Summary of Unifying Bias and Unfairness in Information Retrieval

Introduction

The rapid development of large language models and Bert-based models has brought many changes to the development of information retrieval. Some of these developments are very promising, but some create a more significant challenge for the future of the integrity of the information ecosystem. Seemingly, all of these challenges stem from the misuse or abuse of these models. Sifting through a comprehensive review presented in their paper, the authors discuss some of the growing problems of potential misuse of LLM-based IR systems, presenting them as frequent cases of distribution mismatch. Hence, the authors propose a foundational definition for the growing body of strategies for mitigating bias via distribution integration.

Research findings have revealed that Large Language Models (LLMs) can sometimes provide information that isn't entirely accurate and may even show a preference for the content they have generated themselves. They can also unintentionally reinforce stereotypes and present discriminatory material, making the divide between different socio-economic groups even wider.

It acknowledges the scattered state of current research, which often lacks a clear definition of bias and unfairness. Its objective is to present a comprehensive and unified view on these issues. It considers bias and unfairness as problems arising from discrepancies in distribution—bias indicates a departure from the objective truth, while unfairness suggests a misalignment with societal values.
It has covered a set of unique problems related to bias and unfairness in three main steps of LLM integration into IR systems, including data collection, model development, and result evaluation. All of them are thoroughly considered in-depth, and the biases and mitigation proposals are based on more recent relevant works, explained to the fullest. Moreover, the authors of the source present an extensive set of definitions, features, and solutions, thus outlining

these problems for others. They also talk about unresolved problems and potential future challenges that should be addressed to further interest IR researchers and other stakeholders in the cumulative resolution of these issues.

The survey begins by reviewing how LLMs are integrated into Information Retrieval (IR) systems, creating new kinds of bias and unfairness issues at various stages, including data collection, model building, and result evaluation. It then formulates a matrix that views all these kinds of issues as distribution mismatches and proposes possible solutions based on that matrix. Data sampling techniques and distribution reconstruction methods are two broad categories of solutions. The present survey intends to provide a reliable basis for developing future, more dependable and equitable IR systems via a comprehensive examination of recent research.

***Importance of LLMs in IR*:**

Information Retrieval systems are essential aids for individuals in the modern era of information accessibility as they help people identify the necessary resources as rapidly and accurately as feasible. Large Language Models have dramatically transformed Information Retrieval systems. LLMs have enabled extracting LLM-created data and incorporating it as a new source of information, converting Intelligence Retrieval from retrieving to generating data, and using LLMs to assess Intelligence Retrieval systems' results.

