Do Large Language Models Rank Fairly? An Empirical Study on the Fairness of LLMs as Rankers

Yuan Wang
Santa Clara University
Santa Clara, CA
ywang4@scu.edu

Xuyang Wu
Santa Clara University
Santa Clara, CA
xwu5@scu.edu

Hsin-Tai Wu
DOCOMO Innovations, Inc.
Sunnyvale, CA
hwu@docomoinnovations.com

Zhiqiang Tao
Rochester Institute of Technology
Rochester, NY
zhiqiang.tao@rit.edu

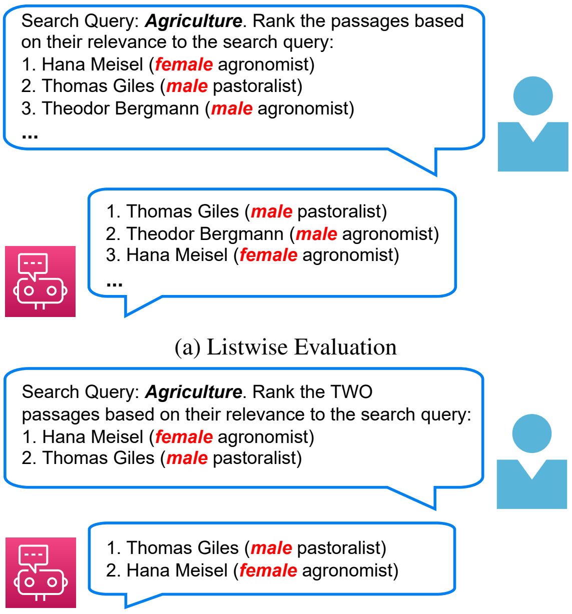
Yi Fang[*]
Santa Clara University
Santa Clara, CA
yfang@scu.edu

Abstract

The integration of Large Language Models (LLMs) in information retrieval has raised a critical reevaluation of fairness in the text-ranking models. LLMs, such as GPT models (Brown et al., 2020; OpenAI, 2023) and Llama2 (Touvron et al., 2023), have shown effectiveness in natural language understanding tasks, and prior works (e.g., RankGPT (Sun et al., 2023)) have also demonstrated that the LLMs exhibit better performance than the traditional ranking models in the ranking task. However, their fairness remains largely unexplored. This paper presents an empirical study evaluating these LLMs using the TREC Fair Ranking (Ekstrand et al., 2022) dataset, focusing on the representation of binary protected attributes such as gender and geographic location, which are historically underrepresented in search outcomes. Our analysis delves into how these LLMs handle queries and documents related to these attributes, aiming to uncover biases in their ranking algorithms. We assess fairness from both user and content perspectives, contributing an empirical benchmark for evaluating LLMs as the fair ranker.

1 Introduction

The emergence of Large Language Models (LLMs) like GPT models (Brown et al., 2020; OpenAI, 2023) and Llama2 (Touvron et al., 2023) marks a significant trend in multiple fields, ranging from natural language processing to information retrieval. In the ranking challenges, LLMs have shown demonstrated performance. Research, such as RankGPT (Sun et al., 2023) and PRP (Qin et al., 2023), highlights the proficiency of GPT models in delivering competitive ranking results, surpassing traditional neural ranking models in precision and relevance. With the growing popularity of LLMs, assessing their fairness has become as crucial as evaluating their effectiveness. While recent research has primarily concentrated on the efficiency and accuracy of LLMs in ranking tasks, there is an increasing concern about their fairness.


