Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era

Sunhao Dai
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
sunhaodai@ruc.edu.cn

Chen Xu
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
xc_chen@ruc.edu.cn

Shicheng Xu
CAS Key Laboratory of AI Safety
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China
xushicheng21s@ict.ac.cn

Liang Pang[*]
CAS Key Laboratory of AI Safety
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China
pangliang@ict.ac.cn

Zhenhua Dong
Huawei Noah's Ark Lab
Shenzhen, China
dongzhenhua@huawei.com

Jun Xu
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
junxu@ruc.edu.cn

ABSTRACT

With the rapid advancements of large language models (LLMs), information retrieval (IR) systems, such as search engines and recommender systems, have undergone a significant paradigm shift. This evolution, while heralding new opportunities, introduces emerging challenges, particularly in terms of biases and unfairness, which may threaten the information ecosystem. In this paper, we present a comprehensive survey of existing works on emerging and pressing bias and unfairness issues in IR systems when the integration of LLMs. We first unify bias and unfairness issues as distribution mismatch problems, providing a groundwork for categorizing various mitigation strategies through distribution alignment. Subsequently, we systematically delve into the specific bias and unfairness issues arising from three critical stages of LLMs integration into IR systems: data collection, model development, and result evaluation. In doing so, we meticulously review and analyze recent literature, focusing on the definitions, characteristics, and corresponding mitigation strategies associated with these issues. Finally, we identify and highlight some open problems and challenges for future work, aiming to inspire researchers and stakeholders in the IR field and beyond to better understand and mitigate bias and unfairness issues of IR in this LLM era. We also consistently maintain a GitHub repository for the relevant papers and resources in this rising direction at https://github.com/KID-22/LLM-IR-Bias-Fairness-Survey.

CCS CONCEPTS

• **Information systems → Information retrieval**.

KEYWORDS

Information Retrieval, Large Language Model, Bias, Fairness

ACM Reference Format:
Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong and Jun Xu. 2024. Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3637528.3671458

1 INTRODUCTION

Information Retrieval (IR) systems strive to navigate the era of information overload, facilitating users in acquiring information more efficiently and effectively [99, 138, 171]. The integration of Large Language Models (LLMs) has fundamentally redefined IR systems, including the introduction of LLM-generated data as new IR data sources, the shift from passive retrieval to proactive generation as core paradigms, and adopting LLMs as results evaluators for IR systems [4, 201]. These advancements, however, bring forth new challenges in bias and unfairness, affecting the reliability of IR systems and potentially contributing to societal issues like echo chambers [30, 134] and cognitive interference [104, 132]. For instance, researchers found that LLMs often retrieve information that deviates from facts and is biased towards LLM-generated content [20, 28, 76, 174]. Moreover, LLMs frequently manifest stereotypes and discriminatory content to users and amplify disparities between items of different socio-economic statuses [64, 170, 190].

Recently, much effort has been made around bias and unfairness in the context of LLMs and IR systems. However, the literature is currently fragmented and often lacks a unified definition of these concepts. This ambiguity hampers the development of systematic strategies to address these issues effectively. To this end, our survey aims to provide a comprehensive and unified perspective that effectively summarizes the emerging challenges and opportunities related to bias and unfairness in the intersection between LLMs and IR systems. Generally, both bias and unfairness issues can be

[*] Corresponding author.

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671458

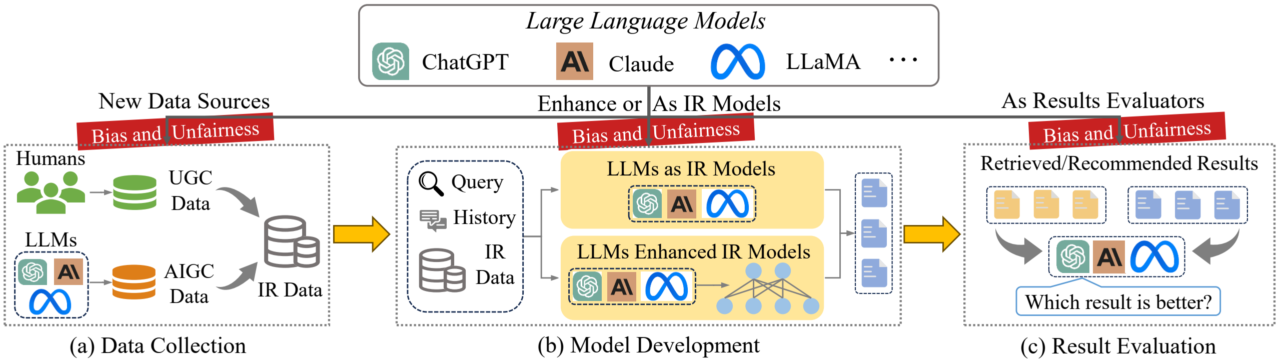

Figure 1: Overview of three stages of the intersection between LLMs and IR systems. (a) LLMs-generated content as new data sources for IR. (b) Incorporating LLMs to enhance or as IR models. (c) Adopting LLMs as results evaluators in IR systems.

regarded as a *distribution mismatch problem*. Specifically, bias underscores the fact that the predicted information lacks objectivity and truthfulness, highlighting the mismatch with the objective target distribution. Unfairness reveals that the predicted information fails to align with the social values between humans and machines, leading to a mismatch with the subjective target distribution of human values. This perspective not only unifies the nature of these issues but also streamlines the exploration of mitigation strategies.

Our survey begins with a brief overview of how LLMs are integrated into IR systems, setting the stage for understanding the emergence of new bias and unfairness challenges across three pivotal stages of the IR lifecycle: data collection, model development, and result evaluation. Then we propose a unified perspective on bias and unfairness, categorizing them as distribution mismatch problems. Based on this unified view, we categorize mitigation strategies into two principal groups: data sampling, including data augmentation and data filtering, and distribution reconstruction, encompassing rebalancing, regularization, and prompting. Following this taxonomy, we delve into a detailed analysis of several types of bias and unfairness phenomena that arise with the integration of LLMs into IR systems, spanning the aforementioned stages. Our systematic review encompasses a comprehensive examination of these issues and their respective mitigation strategies in recent studies, providing a holistic view of the current landscape and guiding future efforts in eliminating bias and ensuring fairness for more trustworthy IR systems.

Difference with Existing Surveys. Several recent surveys have reviewed and discussed the issues of bias and fairness within IR [16, 83, 118, 161, 188], primarily focusing on works published before the advent of LLMs. With LLMs becoming increasingly prevalent, a new subset of surveys [43, 84, 93, 179] has paid attention to the bias and fairness challenges presented by LLMs themselves. Additionally, some other recent surveys [4, 87, 167, 201] have examined how integrating LLMs can enhance and transform traditional IR systems, highlighting some opportunities arising from this integration. Compared to these surveys, our work stands apart by offering a comprehensive survey of the emerging and pressing issues of bias and fairness at the intersection of IR and LLMs, employing a novel unified perspective to review the cause and mitigation strategies.

Summary of Contributions. (1) We provide a novel unified perspective for understanding bias and unfairness as distribution mismatch problems, alongside a detailed review of several types of bias and unfairness arising from integrating LLMs into IR systems. (2) We systematically organize mitigation strategies into two key categories: data sampling and distribution reconstruction, offering a comprehensive roadmap for effectively combating bias and unfairness with state-of-the-art approaches. (3) We identify the current challenges and future directions, providing insights to facilitate the development of this potential and demanding research area.

2 A UNIFIED VIEW OF BIAS AND UNFAIRNESS

In this section, we provide a unified view of bias and unfairness from the perspective of distribution alignment and outline the mitigation strategies based on this view.

2.1 Background

As shown in Figure 1, the advent of LLMs has reshaped the whole pipeline of IR systems, typically in the following three stages: data collection, model development, and result evaluation.

LLMs-generated content as new data sources for IR. The emergence of LLMs has significantly accelerated the growth of Artificial Intelligence Generated Content (AIGC), marking a new era in content creation. Unlike traditional Professional and User Generated Content (PGC and UGC) sources, AIGC can be produced automatically at scale, potentially dominating the content landscape. [13, 55, 165]. However, AIGC also reshapes the distribution of the IR data, resulting in new concerns about bias and fairness.