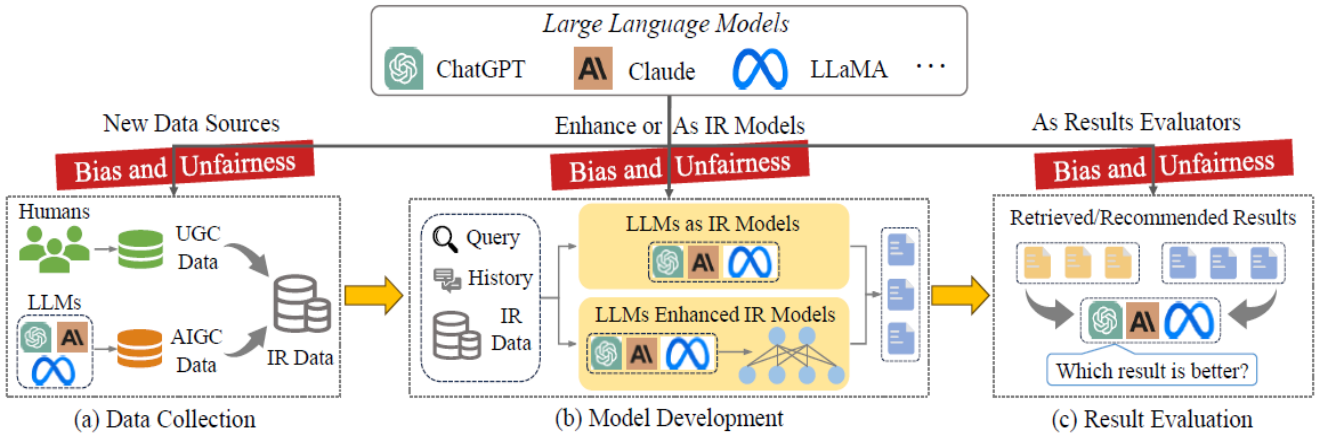

Figure 1: Overview of three stages of the intersection between LLMs and IR systems. (a) LLMs-generated content as new IR data sources. (b) Incorporating LLMs to enhance or as IR models. (c) Adopting LLMs as results evaluators in IR systems.

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.**

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. On one hand, LLMs have been deployed to refine key components of traditional IR systems, 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, directly generating responses to fulfil user queries.

***In-depth Taxonomies of Mitigation Strategies*:**

The main aim of the mitigation strategies is to arrange the distribution of retrieved data with a target distribution defined by either objective criteria or subjective social values.

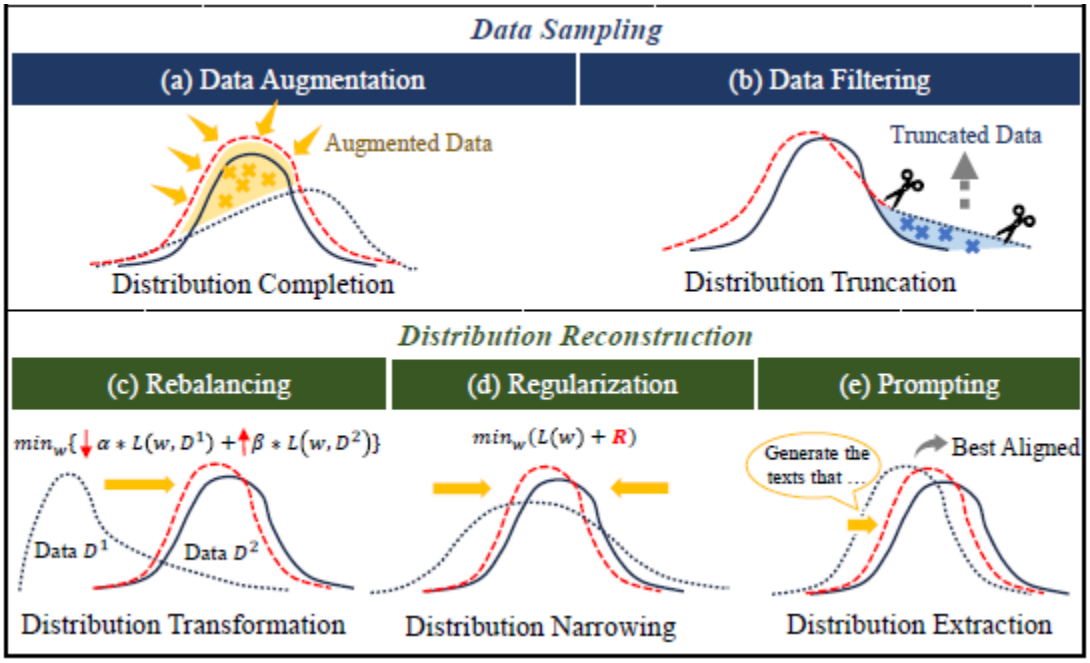

Figure 2: Illustration of different types of mitigation strategies

As shown in Figure 2, The two primary categories and their sub-strategies are:

- **Data Sampling**
  - Data Augmentation
  - Data Filtering
- **Distribution Reconstruction**
  - Rebalancing
  - Regularization
  - Prompting

Data Sampling directly modifies the data by selecting a small amount of data from a large dataset to perform analysis, processing, or testing.