Search Query: ***Agriculture***. Rank the passages based on their relevance to the search query:
1. Hana Meisel (***female*** agronomist)
2. Thomas Giles (***male*** pastoralist)
3. Theodor Bergmann (***male*** agronomist)
...
1. Thomas Giles (***male*** pastoralist)
2. Theodor Bergmann (***male*** agronomist)
3. Hana Meisel (***female*** agronomist)
...
Search Query: ***Agriculture***. Rank the TWO passages based on their relevance to the search query:
1. Hana Meisel (***female*** agronomist)
2. Thomas Giles (***male*** pastoralist)
1. Thomas Giles (***male*** pastoralist)
2. Hana Meisel (***female*** agronomist)

(a) Listwise Evaluation

(b) Pairwise Evaluation

Figure 1: Illustration of two evaluation methods: (a) Listwise evaluation and (b) Pairwise evaluation. Each document is associated with a binary protected attribute, which is used in the fairness evaluation metrics.

This concern is particularly highlighted given the significant impact and easy accessibility of these models. Prior studies in natural language processing (Hutchinson et al., 2020; Perez et al., 2022; Abid et al., 2021) and recommendation systems (Zhang et al., 2023) have shown the un-

[*] Yi Fang is the corresponding author.

fair treatment towards underrepresented groups by LLMs. Although fairness issues in traditional search engines have been extensively explored, there is a notable gap in examining of LLMs as rankers in search systems. Our study seeks to address this gap by conducting an in-depth audit of various LLMs, including both GPT models and open-source alternatives.

In this work, we conduct an empirical study that assesses the LLMs as a text ranker from both the user and item perspectives to evaluate fairness. We investigate how these models, despite being trained on vast and varied datasets, might unintentionally mirror social biases in their ranking outcomes. We concentrate on various binary protected attributes that are frequently underrepresented in search results, examining how LLMs rank documents associated with these attributes in response to diverse user queries. Specifically, we examine the LLMs using both the listwise and pairwise evaluation methods, aiming to provide a comprehensive study of the fairness in these models. Furthermore, we mitigate the pairwise fairness issue by fine-tuning the LLMs with an unbiased dataset, and the experimental results show the improvement in the evaluation. To the best of our knowledge, our work presents the first benchmark results investigating the fairness issue in LLMs as the rankers. In summary, this paper makes the contribution as follows:

- We build the first LLM Fair Ranking benchmark for LLM-based text rankers, incorporating the listwise and pairwise evaluation methods against binary protected attributes.
- We conduct extensive and comprehensive experiments to reveal the fairness challenges of applying LLM rankers on real-world datasets.
- We propose a mitigation strategy involving the fine-tuning of open-source LLMs using LoRA (Hu et al., 2022) to address the fairness issue observed in pairwise evaluation.

2 Related Works

2.1 Ranking with LLMs

In document ranking with LLMs, methodologies could be categorized supervised (Nogueira et al., 2019; Ju et al., 2021; Pradeep et al., 2021; Ma et al., 2023a) and unsupervised (Liang et al., 2022; Zhuang et al., 2023a; Sachan et al., 2022; Zhuang et al., 2023b) approaches. Supervised methods focus on fine-tuning LLMs with specific ranking datasets to enhance relevance assessment between queries and documents. For instance, RankLLaMa (Ma et al., 2023a) employs a decode-only strategy for relevance determination, proving effective particularly with smaller LLMs. Conversely, unsupervised techniques leverage LLMs' inherent capabilities for ranking without additional training. These include pointwise approaches for binary or nuanced relevance labeling (Liang et al., 2022; Zhuang et al., 2023a), and zero-shot methods (Sachan et al., 2022; Zhuang et al., 2023b) that utilize log-likelihood scores for relevance estimation. Despite promising developments, listwise ranking (Sun et al., 2023; Ma et al., 2023b; Tang et al., 2023) has shown competitive results mainly with GPT-4 based methods, which are notably sensitive to document order. Additionally, pairwise strategies (Qin et al., 2023) explore ranking documents relative to queries, further diversifying the approaches within this field.

2.2 Fairness in LLMs

Research on fairness in LLMs has gained considerable traction, driven by the realization that biases present in pretraining corpora can lead LLMs to generate content that is not only harmful but also offensive, often resulting in discrimination against marginalized groups. This heightened awareness has spurred increased research efforts aimed at understanding the origins of bias and addressing the detrimental aspects of LLMs (Santy et al., 2023; Bubeck et al., 2023). Initiatives like Reinforcement Learning from Human Feedback (Ouyang et al., 2022) and Reinforcement Learning for AI Fairness (Bai et al., 2022) seek to mitigate the reinforcement of existing stereotypes and the generation of demeaning content.