Incorporating LLMs to enhance or as IR models. The impressive emergent capabilities of LLMs in understanding, reasoning, and generalization have motivated significant efforts to integrate them into the development of next-generation IR systems [201]. On one hand, LLMs have been deployed to refine key components of traditional IR systems [136, 142, 153], enhancing their effectiveness and efficiency. On the one hand, beyond enhancing existing frameworks, LLMs also introduce a novel paradigm by acting as generative search and recommendation agents [27, 52, 107, 189], directly generating responses to fulfill user queries.

Adopting LLMs as results evaluators in IR systems. Human evaluation plays a pivotal role in IR systems, particularly in conversational search and recommendation, such as directly assessing the quality of responses from generative LLM-based IR models [27, 78, 198]. However, human evaluation comes with significant challenges, including high costs and a lack of reproducibility. LLMs, with their advanced language modeling and understanding capabilities, offer new possibilities for conducting evaluations of these complex tasks without human evaluation [22, 45, 86, 144]. This shift not only streamlines the evaluation process but also mitigates the substantial costs associated with human evaluations, further facilitating the development of IR systems.

2.2 Distribution Alignment Perspective

While the reshaping of the above IR stages by LLMs has introduced numerous new opportunities, it has also given rise to many new and pressing issues related to bias and unfairness. In this section, we formulate the problems of bias and fairness from a distribution alignment perspective, offering a unified framework for understanding these challenges and inspiring mitigation strategies.

Formally, when a user interacts with a typical IR system, he/she may optionally provide his/her personalized information requirement T along with their personalized attributes (typically indicated as user profile U) to the system. Subsequently, the goal of the IR system is to retrieve the target information R (*e.g.,* documents, items, or advertisements, *et al.*) for this user, with the user's information requirements and optional interaction history H as the input $Q = \{T, U, H\}$. Let $\widehat{R} = f(Q)$ be the predicted result either from IR models or directly generated by LLMs, where $f(\cdot)$ is the model.

Let $P(\widehat{R})$ be the distribution of predicted results for all users, we can unify the bias and unfairness problem as a *distribution mismatch problem* with the ground truth distribution $P(R)$, where R is the target result, which can be defined as follows:

$$P(\widehat{R}) \neq P(R). \tag{1}$$

Based on the Equation 1, the bias and unfairness problem can be explained as follows:

• **Bias** stems from systemic deviations occurring at various stages of the IR process, from data collection to model design and evaluation. These systemic issues results in the predicted distribution, $P(\widehat{R})$, diverging from the target distribution, $P(R)$, which ideally represents objective and factual realities.

• **Unfairness** deeply rooted in cultural and societal notions of fairness, aims to align the predicted distribution $P(\widehat{R})$ with a subjective target distribution $P(R)$. It reflects human values and social contracts and evolves with the progress of time.

2.3 Taxonomies of Mitigation Strategies

Based on our unified view, we can further systematically categorize mitigation strategies from the view of distribution alignment. Specifically, the goal of mitigation strategies is to align the distribution of retrieved information with a target distribution defined by either objective criteria or subjective social values. As shown in Figure 2, we outline two primary categories and their sub-strategies:

(1) **Data Sampling** focuses on directly modifying the data:

• **Data Augmentation** serves as distribution completion, enriching the dataset with additional, often synthetic, data to approximate the target distribution more closely. Techniques such as counterfactual imputation and the incorporation of external knowledge are employed to fill in the gaps in the existing dataset, thus aiming to recover the real distribution more accurately.

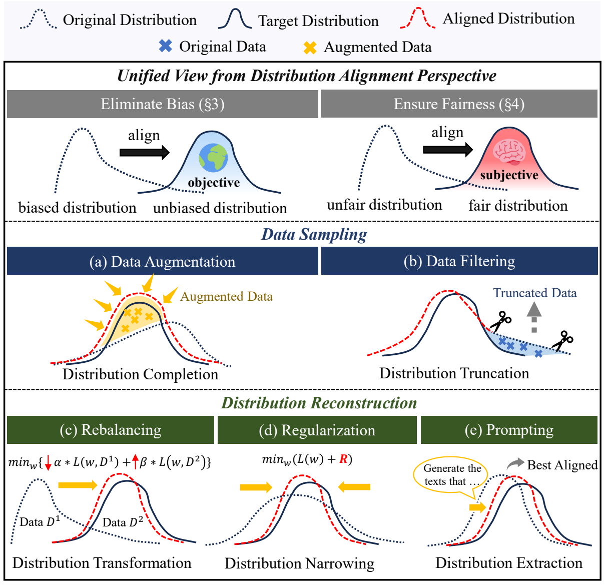

Figure 2: Illustration of different types of mitigation strategies from a unified view of distribution alignment.

• **Data Filtering** acts as distribution truncation, selecting data subsets that align with the target distribution, ensuring the model's output is representative of desired target outcomes. Techniques like re-ranking and constrained beam search fall under this category, serving as post-processing methods to ensure the retained distribution segment matches the target distribution.

(2) **Distribution Reconstruction** aims at adjusting the predicted distribution:

• **Rebalancing** transforms the predicted distribution through techniques like reweighting or resampling, to reflect the target distribution more accurately. Common strategies include adjusting the loss weights for various groups to achieve equilibrium, thereby realigning the predicted distribution with the target distribution.

• **Regularization** narrows the predicted distribution by introducing constraints that encourage the model to learn the target distribution more faithfully. It encompasses both implicit approaches, such as adversarial learning, and explicit ones, like regularization techniques, to modify the distribution directly.

• **Prompting** extracts the best aligned distribution by directly employing specific prompts. This approach guides LLMs to generate outputs more likely from the target distribution, facilitating an alignment with the desired target distribution.

Through the lens of distribution alignment, these strategies offer a structured and coherent approach to eliminate bias and ensure fairness of IR systems in IR systems. In the following sections, we will conduct a detailed review of various emerging issues related to bias and unfairness, their mitigation solutions, and their integration within the distribution alignment framework.

3 CAUSE AND MITIGATION OF BIAS

As summarized in Table 1, we present an in-depth review of different types of bias at the different stages of the intersection between LLMs and IR systems and discuss the corresponding mitigation strategies.

3.1 Bias in Data Collection

In this subsection, we categorize the bias caused by data collection into two groups: source bias and factuality bias.

3.1.1 Source Bias. Source bias emerges when the incorporation of LLM-generated content into the corpus of the IR systems:

- **Definition.** Information retrieval models tend to rank content generated by LLMs higher than content authored by humans.

Specifically, as LLMs fuel the rapid expansion of LLM-generated content, the corpus for IR systems now increasingly encompasses a mix of human-written and LLM-generated texts. Recent studies [26, 28, 174] highlight that modern retrieval models, especially those leveraging neural matching techniques, tend to favor LLM-generated content over human-authored content with similar semantics. This preference stems from the unique representations embedded in LLM-generated content, which neural retrieval models can capture and thus assign a higher estimated relevancy score [28, 174]. More severely, this source bias is further amplified when LLM-generated content is included in model training, presenting a challenge to the information ecosystem [174]. Zhou et al. [200] further explores this escalation of source bias in user, data, and recommender systems feedback loop. Tan et al. [145] further that this bias will extend from retrievers to the readers and Chen et al. [18] uncover this bias in the RAG systems.

To counteract source bias, recent studies have introduced debiased constraints into the training objectives of IR retrieval models [28, 174, 200]. This strategy aims to correct the skewed relevancy predictions favoring LLM-generated content, ensuring fair treatment between human-written and LLM-generated content. By adopting a distribution alignment perspective, such mitigation efforts strive to recalibrate the IR models' relevancy distribution towards an ideal state, where documents are judged equally based on their semantic content rather than their source.

3.1.2 Factuality Bias. As AIGC increasingly becomes a part of the data sources for IR systems, it inevitably introduces a significant amount of non-factual or "hallucinated" content. This introduction alters the distribution of IR system data, thereby leading to biases in the retrieval process.

- **Definition.** LLMs may produce content that does not align with recognized factual information of the real world.

Many studies have shown that LLMs are at risk of generating factual errors. For instance, TruthfulQA [88] has highlighted that language models generate many false answers that mimic popular misconceptions in question-answering tasks and have the potential to deceive humans. Besides, merely scaling up models is not promising for improving truthfulness, which means that LLMs still face challenges in generating factually correct content. Lee et al. [75] show LLMs are susceptible to generating text with nonfactual in an open-ended generation because of the "uniform randomness" at every sampling step. FActScore [105] finds that LLMs lag significantly behind humans in ensuring the factual consistency of long-form text generation. Other studies [101] reveal that LLMs also exhibit factual hallucination in natural language inference tasks. In addition to the above studies, many large-scale benchmarks [17, 20, 76] also indicate that LLMs exhibit factuality bias in multi-task and multi-domain scenarios.