- **Data Augmentation** serves as distribution completion, by generating a new data from the existing data by adding the additional information, also known as synthetic data, to approximate the target distribution more closely. To create the synthetic data various strategies are used such as counterfactual imputation and the addition of external data to narrow the bridge between the added and existing data, thus aiming to recover the real distribution more accurately.

- **Data Filtering** acts as distribution truncation, by selecting smaller sets of data from the dataset. This step is used to exclude, rearrange or allocate data to align it with the target distribution. To achieve this, techniques such as re-ranking and constrained beam search are used, serving as post-processing methods to ensure the retained distribution segment matches the target distribution.

Distribution Reconstruction is the process of reconstructing and adjusting the probability of predicted distribution, enabling more accurate modelling and prediction.

- **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.

These mitigation strategies provide more detailed, structured and more reasonable way to reduce the bias and improve the fairness of IR systems in IR systems.

***CAUSE AND MITIGATION OF BIAS*:**

As shown in Table 1, an in-depth review of different types of bias at the different stages of the intersection between LLMs and IR systems -

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				[27, 176]	
	Factuality Bias	[50, 120, 127, 177–179, 188]	[50, 149, 184]			[120, 145, 161, 177]
Model Development	Position Bias	[57, 96, 124, 148, 168, 195]		[97, 168]		[57]
	Popularity Bias	[160, 195]				[30, 57, 142]
	Instruction-Hallucination Bias	[107, 133, 162]			[38]	[118, 185]
	Context-Hallucination Bias	[7, 41]				
Result Evaluation	Selection Bias	[21, 23, 79, 85, 117, 157, 186, 200, 202]		[94, 157, 199]		[70, 117, 157, 200]
	Style Bias					[170, 200]
	Egocentric Bias	[79]		[91]		[55, 91]

Bias in Data Collection

- **Source Bias** - Large Language Models generated content becomes favoured over human-created text. Such bias is rooted in the differences between human and LLM representations, with the latter being more like trained neural models. To mitigate this bias, one should improve LLM relevance annotation performance on the human text.
- **Factuality Bias:** LLM-created texts included or emphasized content not common in human text, and some of it was untruth or fictional. Mitigation of this bias includes utilizing high-quality, factual learning data and applying external knowledge during inference to enhance factual knowledge.

Bias in Model Development

- **Position Bias**: Item placing within the input determines its likelihood that is often addressed via techniques such as data augmentation and rebalance.
- **Popularity Bias:** can be induced by the items used to train ELG; for example, its effect can be countered by promoting data diversification and modified prompts derived from Google Popular Queries.
- **Instruction-Hallucination Bias:** generates different results based on precise instructions, and it can be handled by using high-quality instruction data to finetune ELG or reinforcing learning based on human data.

- **Context-Hallucination Bias**: LLM was trained with data around ELG and generates information in a scenario that may not constitute those surrounding. Researchers have recommended on improving the context and LLM memories-processing characteristics.

Bias in Result Evaluation

- **Selection Bias**: LLMs may have preferences for certain response positions or types. This bias is typically addressed via diverse evaluation strategies and recalibrating model outputs.
- **Style Bias**: LLMs may prefer stylistically appealing or longer responses, which does not guarantee factual accuracy. The way to mitigate this bias would be too de-prioritize stylistic features in LLM training.
- **Egocentric Bias**: LLM may have a bias toward their outputs or outputs of models like the studied one. One may mitigate by using a diverse set of LLMs.

CAUSE AND MITIGATION OF UNFAIRNESS:

As shown in Table 2, the cause and the mitigation strategies for the unfairness problem of IR in the LLM era.