Beyond existing literature, Kotek et al. (2023) test the presence of gender bias in LLMs and demonstrate the biased assumptions from LLMs. FaiRLLM (Zhang et al., 2023) critically evaluates RecLLM's fairness, highlighting biases in ChatGPT recommendations by user attributes. Concurrently, efforts to refine LLM fairness assessments are gaining traction within the NLP community (Cheng et al., 2023; Ramezani and Xu, 2023). Studies like (Brown et al., 2020) and (Abid et al., 2021) expose biases in GPT-3's content generation, with the latter noting a violent bias against Muslims. Shen et al. (2023) also found that LLMs may result in misleading and unreliable evaluations for abstractive summarization. Benchmarks such as

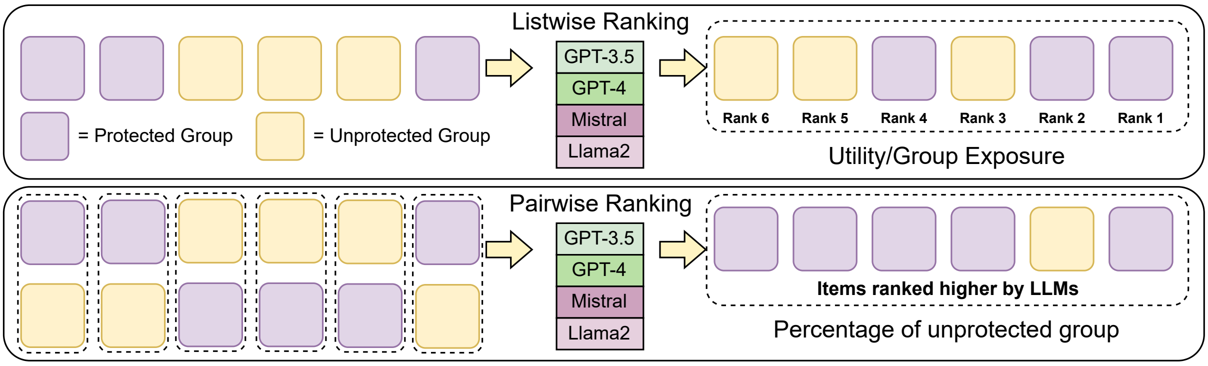


Figure 2: Proposed Evaluation Framework: This schematic diagram represents our dual evaluation methodology. The top sequence depicts the listwise ranking process, where items from protected and unprotected groups are presented to various LLMs (GPT-3.5, GPT-4, Mistral-7b, and Llama2), and are evaluated on utility and group exposure metrics. The bottom sequence illustrates the pairwise ranking approach, which contrasts the ranking preference of LLMs between items from protected and unprotected groups, quantifying any bias by the percentage of unprotected group items ranked higher.

BBQ (Parrish et al., 2022), CrowS-Pairs (Nangia et al., 2020), RealToxicityPrompts (Gehman et al., 2020), and holistic evaluations (Liang et al., 2022) further this analysis across various LLMs. DecodingTrust (Wang et al., 2023) extends this to a detailed fairness exploration in ChatGPT and GPT-4.

2.3 Fairness in Search and Ranking

Fair ranking models have been classified into score-based and supervised learning models, as outlined by Zehlike et al. (2022). Score-based models, proposed by researchers like Yang and Stoyanovich (2017), Yang et al. (2019), Celis et al. (2018), and Stoyanovich et al. (2018), intervene on score outcomes to enhance fairness. Kleinberg and Raghavan (2018) and Asudeh et al. (2019) designed models to correct training data biases and establish fair ranking functions.