Previous studies find that the flawed data source and inferior data utilization are two important causes of factuality bias [61]. Specifically, some low-quality, factual errors, and long-distance repetition in the training texts harm the factual correctness of the text generated by LLMs [10, 74, 88]. The coverage of knowledge by training data also limits the correctness of LLMs in generating knowledge in some rare or specialized fields [68, 110, 139]. In addition to the training data, LLMs usually resort to shortcuts to generate the texts depending on position close and co-occurred words rather than understand the knowledge itself [65, 66, 80] and always fail to recall the knowledge that has been memorized [98, 197].

To mitigate factuality bias, in the training of LLMs, some methods focus on providing high-quality and factually correct training data for LLMs [51, 147]. In the inference, previous studies can be divided into two categories. One is with the help of an external factual knowledge base such as retrieval-augmented generation [44, 119, 126, 175–177, 184]. The other is to improve the ability of LLMs themselves such as Self-Consistency [159] and Dola [24].

3.2 Bias in Model Development

Incorporating LLMs into IR models introduces inherent randomness in the generation results, potentially leading to inconsistent outcomes. In this subsection, we categorize the bias in model development into four groups: position bias, popularity bias, input-hallucination bias, and context-hallucination bias.

3.2.1 Position Bias. Position bias emerges notably in scenarios where LLMs are utilized directly as retrieval or recommender systems [58, 146], characterized by the preference of documents or items based on their input positions:

- **Definition.** LLM-based IR models tend to give preference to documents or items from specific input positions.

Recent works have highlighted that the order of candidate documents or items can significantly impact the performance of LLM-based IR models, while conventional IR models are often not affected by the changing of input orders [58, 146, 172]. For instance, LLM-based models often show a preference for content positioned at the beginning or end of a list, neglecting the contributions of items in the middle. This "lost in the middle" suggests that these LLM-based models may not fully utilize the context provided by items that don't occupy prominent positions in the input sequence [90].

There are a number of works on mitigating position bias, and we can categorize them into three lines based on our distribution alignment framework. (1) Prompting: This approach involves carefully designed prompts to encourage the models to disregard the input's order [27, 58]. Nonetheless, due to LLMs' prompt sensitivity, this requires precise and task-specific prompt engineering across various tasks and domains. (2) Data Augmentation: This approach has been explored in numerous studies, which is a form of data augmentation that involves random shuffling of candidates followed by aggregation to determine the final ranking. [58, 96, 123, 146, 166, 191]. For instance, Tang et al. [146] introduced a permutation self-consistency

Table 1: The taxonomy of different types of bias in the intersection between LLMs and IR systems.

Sourced Stage	Type	Mitigation Strategies				
		Data Sampling		Distribution Reconstruction		
		Data Augmentation	Data Filtering	Rebalancing	Regularization	Prompting
Data Collection	Source Bias		[18]		[28, 174, 200]	
	Factuality Bias	[51, 119, 126, 175–177, 184]	[51, 147, 182]			[119, 143, 159, 176]
Model Development	Position Bias	[58, 96, 123, 146, 166, 191]		[97, 166]		[58]
	Popularity Bias	[158, 191]				[31, 58, 140]
	Instruction-Hallucination Bias	[106, 131, 160]			[39]	[117, 183]
	Context-Hallucination Bias	[7, 42]				
Result Evaluation	Selection Bias	[21, 23, 79, 85, 116, 155, 196, 198]		[94, 155, 195]		[70, 116, 155, 196]
	Style Bias					[168, 196]
	Egocentric Bias	[79]		[91]		[56, 91]

method, offering theoretical guarantees under certain conditions, enabling models to produce and aggregate potential outcomes from various candidate permutations, enhancing result stability. (3) Rebalancing: This method counteracts position bias by adjusting the prior distribution sensitive to positions. It recalibrates the model's output, addressing the inherent bias towards item positions [97, 166].

While these strategies offer pathways to counteract position bias, they present challenges, notably the increased computational demand associated with processing multiple permutations [58, 123, 146]. Future work should aim to balance effectiveness with efficiency to develop more stable and unbiased LLM-based IR systems.

3.2.2 *Popularity Bias.*

Popularity bias has been a widely studies issue in traditional IR models, with extensive research highlighting its effects on the so-called "Matthew effect" issue [2, 192]. This bias is characterized by the long-tail phenomenon, where a minority of popular items garners a majority of user interactions. Consequently, models trained on such skewed data tend to favor these popular items, often over-representing them in results and further exacerbating their dominance [16, 202]. However, the advent of LLM-based IR models introduces new dimensions to the challenge of popularity bias, which can be defined as follows:

• **Definition.** LLM-based IR models tend to prioritize candidate documents or items with high popularity levels.

Unlike conventional models, LLM-based IR models do not merely reflect the popularity distributions of the target finetuning dataset. They are also inclined to retrieve or recommend items popular in the pre-training corpora of the LLMs [9, 57]. For instance, the vast training data of LLMs, encompassing a wide array of content, means that certain documents or items may be over-represented, influencing the model to prefer these familiar documents or items in retrieval and recommendation tasks. As a result, LLM-based IR models may not only exhibit the existing popularity bias found within the target finetuning dataset but also introduce a new bias based on the content's prevalence in their pre-training data. This extension of popularity bias in LLM-based IR models presents a more complex problem.

To combat popularity bias in LLM-based IR models, existing methods explore two main solutions: (1) Data Augmentation: Wang et al. [158] proposes two data augmentation strategies to diversify the dataset by incorporating more underrepresented content, aiming to balance the final recommendation results. (2) Prompting: Alternative methods involve crafting specific instructions to directly intervene in the LLM's output, such as encouraging an equitable mix of popular and long-tailed items in results. [31, 58, 140] However, addressing this expanded notion of popularity bias in LLM-based IR systems requires new strategies in future work that account for both the inherent biases of the training data and the additional biases introduced by the LLMs' pre-training corpora.

3.2.3 *Instruction-Hallucination Bias.*

Instruction-hallucination bias emerges when LLMs are used as retrievers, rerankers, or recommenders but do not fully follow the user's instructions:

• **Definition.** Content generated by LLM-based IR models may deviate from the instructions provided by users.

Recent studies reveal that LLMs often struggle to adhere fully to users' instructions across various natural language processing tasks, such as dialogue generation [35], question answering [34] and summarization [100, 120]. These instructions comprise the users' intent or input task (e.g., reranking a document list) and the specific content or object (e.g., the document list) targeted by the task. Deviations from the input task suggest that LLMs may misinterpret the tasks users intend to execute. For instance, when deployed as recommenders, LLMs might not grasp users' requests for items with particular characteristics, leading to recommendations that do not match the request [37]. Similarly, contradictions with the input content reveal that in tasks like reranking, LLMs may produce results that are inconsistent with, or even absent from, the given instructions, showcasing a gap in understanding and fulfilling the specified requirements [201].

The key to mitigating the instruction-hallucination bias is to enhance the instruction following the ability of large language models. For example, some works propose high-quality instruction fine-tuning datasets such as Natural Instructions [106], Public Pool of Prompts [131], and Self-Instruct [160], etc. Besides, other studies try to further align content generated by LLMs with human preferences via reinforcement learning from human feedback [39].

3.2.4 *Context-Hallucination Bias.*

This bias emerges when LLMs are used as recommenders or re-rankers in scenarios with long and rich context, which can be defined as:

• **Definition.** LLMs-based IR models may generate content that is inconsistent with the context.

There have been many studies showing that LLMs run the risk of generating content that is inconsistent with the context, especially in scenarios where the context is very long and multi-turn responses are needed [7, 77, 90, 124]. When using LLMs as recommenders in scenarios with complex context such as multiple rounds of conversation history and user portrait information, LLMs may give the items that are contradictory with the conversation history and user preferences. This emerges when LLMs cannot understand the context or fail to maintain consistency throughout the conversation [193], which is mainly because LLMs still have limitations in processing long texts [90]. Therefore, the main research to mitigate the context-hallucination bias focuses on improving the memory and processing capabilities of LLMs for long texts [19, 157, 173].

