rating distribution calibration for selection bias mitigation in recommendations. *Proc. ACM Web Conf.* **2022**, 2048–2057 (2022).
- Li, Y., Liu, K., Satapathy, R., Wang, S. & Cambria, E. Recent developments in recommender systems: A survey. *IEEE Comput. Intell. Mag.* **19**, 78–95 (2024).
- Li, Z., Chen, Y., Zhang, X. & Liang, X. Bookgpt: A general framework for book recommendation empowered by large language model. *Electronics* **12**, 4654 (2023).
- Zhang, J. et al. Is chatgpt fair for recommendation? Evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems* 993–999 (2023).
- Deldjoo, Y. & Di Noia, T. Cfairllm: Consumer fairness evaluation in large-language model recommender system. Preprint at <http://arxiv.org/abs/2403.05668> (2024).
- Deldjoo, Y. Fairevallm: A comprehensive framework for benchmarking fairness in large language model recommender systems. Preprint at <http://arxiv.org/abs/2405.02219> (2024).
- Jiang, M. et al. Item-side fairness of large language model-based recommendation system. *Proc. ACM Web Conf.* **2024**, 4717–4726 (2024).
- Xu, C. et al. Do llms implicitly exhibit user discrimination in recommendation? An empirical study. Preprint at <http://arxiv.org/abs/2311.07054> (2023).
- Li, T. et al. Your large language model is secretly a fairness proponent and you should prompt it like one. Preprint at <http://arxiv.org/abs/2402.12150> (2024).
- Shiffrin, R. & Mitchell, M. Probing the psychology of ai models. *Proc. Natl. Acad. Sci.* **120**, e2300963120 (2023).
- Binz, M. & Schulz, E. Using cognitive psychology to understand gpt-3. *Proc. Natl. Acad. Sci.* **120**, e2218523120 (2023).
- Lampinen, A. K. et al. Language models show human-like content effects on reasoning tasks. Preprint at <http://arxiv.org/abs/2207.07051> (2022).
- Jiang, G. et al. Mewl: Few-shot multimodal word learning with referential uncertainty. In *International Conference on Machine Learning* 15144–15169 (PMLR, 2023).
- Aher, G. V., Arriaga, R. I. & Kalai, A. T. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning* 337–371 (PMLR, 2023).
- Rosenfeld, R. Two decades of statistical language modeling: Where do we go from here? *Proc. IEEE* **88**, 1270–1278 (2000).
- Mikolov, T., Karafiát, M., Burget, L., Cernocký, J. & Khudanpur, S. Recurrent neural network based language model. In *Interspeech* 1045–1048 (Makuhari, 2010).
- Sarzynska-Wawer, J. et al. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Res.* **304**, 114135 (2021).
- Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
- Chowdhery, A. et al. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.* **24**, 1–113 (2023).
- Touvron, H. et al. Llama: Open and efficient foundation language models. Preprint at <http://arxiv.org/abs/2302.13971> (2023).
- Zeng, A. et al. Glm-130b: An open bilingual pre-trained model. Preprint at <http://arxiv.org/abs/2210.02414> (2022).
- Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 1 (2017).
- Wei, J. et al. Emergent abilities of large language models. Preprint at <http://arxiv.org/abs/2206.07682> (2022).
- Jiang, H. A latent space theory for emergent abilities in large language models. Preprint at <http://arxiv.org/abs/2304.09960> (2023).
- Webb, T., Holyoak, K. J. & Lu, H. Emergent analogical reasoning in large language models. *Nat. Hum. Behav.* **7**, 1526–1541 (2023).
- Wong, L. et al. From word models to world models: Translating from natural language to the probabilistic language of thought. Preprint at <http://arxiv.org/abs/2306.12672> (2023).
- Jones, E. & Steinhardt, J. Capturing failures of large language models via human cognitive biases. *Adv. Neural Inf. Process. Syst.* **35**, 11785–11799 (2022).
- Ziems, C. et al. Can large language models transform computational social science? *Comput. Linguist.* **50**, 237–291 (2024).
- Park, J. S. et al. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* 1–22 (2023).
- Jiang, G. et al. Evaluating and inducing personality in pre-trained language models. *Adv. Neural Inf. Process. Syst.* **36**, 1 (2024).
- Kingma, D. P. & Welling, M. Auto-encoding variational bayes. Preprint at <http://arxiv.org/abs/1312.6114> (2013).
- Liang, D., Krishnan, R. G., Hoffman, M. D. & Jebara, T. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 World Wide Web Conference* 689–698 (2018).
- Borges, R. & Stefanidis, K. Feature-blind fairness in collaborative filtering recommender systems. *Knowl. Inf. Syst.* **64**, 943–962 (2022).
- Vassøy, B., Langseth, H. & Kille, B. Providing previously unseen users fair recommendations using variational autoencoders. In *Proceedings of the 17th ACM Conference on Recommender Systems* 871–876 (2023).
- Park, C., Kim, D., Oh, J. & Yu, H. Do “also-viewed” products help user rating prediction? In *Proceedings of the 26th International Conference on World Wide Web* 1113–1122 (2017).
- Yin, H., Wang, W., Wang, H., Chen, L. & Zhou, X. Spatial-aware hierarchical collaborative deep learning for poi recommendation. *IEEE Trans. Knowl. Data Eng.* **29**, 2537–2551 (2017).
- Harper, F. M. & Konstan, J. A. The movielens datasets: History and context. *Acm Trans. Interact. Intell. Syst.* **5**, 1–19 (2015).
- Chen, L. et al. Improving recommendation fairness via data augmentation. *Proc. ACM Web Conf.* **2023**, 1012–1020 (2023).

43. Du, Z. et al. Gln: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 320–335 (2022).
44. Touvron, H. et al. Llama 2: Open foundation and fine-tuned chat models. Preprint at <http://arxiv.org/abs/2307.09288> (2023).

Author contributions

Wei Liu carried out the experiment and result analysis, wrote the main manuscript text. All authors reviewed the manuscript.

Funding

This work was supported in part by the Natural Science Foundation of Zhejiang Province (No. LZ20F020001), Science and Technology Innovation 2025 Major Project of Ningbo (No. 20211ZDYF020036), Natural Science Foundation of Ningbo (No. 2021J091), and the Research and Development of a Digital Infrastructure Cloud Operation and Maintenance Platform Based on 5G and AI (No. HK2022000189).

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval and consent to participate

This work did not involve any human participants, their data, or biological materials, and therefore did not require ethical approval.

Additional information

Correspondence and requests for materials should be addressed to B.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025