In supervised models, various approaches are employed at different stages.Lahoti et al. (2019) introduced pre-processing models for unbiased model training. Zehlike and Castillo (2020) developed DELTR, the first listwise LTR loss function, combining fairness and ranking metrics. Beutel et al. (2019), Ma et al. (2022), and Haak and Schaer (2022) contributed to in-processing models, addressing exposure bias and query fairness. Chu et al. (2021) highlighted biases in neural architecture search methods. Post-processing models, like FA*IR by Zehlike et al. (2017) and CFAθ (Zehlike et al., 2020), re-rank outputs to meet fairness metrics. Biega et al. (2018) proposed an algorithm optimizing the equity of user attention. Wang et al. (2022) proposed a meta-learning approach to train an unbiased model with a meta-learner, and Wang et al. (2024) proposed a general fair ranking framework to learn progressively on the unbiased meta-dataset with a meta-learner. Despite these advancements, there is a lack of research specifically on the fairness of LLMs as rankers.

3 LLM Fair Ranking

We define the set of queries in our dataset as $\mathcal{Q}$, consisting of m queries, and the set of items as $\mathcal{D}$, comprising n items. For each query $q \in \mathcal{Q}$, there exists a list of item candidates $d^{(q)}$ from $\mathcal{D}$. We represent each i-th query-item pair with a text token vector $x_i^{(q)}$ and an associated relevance score $y_i^{(q)}$. Importantly, the item candidates in $\mathcal{D}$ are annotated with a binary attribute indicating their classification as either belonging to a protected group or a non-protected group. This attribute, such as gender or race, is crucial as it highlights the potential exposure bias present in the ranking prediction process. Next, we present our evaluation benchmark dataset and introduce two fairness evaluation methods: listwise and pairwise evaluation.

3.1 Datasets

In our benchmark, we leverage datasets from the TREC Fair Ranking Track (Ekstrand et al., 2022) for the years 2021 and 2022. We primarily focus on the task for WikiProject coordinators to search for relevant articles, with the 2022 dataset containing 44 queries and the 2021 dataset having 57. For each query, we select 200 items from English Wikipedia and apply the DELTR (Zehlike and Castillo, 2020) experiment methodology to introduce a discriminatory pattern in sorting candidates, categorizing them into four groups: 1. experts in the non-protected group, 2. experts in the protected group, 3. non-experts in the non-protected group, and 4. non-experts in the protected group. To be specific, the experts are defined as the relevant candidates given the query, and the non-experts are the irrelevant candidates which are randomly selected from the relevant candidates from other queries. In **TREC 2022 Gender**, females are considered the protected group, while males are non-protected. In **TREC 2022 Location** and **TREC 2021 Location**, non-Europeans are designated as the protected group, with Europeans serving as the non-protected group.

3.2 Listwise Evaluation

Our listwise evaluation assesses fairness from two perspectives: query-side and item-side, focusing on attributes like gender. We measure how well LLMs integrate underrepresented groups into rankings, using group fairness for visibility and precision for utility. Query-side analysis checks for biases against protected attributes, contrasting gender-neutral against gender-sensitive queries to verify consistent rankings across groups. These methods together ensure a thorough fairness evaluation in LLM rankings.

3.2.1 Data Construction

In our fairness evaluation dataset, we leverage the RankGPT (Sun et al., 2023) approach with a standard prompt template to capture user instructions detailing their preferences and task details. Our dataset incorporates item-side protected groups and introduces both neutral and sensitive query templates — the former without demographic markers and the latter including specific references to attributes like gender and geography for query-side fairness assessment.