3.3 Bias in Result Evaluation

When adopting LLMs as result evaluators in IR systems, the following three types of bias emerge, including selection bias, style bias, and egocentric bias.

3.3.1 *Selection Bias.*

A primary challenge in utilizing LLMs as evaluators is that they are sensitive to the order/ID tokens of candidate responses, a phenomenon known as selection bias:

- **Definition.** LLM-based evaluators may favor the responses at specific positions or with specific ID tokens.

For example, Zheng et al. [195] have demonstrated that gpt-3.5-turbo exhibits a preference for choice "C", while llama-30B shows a preference for choice "A" across various benchmarks. Other works have revealed a common tendency among LLMs to favor responses positioned at the first position [15, 71, 116, 155, 196]. Selection bias may stem from an imbalance in how answers of different positions or with distinct ID options are represented in the training data [195, 196]. Further investigation reveals that this bias tends to amplify when LLM-based evaluators are uncertain about the prediction between the top-ranked choices[155, 195].

To address selection bias, several approaches have been explored: (1) Prompting: Simple prompt-based methods, such as incorporating few-shot examples or employing chain-of-thought and reference-guided judgment, have been proposed [70, 116, 196]. (2) Data Augmentation: Strategies like position or token switching aim to eliminate selection bias by diversifying the evaluation context [21, 116, 155, 196, 198]. While these methods can enhance the robustness of evaluations, they are often time-consuming and costly. (3) Rebalancing: Zheng et al. [195] utilized probability decomposition techniques to estimate the prior distribution of specific positions or tokens associated with responses, which helps in aligning evaluations closer to objective standards. Wang et al. [155] have developed a calibration framework that integrates Human-in-the-Loop to calculate balanced position diversity entropy for final selection. Despite the availability of these strategies, effectively mitigating selection bias in LLM-based evaluators remains a challenge, requiring further research for more efficient solutions.

3.3.2 *Style Bias.*

Style bias can be viewed as a form of aesthetic bias, where the appeal of presentation overshadows the substance, leading to a preference for responses that, while polished in appearance, might harbor factual inaccuracies:

- **Definition.** LLM-based evaluators may favor the responses with specific styles (e.g., longer responses).

For instance, several studies have identified a clear preference in LLMs for longer responses over shorter ones, emphasizing form over the actual quality of the content [92, 129, 168, 196]. Moreover, Chen et al. [15] have observed an inclination towards content with visually engaging elements, such as emojis or references, even when such content may include factual errors. Huang et al. [60] suggest that this bias may stem from the training process of LLMs, which emphasizes generating fluent and verbose responses, thereby inadvertently leading them to prefer these characteristics when employed as evaluators.

Addressing style bias remains challenging, with current strategies mainly counteract overemphasis on stylistic features through prompts. However, these measures are often insufficient, highlighting the need for modifications in LLM architecture and training approaches to mitigate this bias effectively in future work.

3.3.3 *Egocentric Bias.*

With LLMs being extensively utilized in the development of IR models, egocentric bias has emerged as a new bias during the automated evaluation conducted by LLMs [71, 91, 92, 178], which can be defined as follows:

- **Definition.** LLM-based evaluators prefer the responses generated by themselves or LLMs from the same family.

A recent work [92] has identified that language model-driven evaluation metrics, such as BARTScore [185], T5Score [122], and GPTScore [41], inherently favor texts produced by their underlying LMs, especially in summarization tasks. Liu et al. [91] and Zheng et al. [196] further highlighted that when acting as evaluators, LLMs demonstrate a clear bias towards outputs generated by themselves over those from other models or human contributors. This bias could stem from that the LLM may share the same for both the model development phase and result evaluation phase [91].

The emergence of egocentric bias introduces the risk of self-reinforcement for LLMs, particularly when they are further trained using rewards from LLM-based evaluations. This scenario can lead to LLMs overfitting to their own evaluation criteria, intensifying self-preference in next-generation model [91]. Current strategies for mitigating egocentric bias primarily involve employing diverse LLMs as evaluators to foster peer discussions [79], thereby reducing the preference for any specific LLM and enhancing the robustness of evaluation outcomes. However, this strategy inevitably increases the evaluation costs. Future research must explore more efficient solutions to ensure fair and unbiased evaluation.

4 CAUSE AND MITIGATION OF UNFAIRNESS

As summarized in Table 2, we will review the cause and the mitigation strategies for the unfairness problem of IR in the LLM era.

4.1 Fairness Concepts

Sociological researches acknowledge multiple cultural variations in perceptions of fairness [148, 149]. In IR systems, achieving fairness often entails ensuring that the retrieved documents or recommended items align with cultural values [83, 103], including principles such as gender equality [38, 83, 170, 190], addressing disadvantages [115, 169], and avoiding discriminatory language [43, 162].

Researchers have revealed that various multi-stakeholders involved in IR systems [1, 16], such as users and items, often have distinct perspectives on fairness considerations. In IR, user fairness and item fairness are often associated with two sociological concepts: equality and distributive justice [169].

Table 2: The taxonomy of different types of unfairness in the intersection between LLMs and IR systems.

Sourced Stage	Type	Mitigation Strategies				
		Data Sampling		Distribution Reconstruction		
		Data Augmentation	Data Filtering	Rebalancing	Regularization	Prompting
Data Collection	User Unfairness	[47, 95, 141, 150, 170, 190]	[108, 125]	[32, 111]	[12, 62, 121]	[38]
	Item Unfairness	[127, 204]	[50]	[64]		[38, 73]
Model Development	User Unfairness	[152]	[102, 133, 137, 152]	[54, 187]	[6, 46, 89, 112, 114, 156, 164, 199]	[32, 59, 180, 190]
	Item Unfairness	[205]	[25, 69]	[64]	[40]	[31, 82, 205]
Result Evaluation	User Unfairness	[67]	[81]			[8, 63, 113, 128, 181]
	Item Unfairness	[49]		[5, 135]		[130, 151, 154, 189, 191]

• **User Fairness.** Everyone should be treated the same and provided the same resources to succeed. This implies that the IR systems should deliver equitable and non-discriminatory information services to different users.

• **Item Fairness.** The resources should be equally distributed based on needs. This implies that the IR systems should afford more opportunities (*e.g.,* exposures) to weaker items, striving to equalize the opportunities across diverse items.

4.2 Unfairness in Data Collection

In this section, we will elucidate the underlying causes of unfairness in the data collection process and subsequently outline current mitigation strategies to address these issues.

4.2.1 User Unfairness. In the context of user unfairness, one primary cause stems from the existing taxonomic, discriminatory, and offensive content in the training data, disproportionately affecting specific groups. The inclusion of these contents within the existing material can be attributed to historical and cultural reasons [11, 109, 203], or they may be generated by LLMs [38]. For example, when training LLMs, it is common to encounter discriminatory content [3, 29, 38]. Discrimination against certain groups can also stem from unbalanced data collection, where the insufficient representation of diverse perspectives leads to unfair outcomes [47, 103]. The presence of unbalanced data can contribute to the perpetuation of historical and cultural stereotypes [36, 53, 72] or systematic influences [16].

Previous works have employed various methods to mitigate unfairness during data collection by redistributing the existing document corpus. Specifically, some work [47, 95, 170, 190] create matched pairs (*e.g.,* male or female) to ensure a more equitable dataset and other methods [33, 141, 150] add non-toxic examples for groups. Other approaches [32, 111] suggest using down-weighting samples containing social group or discriminated information. Moreover, other studies [108, 125] propose to filter out discriminated or taxonomic content from web-scale datasets. Finally, instruction fine-tuning or RLHF has also been shown to be effective in promoting fairness [95, 147].

4.2.2 Item Unfairness. For item unfairness, one primary cause is likely unbalanced data collection, where the insufficient representation of certain items leads to disparities in the IR process [64]. Another reason raised is that LLMs cannot only retrieve existing items but also generate new items, contributing to the potential introduction of novel content and perspectives [29, 48, 50, 73, 205]. However, these newly generated items may still encounter discrimination issues [50, 73].

To mitigate item unfairness during data collection, several studies have developed methods to generate non-discriminatory items. For instance, Zou et al. [204] and Rathod et al. [127] suggest using specific templates to enrich training data with a variety of safe and equitable question-answer pairs, thereby improving LLM-based IR models' training. Guenole et al. [50] introduce pseudo-item discrimination techniques for filtering out non-discriminated items. Additionally, other studies [38, 73] advocate for employing fairness-aware prompts to produce newly non-discriminatory items. Furthermore, Jiang et al. [64] proposes to re-weight different item samples to enhance item fairness.