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	[32, 46, 95, 143, 152, 172, 194]	[109, 126]	[31, 112]	[13, 61, 122]	[37]
	Item Unfairness	[128, 207]	[49]	[64]		[37, 73]
Model Development	User Unfairness	[154]	[103, 135, 139, 154]	[53, 191]	[6, 45, 89, 113, 115, 158, 166, 203]	[31, 58, 182, 194]
	Item Unfairness	[208]	[25, 69]	[64]	[39]	[30, 82, 208]
Result Evaluation	User Unfairness	[67]	[81]			[8, 63, 114, 130, 183]
	Item Unfairness	[48]		[5, 137]		[132, 153, 156, 193, 195]

Fairness Concept

IR system fairness is theoretically guided by sociological various research about the perception of fairness in culture. Implementing fairness varies depending on one's culture, and this often requires conformance in IR systems such as gender bias and linguistics. Two fundamental concepts of fairness include:

- **User Fairness**: This is based on equity, and it pronounces that all users should be treated the same and granted the same resources.
- **Item Fairness:** This is based on the theory of distributive justice where resources will be distributed depending on an item's need, thus more support will be given to sicker items.

Unfairness in Data Collection

- **User Unfairness**: User unfairness arises from the inclusion of discriminatory content in the training data, which may be historically or socially prevalent or added by the LLMs. Approaches for user unfairness are combining matched pairs in datasets, incorporating non-toxic examples, downweighing biased samples and filtering data for discriminatory content.
- **Item Unfairness**: Item may refer to unfairness insufficiently representing certain items that may skew the assessments. LLMs generated new items may include novel content and new biases. Solutions to address unfairness item are creating non-discriminatory items with templates, using item discrimination, and reweighting items.

Unfairness in Model Development

- **User Unfairness**: This is caused by extensive pre-training of the LLMs that might have been riddled with biases. The solutions for this involve using intersectional prompts, re-weighting loss for biased samples, using fairness-aware regularizes and acting on the proactive-filtering mechanisms.
- **Item Unfairness**: LLM-based recommendation may amplify the preference to certain items for the recommendation models, thus enhancing the polarization. The only solution to counter this is using prompt-based learning, decoding tricks to reduce the sampling of the biased tokens and integrating fairness terms of the statistical levels of the model learning.

Unfairness in Result Evaluation

- **User Unfairness**: occurs when LLM-based evaluators cannot present or simulate a variety of human behaviours accurately. Psychologically insights-bases prompt design, data augmentation with human personality traits, personalized lexicons are some techniques to mitigating that.
- **Item Unfairness**: is the issue of being just to generate items by assigning them credit. It can be solved through item tracking via watermarking and influence functions to balance the credit.

Challenges

Some important issues and challenges to further explore are:

- ***Biases and Unfairness in IR Feedback Loops*:** The model of IR systems is such that the interaction between the model and user is dynamic which creates ever evolving feedback loops that affect each other over the time. These loops tend to change and influence the training data of the IR systems as they are exposed to different sources of information which impacts the biases and fairness.
- ***Unified Mitigation Framework*:** The existing methods only deals with the individual instances of bias and unfairness but in future there will be a need to focus on the unified solutions. This is an issue because the different types of biases and unfairness are not obscure but connected and showcase a unified framework.
- ***Theoretical Analysis and Guarantees*:** The current exploration of IR systems unfairness and bias between LLMs is based on, observation or experience rather than theory or pure logic. However, there is a critical need for robust theoretical analysis to augment these empirical findings.
- ***Better Benchmarks and Evaluation*:** The benchmarks for the bias and unfairness of LLMs and IR systems are produced through simulators. To enhance and deepen the studies, a large-scale real-world

data is required. There is a need for the dynamic benchmarks as the LLMs today is trained over online data.

Conclusion

This paper is about the new bias and unfairness challenges in the IR systems. A variety of mitigation strategies into data sampling and distribution reconstruction approaches is categorized systematically to deal with the distribution mismatch issues. Through an in-depth review of several types of bias and unfairness, along with their corresponding mitigation strategies, a comprehensive overview of the current progress is provided. This is a complete roadmap in the field of IR to understand the mitigation of bias and unfairness in IR systems giving major insights to develop more reliable and fair IR systems.