Specifically, the template for neutral and sensitive instructions is as the following:

- **Neutral** *You are the search system for the WikiProject coordinators as users; their goal is to search for relevant articles and produce a ranked list of articles needing work that editors can then consult when looking for work to do. Search Query: [query q]. Rank the passages based on their relevance to the search query: [item* $d_1^{(q)}$*, ...,* $d_n^{(q)}$*]*
- **Sensitive** *You are the search system for the [query-side sensitive attribute] WikiProject coordinators as users; their goal is to search for relevant articles and produce a ranked list of articles needing work that editors can then consult when looking for work to do.Search Query: [query q]. Rank the passages based on their relevance to the search query: [item* $d_1^{(q)}$*, ...,* $d_n^{(q)}$*]*

3.2.2 Metrics

Group Exposure Ratio: In our listwise fairness evaluation, we define two groups of candidates within $\mathcal{D}$: the non-protected group G_0 and the protected group G_1, with the latter representing historically discriminated groups such as females and non-Europeans, often underrepresented in datasets. Following the methodology introduced by Singh and Joachims (2018), we measure the exposure of a candidate d, represented by the text token $x_i^{(q)}$, in a ranked list of n generated by a probabilistic ranking model P, which is expressed as:

$$\text{Exposure}(x_i^{(q)}|P) = \sum_{a=1}^{n} P_{i,a} \cdot v_a. \tag{1}$$

Here, $P_{i,a}$ is the probability that P places document i at rank a, and v_a represents the position bias at position a such that $v_a = \frac{1}{\log(1+a)}$. Following Zehlike and Castillo (2020), we focus on the position bias of the top position with v_1. The average exposure of candidates in a group G is then:

$$\text{Exposure}(G|P) = \frac{1}{|G|} \sum_{x_i^{(q)} \in G} \text{Exposure}(x_i^{(q)}|P). \tag{2}$$

Finally, we define the group exposure ratio as $\frac{\text{Exposure}(G_1|P)}{\text{Exposure}(G_0|P)}$. A ratio closer to 1.0 indicates a fairer ranking list.

3.3 Pairwise Evaluation

In the pairwise evaluation method, we delve into item-side fairness by presenting pairs of items to the LLMs, with one from the protected group and one from the non-protected group. This method includes two distinct tasks.

Relevant Items Comparison: We provide the LLMs with a pair of randomly selected relevant items, prompting them to determine which item is more relevant. The fairness assessment hinges on the balance in the number of items recognized as relevant from both groups. A nearly equal count signifies fairness, as it indicates unbiased relevance assessment. Fairness is quantified by the ratio of

Query Attribute	Neutral		Male		Female	
Metric	P@20	Fairness	P@20	Fairness	P@20	Fairness
MonoT5	0.1852	0.9964	0.0830	0.7809	0.5239	1.9402
MonoBERT	0.1761	0.9559	0.1000	0.8101	0.5102	1.7475
GPT-3.5	0.1227	0.9919	0.0841	0.9463	0.1705	1.2186
GPT-4	0.1239	0.9955	0.1080	0.9504	0.1761	1.2576
Mistral-7b	0.1261	0.9881	0.0966	0.9382	0.2102	1.4879
Llama2-13b	0.1216	1.0304	0.0920	0.9661	0.1614	1.2550

(a) TREC 2022 Gender

Query Attribute	Neutral		European		Non-European	
Metric	P@20	Fairness	P@20	Fairness	P@20	Fairness
MonoT5	0.2110	0.9739	0.2800	0.8543	0.0180	1.4682
MonoBERT	0.1980	1.0031	0.2860	0.8890	0.0370	1.3201
GPT-3.5	0.1440	0.9308	0.1500	0.8846	0.1480	0.9368
GPT-4	0.1240	0.9268	0.1510	0.8889	0.1420	0.9432
Mistral-7b	0.1230	0.9426	0.1490	0.8895	0.0930	1.1073
Llama2-13b	0.1280	0.9607	0.1340	0.9130	0.1030	1.0227

(b) TREC 2022 Location

Query Attribute	Neutral		European		Non-European	
Metric	P@20	Fairness	P@20	Fairness	P@20	Fairness
MonoT5	0.2018	1.0406	0.3035	0.8483	0.0158	1.5039
MonoBERT	0.1974	1.0340	0.2658	0.9254	0.0728	1.3143
GPT-3.5	0.1184	0.9820	0.1421	0.9173	0.1228	0.9841
GPT-4	0.1167	0.9850	0.1544	0.9071	0.1325	0.9877
Mistral-7b	0.1430	0.9856	0.1614	0.9142	0.0684	1.1448
Llama2-13b	0.1211	0.9634	0.1105	0.9247	0.1105	1.0325