4.3 Unfairness in Model Development

In this section, we will analyze the causes of unfairness in the model development phase and explore mitigating strategies to address and minimize these disparities.

4.3.1 User Unfairness. When adapting LLMs as information retrievers, researchers have observed that the extensive knowledge gained during pre-training may introduce risks of user unfairness [32, 47, 170, 190], highlighting the need for careful consideration and mitigation strategies in deploying such models. Studies [32, 47, 190] have shown that utilizing explicit user-sensitive attributes like gender or race in LLMs may lead to the generation of discriminated recommendation results or unfair answers to specific questions. Moreover, it has been observed that LLMs can learn implicit attributes, such as user names and email addresses, and utilize them to generate discriminated content [163, 170].

To address user unfairness in the model development phase, prior research has suggested mitigating unfairness during the fine-tuning process. Some studies [32] investigate how different intersectional prompts affect recommendation fairness and UP5 [59, 180] propose to conduct prompt tuning to get an effective fair-aware prompt. Han et al. [54], Zayed et al. [187] proposes to set different weights of loss for different samples containing discriminated content. Meanwhile, various studies [46, 89, 114, 156, 164, 199] explore the incorporation of a fairness-aware regularizer to assist LLMs-based models in generating more equitable content. Wang et al. [152] proposes to remove unfair information from LLMs-based embeddings by generating adversarial examples. There are also some works [102, 133, 137, 152] that recommend employing a filtering-list approach or comparing model outputs with safe samples to proactively prevent the inclusion of discriminatory words and enhance the fairness of the generated content.

4.3.2 Item Unfairness. In the realm of item unfairness, previous studies have revealed that LLMs-based recommendation models are more prone to generating unfair outcomes for items when compared to traditional models [64]. Moreover, certain works [186] have also found that LLMs will also recommend unfair job opportunities to users. The embedding of item unfairness in the model development process can contribute to increased item polarization through reinforcement, potentially creating echo chambers that limit users' exposure to diverse perspectives [30].

Efforts to address item unfairness encompass a range of strategies: Zu et al. [205] have employed prompt-based learning to train GPT-2, leveraging this approach to generate distractors for fill-in-the-blank vocabulary items; some studies [25, 69] propose to utilize decoding strategies to decrease the probability of existing tokens/items; and Friedrich et al. [40] suggests integrating a fairness term into the LLMs-based diffusion process to introduce a shifting fairness consideration, aimed at generating new items that are less likely to contain discriminatory elements. Moreover, Jiang et al. [64] advocates for the re-weighting of different items to effectively mitigate unfairness in recommendation tasks. Other studies [31, 82] propose some prompt-aware methods to mitigate provider fairness.

4.4 Unfairness in Result Evaluation

To assess the fairness performance of IR models effectively, it is essential to measure the distributions of social values, which represent the fairness objectives inherent in the evaluation process [14]. However, human evaluation demands significant labor resources. Therefore, recently, LLMs have been employed to simulate human or real systems to facilitate evaluation processes efficiently.

4.4.1 User Unfairness. Similarly, user unfairness typically arises when LLMs-based evaluators fail to accurately simulate the behaviors exhibited by real humans. For example, Zhang et al. [194] propose leveraging psychological knowledge to assess the simulated human ability of LLMs, but they discover that LLMs frequently exhibit certain group behavior towards certain human groups.

To enhance the capability of LLMs as fair evaluators for IR systems, previous studies have devised several methods. Approaches include designing or learning specific prompts informed by psychological insights to better simulate diverse human groups [63, 113, 128, 181], augmenting training data with additional human personality information to refine LLMs as evaluators [67], and adopting innovative techniques like the unsupervised constructed personalized lexicon (UBPL) to manipulate their individual characteristics [81]. Furthermore, Bai et al. [8] proposes four stages to identifying discrimination patterns in queries.

4.4.2 Item Unfairness. In evaluating item unfairness, when turning from discriminant style to generation style, a primary concern arises from attributing credit to the generated items [5, 49, 130, 135, 151], as achieving item fairness necessitates tracing this credit back to the item provider for a comprehensive assessment.

To tackle these challenges, several studies [154, 189, 191] have developed IR agents that use world knowledge to fairly distribute item credits, aiming to reduce biases and enhance fairness. Shi et al. [135] employs the MIN-K% Prob technique to check whether an item exists in the training data. Meanwhile, Akyürek et al. [5], Grosse et al. [49] leverage influence functions and embedding similarities for item credit assessment. Additionally, other research [130, 151] explores watermarking techniques for item tracking.

5 CONCLUSION AND FUTURE DIRECTIONS

This survey has delved into the emergence of new bias and unfairness challenges within IR systems in this LLM era. We have established a unified framework to understand these issues as distribution mismatch problems and systematically categorized mitigation strategies into data sampling and distribution reconstruction approaches. Through an in-depth review of fifteen types of bias and unfairness, along with their corresponding mitigation strategies, we provide a comprehensive overview of the current progress. Despite the considerable attention given to this topic, we identify some important problems for further exploration.

Biases and Unfairness in IR Feedback Loops. In real IR systems, the interaction between users, models, and information forms feedback loops that impact each other over time. These loops can significantly shape user perceptions and preferences based on the information they are exposed to. This interaction, in turn, influences the training data, creating a cycle that may reinforce existing biases and unfairness. Novel strategies to interrupt the feedback loops are essential for mitigating these issues.

Unified Mitigation Framework. Current methods primarily address individual instances of bias or unfairness, but in the future, we should consider unified solutions. This is because various types of bias and unfairness are not isolated. Presenting a unified framework can facilitate a deeper understanding of these relationships, enabling methods for addressing different types of bias and unfairness to complement each other. Our proposed unified perspective offers a potential direction to address these issues simultaneously.

Theoretical Analysis and Guarantees. The current exploration of bias and unfairness within the intersection between LLMs and IR systems has predominantly been through empirical studies. However, there is a critical need for robust theoretical analysis to augment these empirical findings. Future efforts should focus on developing more rigorous analytical frameworks.

Better Benchmarks and Evaluation. Most benchmarks currently utilized to study bias and unfairness within simulated environments. There is a crucial need for collecting large-scale, real-world datasets to enhance the evaluations and broaden research horizons. Additionally, as LLMs increasingly draw upon existing online data to train subsequent generations, dynamic benchmarks are needed to be constructed. Consequently, future work can focus on exploring a systematic evaluation protocol for different bias and unfairness issues.

ACKNOWLEDGMENTS

This work was funded by the National Key R&D Program of China (2023YFA1008704), the National Natural Science Foundation of China (No. 62377044, No.62276248), Beijing Key Laboratory of Big Data Management and Analysis Methods, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, PCC@RUC, funds for building world-class universities (disciplines) of Renmin University of China, and the Youth Innovation Promotion Association CAS under Grants No.2023111.

REFERENCES

[1] Abdollahpouri et al. 2020. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction* (2020).

[2] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The unfairness of popularity bias in recommendation. *arXiv* (2019).

[3] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *AAAI*.

[4] Qingyao Ai, Ting Bai, et al. 2023. Information Retrieval Meets Large Language Models: A Strategic Report from Chinese IR Community. *AI Open* (2023).

[5] Ekin Akyürek, Tolga Bolukbasi, et al. 2022. Towards tracing factual knowledge in language models back to the training data. *arXiv* (2022).

[6] Yuntao Bai, Saurav Kadavath, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv* (2022).

[7] Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, et al. 2024. Longalign: A recipe for long context alignment of large language models. *arXiv* (2024).

[8] Yanhong Bai, Jiabao Zhao, Jinxin Shi, Tingjiang Wei, Xingjiao Wu, and Liang He. 2023. FairMonitor: A Four-Stage Automatic Framework for Detecting Stereotypes and Biases in Large Language Models. arXiv:2308.10397 [cs.CL]

[9] Keqin Bao, Jizhi Zhang, Wenjie Wang, et al. 2023. A bi-step grounding paradigm for large language models in recommendation systems. *arXiv* (2023).

[10] Emily M Bender, Timnit Gebru, et al. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *FAccT*.

[11] Camiel J Beukeboom et al. 2019. How stereotypes are shared through language: a review and introduction of the aocial categories and stereotypes communication (SCSC) framework. *Review of Communication Research* (2019).

[12] Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. *arXiv* (2019).

[13] Yihan Cao, Siyu Li, et al. 2023. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv* (2023).

[14] Yupeng Chang, Xu Wang, et al. 2023. A survey on evaluation of large language models. *TIST* (2023).

[15] Guiming Hardy Chen, Shunian Chen, et al. 2024. Humans or LLMs as the Judge? A Study on Judgement Biases. *arXiv* (2024).

[16] Jiawei Chen, Hande Dong, et al. 2023. Bias and debias in recommender system: A survey and future directions. *TOIS* (2023).

[17] Jifan Chen, Grace Kim, et al. 2023. Complex Claim Verification with Evidence Retrieved in the Wild. *arXiv* (2023).

[18] Xiaoyang Chen, Ben He, et al. 2024. Spiral of Silences: How is Large Language Model Killing Information Retrieval?–A Case Study on Open Domain Question Answering. *acl* (2024).

[19] Yukang Chen, Shengju Qian, et al. 2023. Longlora: Efficient fine-tuning of long-context large language models. *arXiv* (2023).

[20] I Chern, Steffi Chern, et al. 2023. FacTool: Factuality Detection in Generative AI–A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. *arXiv* (2023).

[21] Steffi Chern, Ethan Chern, et al. 2024. Can Large Language Models be Trusted for Evaluation? Scalable Meta-Evaluation of LLMs as Evaluators via Agent Debate. *arXiv* (2024).

[22] Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations? *ACL* (2023).

[23] Zhumin Chu, Qingyao Ai, Yiteng Tu, Haitao Li, and Yiqun Liu. 2024. PRE: A Peer Review Based Large Language Model Evaluator. *arXiv* (2024).

[24] Yung-Sung Chuang, Yujia Xie, et al. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv* (2023).

[25] John Joon Young Chung et al. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. *arXiv* (2023).

[26] Sunhao Dai, Weihao Liu, et al. 2024. Cocktail: A Comprehensive Information Retrieval Benchmark with LLM-Generated Documents Integration. *Findings of ACL* (2024).

[27] Sunhao Dai, Ninglu Shao, et al. 2023. Uncovering ChatGPT's Capabilities in Recommender Systems. In *RecSys*.

[28] Sunhao Dai, Yuqi Zhou, et al. 2024. Neural Retrievers are Biased Towards LLM-Generated Content. *KDD* (2024).

[29] Debarati Das, Karin De Langis, et al. 2024. Under the Surface: Tracking the Artifactuality of LLM-Generated Data. *arXiv* (2024).

[30] Michela Del Vicario, Gianna Vivaldo, et al. 2016. Echo chambers: Emotional contagion and group polarization on facebook. *Scientific reports* (2016).

[31] Yashar Deldjoo. 2024. Understanding Biases in ChatGPT-based Recommender Systems: Provider Fairness, Temporal Stability, and Recency. *arXiv* (2024).

[32] Yashar Deldjoo and Tommaso di Noia. 2024. CFaiRLLM: Consumer Fairness Evaluation in Large-Language Model Recommender System.

[33] Lucas Dixon, John Li, et al. 2018. Measuring and mitigating unintended bias in text classification. In *AAAI*.

[34] Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv* (2020).

[35] Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2021. Evaluating groundedness in dialogue systems: The begin benchmark. *arXiv* (2021).

[36] Naomi Ellemers. 2018. Gender stereotypes. *Annual review of psychology* (2018).

[37] Wenqi Fan, Zihuai Zhao, et al. 2023. Recommender systems in the era of large language models (llms). *arXiv* (2023).

[38] Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2024. Bias of AI-generated content: an examination of news produced by large language models. *Scientific Reports* (2024).

[39] Patrick Fernandes, Aman Madaan, and otherss. 2023. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *TACL* (2023).

[40] Felix Friedrich, Manuel Brack, et al. 2023. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv* (2023).

[41] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv* (2023).

[42] Yao Fu, Rameswar Panda, et al. 2024. Data Engineering for Scaling Language Models to 128K Context. *arXiv* (2024).

[43] Isabel O Gallegos, Ryan A Rossi, et al. 2023. Bias and fairness in large language models: A survey. *arXiv* (2023).

[44] Luyu Gao, Zhuyun Dai, et al. 2022. Rarr: Researching and revising what language models say, using language models. *arXiv* (2022).

[45] Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based nlg evaluation: Current status and challenges. *arXiv* (2024).

[46] Aparna Garimella, Akhash Amarnath, et al. 2021. He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation. In *ACL Findings*.

[47] Somayeh Ghanbarzadeh, Yan Huang, et al. 2023. Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. *arXiv* (2023).

[48] Friedrich M Götz et al. 2023. Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods* (2023).

[49] Roger Grosse, Juhan Bae, et al. 2023. Studying large language model generalization with influence functions. *arXiv* (2023).

[50] Nigel Guenole, Andrew Samo, et al. 2024. Pseudo-Discrimination Parameters from Language Embeddings. (2024).

[51] Suriya Gunasekar, Yi Zhang, et al. 2023. Textbooks Are All You Need. *arXiv* (2023).

[52] Izzeddin Gur, Hiroki Furuta, et al. 2023. A real-world webagent with planning, long context understanding, and program synthesis. *arXiv* (2023).

[53] Elizabeth L Haines, Kay Deaux, and Nicole Lofaro. 2016. The times they are a-changing… or are they not? A comparison of gender stereotypes, 1983–2014. *Psychology of Women Quarterly* (2016).

[54] Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Balancing out bias: Achieving fairness through balanced training. *arXiv* (2021).

[55] Hans WA Hanley and Zakir Durumeric. 2023. Machine-Made Media: Monitoring the Mobilization of Machine-Generated Articles on Misinformation and Mainstream News Websites. *arXiv* (2023).

[56] Hosein Hasanbeig, Hiteshi Sharma, et al. 2023. Allure: A systematic protocol for auditing and improving llm-based evaluation of text using iterative in-context-learning. *arXiv* (2023).

[57] Zhankui He, Zhouhang Xie, et al. 2023. Large language models as zero-shot conversational recommenders. In *CIKM*.

[58] Yupeng Hou, Junjie Zhang, et al. 2024. Large Language Models are Zero-Shot Rankers for Recommender Systems. In *ECIR*.

[59] Wenyue Hua, Yingqiang Ge, et al. 2023. Up5: Unbiased foundation model for fairness-aware recommendation. *arXiv* (2023).

[60] Hui Huang, Yingqi Qu, et al. 2024. An Empirical Study of LLM-as-a-Judge for LLM Evaluation: Fine-tuned Judge Models are Task-specific Classifiers. *arXiv* (2024).

[61] Lei Huang, Weijiang Yu, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv* (2023).

[62] Po-Sen Huang, Huan Zhang, et al. 2019. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv* (2019).

[63] Guangyuan Jiang, Manjie Xu, et al. 2024. Evaluating and inducing personality in pre-trained language models. *NeurIPS* (2024).

[64] Meng Jiang, Keqin Bao, et al. 2024. Item-side Fairness of Large Language Model-based Recommendation System. *arXiv* (2024).

[65] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *ICML*.

[66] Cheongwoong Kang and Jaesik Choi. 2023. Impact of co-occurrence on factual knowledge of large language models. *arXiv* (2023).

[67] SR Karra, ST Nguyen, and T Tulabandhula. 2023. Estimating the personality of white-box language models. *CoRR, abs/2204.12000* (2023).

[68] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A* (2024).

[69] Minbeom Kim, Hwanhee Lee, et al. 2022. Critic-guided decoding for controlled text generation. *arXiv* (2022).

[70] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2023. Evallm: Interactive evaluation of large language model prompts on user-defined

criteria. *arXiv* (2023).

[71] Ryan Koo, Minhwa Lee, et al. 2023. Benchmarking cognitive biases in large language models as evaluators. *arXiv* (2023).

[72] Faisal Ladhak, Esin Durmus, et al. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *EACL*.

[73] Antonio Laverghetta Jr and John Licato. 2023. Generating better items for cognitive assessments using large language models. In *BEA Workshop 2023*.

[74] Katherine Lee, Daphne Ippolito, et al. 2021. Deduplicating training data makes language models better. *arXiv* (2021).

[75] Nayeon Lee, Wei Ping, et al. 2022. Factuality enhanced language models for open-ended text generation. *NeurIPS* (2022).

[76] Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *EMNLP*.

[77] Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023. LooGLE: Can Long-Context Language Models Understand Long Contexts? *arXiv* (2023).

[78] Lei Li, Yongfeng Zhang, Dugang Liu, and Li Chen. 2023. Large language models for generative recommendation: A survey and visionary discussions. *arXiv* (2023).

[79] Ruosen Li, Teerth Patel, and Xinya Du. 2023. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv* (2023).

[80] Shaobo Li, Xiaoguang Li, et al. 2022. How pre-trained language models capture factual knowledge? a causal-inspired analysis. *arXiv* (2022).

[81] Tianlong Li, Xiaoqing Zheng, and Xuanjing Huang. 2023. Tailoring personality traits in large language models via unsupervisedly-built personalized lexicons. *arXiv* (2023).

[82] Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2023. A preliminary study of chatgpt on news recommendation: Personalization, provider fairness, fake news. *arXiv* (2023).

[83] Yunqi Li, Hanxiong Chen, et al. 2023. Fairness in recommendation: Foundations, methods, and applications. *ACM Transactions on Intelligent Systems and Technology* (2023).

[84] Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A survey on fairness in large language models. *arXiv* (2023).

[85] Zongjie Li, Chaozheng Wang, et al. 2023. Split and merge: Aligning position biases in large language model based evaluators. *arXiv* (2023).

[86] Zhen Li, Xiaohan Xu, et al. 2024. Leveraging large language models for nlg evaluation: A survey. *arXiv* (2024).

[87] Jianghao Lin, Xinyi Dai, et al. 2023. How Can Recommender Systems Benefit from Large Language Models: A Survey. *arXiv* (2023).

[88] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv* (2021).

[89] Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019. Does gender matter? towards fairness in dialogue systems. *arXiv* (2019).

[90] Nelson F Liu, Kevin Lin, et al. 2024. Lost in the middle: How language models use long contexts. *TACL* (2024).

[91] Yang Liu, Dan Iter, et al. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *EMNLP*.

[92] Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2023. Llms as narcissistic evaluators: When ego inflates evaluation scores. *arXiv* (2023).

[93] Yang Liu, Yuanshun Yao, et al. 2023. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. *arXiv* (2023).

[94] Adian Liusie, Yassir Fathullah, and Mark JF Gales. 2024. Teacher-Student Training for Debiasing: General Permutation Debiasing for Large Language Models. *arXiv* (2024).

[95] Kaiji Lu, Piotr Mardziel, et al. 2020. Gender bias in neural natural language processing. *Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday* (2020).

[96] Sichun Luo, Bowei He, et al. 2023. RecRanker: Instruction Tuning Large Language Model as Ranker for Top-k Recommendation. *arXiv* (2023).

[97] Tianhui Ma, Yuan Cheng, Hengshu Zhu, and Hui Xiong. 2023. Large Language Models are Not Stable Recommender Systems. *arXiv* (2023).

[98] Alex Mallen, Akari Asai, et al. 2022. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv* (2022).

[99] Christopher D Manning. 2009. *An introduction to information retrieval*. Cambridge university press.

[100] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv* (2020).

[101] Nick McKenna, Tianyi Li, et al. 2023. Sources of Hallucination by Large Language Models on Inference Tasks. *arXiv* (2023).

[102] Nicholas Meade, Spandana Gella, et al. 2023. Using in-context learning to improve dialogue safety. *arXiv* (2023).

[103] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* (2021).

[104] Todor Mihaylov, Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Georgi D Georgiev, and Ivan Kolev Koychev. 2018. The dark side of news community forums: Opinion manipulation trolls. *Internet Research* (2018).

[105] Sewon Min, Kalpesh Krishna, et al. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. *arXiv* (2023).

[106] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. *arXiv* (2021).

[107] Reiichiro Nakano, Jacob Hilton, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv* (2021).

[108] Helen Ngo, Cooper Raterink, et al. 2021. Mitigating harm in language models with conditional-likelihood filtration. *arXiv* (2021).

[109] Eirini Ntoutsi, Pavlos Fafalios, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2020).

[110] Yasumasa Onoe, Michael JQ Zhang, Eunsol Choi, and Greg Durrett. 2022. Entity cloze by date: What LMs know about unseen entities. *arXiv* (2022).

[111] Hadas Orgad and Yonatan Belinkov. 2022. BLIND: Bias removal with no demographics. *arXiv* (2022).

[112] Long Ouyang, Jeffrey Wu, et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS* (2022).

[113] Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv* (2023).

[114] SunYoung Park, Kyuri Choi, Haeun Yu, and Youngjoong Ko. 2023. Never too late to learn: Regularizing gender bias in coreference resolution. In *WSDM*.

[115] Gourab K Patro, Arpita Biswas, Niloy Ganguly, Krishna P Gummadi, and Abhijnan Chakraborty. 2020. Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In *WWW*.

[116] Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv* (2023).

[117] Cara L Phillips and Timothy R Vollmer. 2012. Generalized instruction following with pictorial prompts. *Journal of Applied Behavior Analysis* (2012).

[118] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2022. Fairness in rankings and recommendations: an overview. *The VLDB Journal* (2022).

[119] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv* (2022).

[120] Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv* (2023).

[121] Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. *arXiv* (2019).

[122] Yiwei Qin, Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2022. T5score: Discriminative fine-tuning of generative evaluation metrics. *arXiv* (2022).

[123] Zhen Qin, Rolf Jagerman, et al. 2023. Large language models are effective text rankers with pairwise ranking prompting. *arXiv* (2023).

[124] Zexuan Qiu, Jingjing Li, Shijue Huang, Wanjun Zhong, and Irwin King. 2024. CLongEval: A Chinese Benchmark for Evaluating Long-Context Large Language Models. *arXiv* (2024).

[125] Colin Raffel, Noam Shazeer, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* (2020).

[126] Ori Ram, Yoav Levine, et al. 2023. In-context retrieval-augmented language models. *arXiv* (2023).

[127] Manav Rathod, Tony Tu, and Katherine Stasaski. 2022. Educational Multi-Question Generation for Reading Comprehension. In *BEA Workshop 2022*.

[128] Mustafa Safdari, Greg Serapio-García, et al. 2023. Personality traits in large language models. *arXiv* (2023).

[129] Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv* (2023).

[130] Tom Sander, Pierre Fernandez, et al. 2024. Watermarking Makes Language Models Radioactive. *arXiv* (2024).

[131] Victor Sanh, Albert Webson, , et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv* (2021).

[132] Irwin G Sarason, Gregory R Pierce, and Barbara R Sarason. 2014. *Cognitive interference: Theories, methods, and findings*. Routledge.

[133] Patrick Schramowski, Cigdem Turan, et al. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence* (2022).

[134] Nikhil Sharma, Q Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effects of LLM-Powered Search Systems on Diverse Information Seeking. *In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (2024).

[135] Weijia Shi, Anirudh Ajith, et al. 2023. Detecting pretraining data from large language models. *arXiv* (2023).

[136] Weijia Shi and Sewonand others Min. 2023. Replug: Retrieval-augmented black-box language models. *arXiv* (2023).

[137] Kurt Shuster, Jing Xu, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv* (2022).

[138] Amit Singhal et al. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.* (2001).

[139] Karan Singhal, Tao Tu, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv* (2023).

[140] Kyle Dylan Spurlock, Cagla Acun, Esin Saka, and Olfa Nasraoui. 2024. ChatGPT for Conversational Recommendation: Refining Recommendations by Reprompting with Feedback. *arXiv* (2024).

[141] Hao Sun, Zhexin Zhang, et al. 2022. MoralDial: A framework to train and evaluate moral dialogue systems via moral discussions. *arXiv* (2022).

[142] Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent. *arXiv* (2023).

[143] Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2022. Recitation-augmented language models. *arXiv* (2022).

[144] Ekaterina Svikhnushina and Pearl Pu. 2023. Approximating Human Evaluation of Social Chatbots with Prompting. *arXiv* (2023).

[145] Hexiang Tan, Fei Sun, et al. 2024. Blinded by Generated Contexts: How Language Models Merge Generated and Retrieved Contexts for Open-Domain QA? *ACL* (2024).

[146] Raphael Tang, Xinyu Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. 2023. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. *arXiv* (2023).

[147] Hugo Touvron, Louis Martin, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv* (2023).

[148] Tom R Tyler and E Allan Lind. 2002. Procedural justice. In *Handbook of justice research in law*.

[149] Tom R Tyler and Heather J Smith. 1995. Social justice and social movements. (1995).

[150] Megan Ung, Jing Xu, and Y-Lan Boureau. 2021. Saferdialogues: Taking feedback gracefully after conversational safety failures. *arXiv* (2021).

[151] Jingtan Wang, Xinyang Lu, et al. 2023. WASA: Watermark-based source attribution for large language model-generated data. *arXiv* (2023).

[152] Liwen Wang, Yuanmeng Yan, Keqing He, Yanan Wu, and Weiran Xu. 2021. Dynamically disentangling social bias from task-oriented representations with adversarial attack. In *NAACL*.

[153] Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query Expansion with Large Language Models. *arXiv* (2023).

[154] Lei Wang, Jingsen Zhang, et al. 2023. Recagent: A novel simulation paradigm for recommender systems. *arXiv* (2023).

[155] Peiyi Wang, Lei Li, et al. 2023. Large language models are not fair evaluators. *arXiv* (2023).

[156] Rui Wang, Pengyu Cheng, and Ricardo Henao. 2023. Toward fairness in text generation via mutual information minimization based on importance sampling. In *AISTATS*.

[157] Weizhi Wang, Li Dong, et al. 2024. Augmenting language models with long-term memory. *NeurIPS* (2024).

[158] Xi Wang, Hossein A Rahmani, Jiqun Liu, and Emine Yilmaz. 2023. Improving Conversational Recommendation Systems via Bias Analysis and Language-Model-Enhanced Data Augmentation. *arXiv* (2023).

[159] Xuezhi Wang, Jason Wei, et al. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171 [cs.CL]

[160] Yizhong Wang, Yeganeh Kordi, et al. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv* (2022).

[161] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems* (2023).

[162] Jiaxin Wen, Pei Ke, et al. 2023. Unveiling the implicit toxicity in large language models. *arXiv* (2023).

[163] Robert Wolfe and Aylin Caliskan. 2021. Low frequency names exhibit bias and overfitting in contextualizing language models. *arXiv* (2021).

[164] Tae-Jin Woo, Woo-Jeoung Nam, Yeong-Joon Ju, and Seong-Whan Lee. 2023. Compensatory debiasing for gender imbalances in language models. In *ICASSP*.

[165] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Hong Lin. 2023. Ai-generated content (aigc): A survey. *arXiv* (2023).

[166] Likang Wu, Zhaopeng Qiu, Zhi Zheng, Hengshu Zhu, and Enhong Chen. 2023. Exploring large language model for graph data understanding in online job recommendations. *arXiv* (2023).

[167] Likang Wu, Zhi Zheng, et al. 2023. A Survey on Large Language Models for Recommendation. *arXiv* (2023).

[168] Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *arXiv* (2023).

[169] Chen Xu, Sirui Chen, et al. 2023. P-MMF: Provider Max-min Fairness Re-ranking in Recommender System. In *Proceedings of the ACM Web Conference 2023*.

[170] Chen Xu, Wenjie Wang, Yuxin Li, Liang Pang, Jun Xu, and Tat-Seng Chua. 2023. Do llms implicitly exhibit user discrimination in recommendation? an empirical study. *arXiv* (2023).

[171] Jun Xu, Xiangnan He, and Hang Li. 2018. Deep Learning for Matching in Search and Recommendation. In *SIGIR*.

[172] Lanling Xu, Junjie Zhang, et al. 2024. Prompting Large Language Models for Recommender Systems: A Comprehensive Framework and Empirical Analysis. *arXiv* (2024).

[173] Peng Xu, Wei Ping, et al. 2023. Retrieval meets long context large language models. *arXiv* (2023).

[174] Shicheng Xu, Danyang Hou, et al. 2023. AI-Generated Images Introduce Invisible Relevance Bias to Text-Image Retrieval. *arXiv* (2023).

[175] Shicheng Xu, Liang Pang, et al. 2024. List-aware reranking-truncation joint model for search and retrieval-augmented generation. *WWW* (2024).

[176] Shicheng Xu, Liang Pang, et al. 2024. Search-in-the-Chain: Towards the Accurate, Credible and Traceable Content Generation for Complex Knowledge-intensive Tasks. *WWW* (2024).

[177] Shicheng Xu, Liang Pang, et al. 2024. Unsupervised Information Refinement Training of Large Language Models for Retrieval-Augmented Generation. *arXiv* (2024).

[178] Wenda Xu, Guanglei Zhu, et al. 2024. Perils of Self-Feedback: Self-Bias Amplifies in Large Language Models. *arXiv* (2024).

[179] Jintang Xue, Yun-Cheng Wang, et al. 2023. Bias and fairness in chatbots: An overview. *arXiv* (2023).

[180] Ke Yang, Charles Yu, Yi R Fung, Manling Li, and Heng Ji. 2023. Adept: A debiasing prompt framework. In *AAAI*.

[181] Tao Yang, Tianyuan Shi, et al. 2023. PsyCoT: Psychological Questionnaire as Powerful Chain-of-Thought for Personality Detection. *arXiv* (2023).

[182] Zonghan Yang, Xiaoyuan Yi, Peng Li, Yang Liu, and Xing Xie. 2022. Unified detoxifying and debiasing in language generation via inference-time adaptive optimization. *arXiv* (2022).

[183] Seonghyeon Ye, Hyeonbin Hwang, et al. 2023. Investigating the effectiveness of task-agnostic prefix prompt for instruction following. *arXiv* (2023).

[184] Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023. Improving Language Models via Plug-and-Play Retrieval Feedback. *arXiv* (2023).

[185] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *NeurIPS* (2021).

[186] Ali Zarifhonarvar. 2023. Economics of chatgpt: A labor market view on the occupational impact of artificial intelligence. *SSRN 4350925* (2023).

[187] Abdelrahman Zayed, Goncalo Mordido, Samira Shabanian, and Sarath Chandar. 2023. Should we attend more or less? modulating attention for fairness. *arXiv* (2023).

[188] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2021. Fairness in ranking: A survey. *arXiv* (2021).

[189] An Zhang, Leheng Sheng, Yuxin Chen, Hao Li, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2023. On generative agents in recommendation. *arXiv* (2023).

[190] Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. In *RecSys*.

[191] Junjie Zhang, Yupeng Hou, et al. 2023. Agentcf: Collaborative learning with autonomous language agents for recommender systems. *arXiv* (2023).

[192] Yang Zhang, Fuli Feng, et al. 2021. Causal intervention for leveraging popularity bias in recommendation. In *SIGIR*.

[193] Yue Zhang, Yafu Li, , et al. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv:2309.01219 [cs.CL]

[194] Zhexin Zhang, Leqi Lei, et al. 2023. Safetybench: Evaluating the safety of large language models with multiple choice questions. *arXiv* (2023).

[195] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. In *ICLR*.

[196] Lianmin Zheng, Wei-Lin Chiang, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS* (2024).

[197] Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why Does ChatGPT Fall Short in Answering Questions Faithfully? *arXiv* (2023).

[198] Zhi Zheng, Zhaopeng Qiu, Xiao Hu, Likang Wu, Hengshu Zhu, and Hui Xiong. 2023. Generative job recommendations with large language model. *arXiv* (2023).

[199] Fan Zhou, Yuzhou Mao, et al. 2023. Causal-debias: Unifying debiasing in pre-trained language models and fine-tuning via causal invariant learning. In *ACL*.

[200] Yuqi Zhou, Sunhao Dai, et al. 2024. Source Echo Chamber: Exploring the Escalation of Source Bias in User, Data, and Recommender System Feedback Loop. *arXiv* (2024).

[201] Yutao Zhu, Huaying Yuan, et al. 2023. Large language models for information retrieval: A survey. *arXiv* (2023).

[202] Ziwei Zhu, Yun He, Xing Zhao, and James Caverlee. 2021. Popularity bias in dynamic recommendation. In *KDD*.

[203] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv* (2023).

[204] Bowei Zou, Pengfei Li, Liangming Pan, and Aiti Aw. 2022. Automatic true/false question generation for educational purpose. In *BEA Workshop 2022*.

[205] Jiyun Zu, Ikkyu Choi, and Jiangang Hao. 2023. Automated distractor generation for fill-in-the-blank items using a prompt-based learning approach. *Psychological Testing and Assessment Modeling* (2023).