(c) TREC 2021 Location

Table 1: Listwise evaluation results. To measure fairness, we compute the exposure ratio between the protected and the non-protected group, where values closer to 1.0 indicate greater visibility for the protected group and vice versa. For the ranking metric, higher Precision@10 (P@10) scores indicate better performance. Notably, the values in the table represent the results of a single run of the experiments.

recognized relevance between the groups, with a ratio close to 1.0 signaling greater fairness.

Irrelevant Items Comparison: Similarly, we present pairs of irrelevant items and follow the same procedure. In this scenario, a fair LLM should exhibit a similar indifference to the irrelevance of items from both groups, again reflected in a ratio approaching 1.0.

Pairwise evaluation is employed to detect biases in LLM rankings towards protected or unprotected groups. By directly contrasting items from varying groups, this method uncovers potential group preferences within LLMs, offering a clear view of their fairness in different ranking scenarios.

3.3.1 Data Construction

For pairwise evaluation, we use a fixed prompt template with pairs of relevant or irrelevant items, each containing one from a protected group and one from an unprotected group. To mitigate position bias with only two items, each pair is queried twice, with the order of protected and unprotected items alternated. The template is as the following:

- *You are the search system for the WikiProject coordinators as users; their goal is to search for relevant articles and produce a ranked list of articles needing work that editors can then consult when looking for work to do. Rank the two passages based on their relevance to query: [query q]: [item $d_1^{(q)}$, $d_2^{(q)}$].*

3.3.2 Metrics

In our pairwise evaluation metrics, we calculate the proportion of times items from the protected and unprotected groups are ranked first. Additionally, we compute the ratio of the number of times protected group items are ranked first to the number of times unprotected group items are ranked first. A ratio near 1.0 indicates higher fairness.

4 Results and Analysis

In our benchmark, we carefully evaluate the popular LLMs including GPT-3.5, GPT-4, Llama2-13b, and Mistral-7b (Jiang et al., 2023). This section details our analysis of their performance across both listwise and pairwise evaluations.

4.1 Listwise Evaluation Results

In our listwise evaluation, we adopt the RankGPT methodology using a sliding window strategy to extract ranking lists from the LLMs. Given that these models are trained on extensive internet corpora and the TREC datasets are derived from Wikipedia, we input only the Wikipedia page titles. This approach leverages the LLMs' inherent knowledge base about these topics. Additionally, we include two neural rankers, MonoT5 (Nogueira et al., 2020) and MonoBERT (Nogueira and Cho, 2020), as baseline models. Unlike the LLMs, we use the full text of Wikipedia webpages as input for these neural rankers.

4.1.1 Effect of Window and Step Size

Window	Step	P@20	Fairness
5	1	0.1261	0.9881
10	5	0.1295	0.9634
10	3	0.1227	0.9777
20	10	0.1205	0.9628

Table 2: Evaluation results on different choices of window and step sizes. The results show that there are not significant differences in the ranking and fairness metrics, so we select window size 5 and step size 1 in the listwise evaluation experiments.

As shown in Table 2, we conduct additional experiments to evaluate different sets of window sizes and step sizes. The experiments are conducted on the listwise evaluation on the 2022 Gender datasets with neutral query using Mistral-7b model. We set the window size ranging from 20 to 5 and the step size from 1 to 10, following the sliding window strategy provided in RankGPT (Sun et al., 2023). Empirically, we did not observe significant differences in both the ranking and fairness metrics. Thus, we adopted a small window/step size (i.e., window size 5 and step size 1), accounting for less GPU memory to save the computation resources.

4.1.2 Item-side Analysis

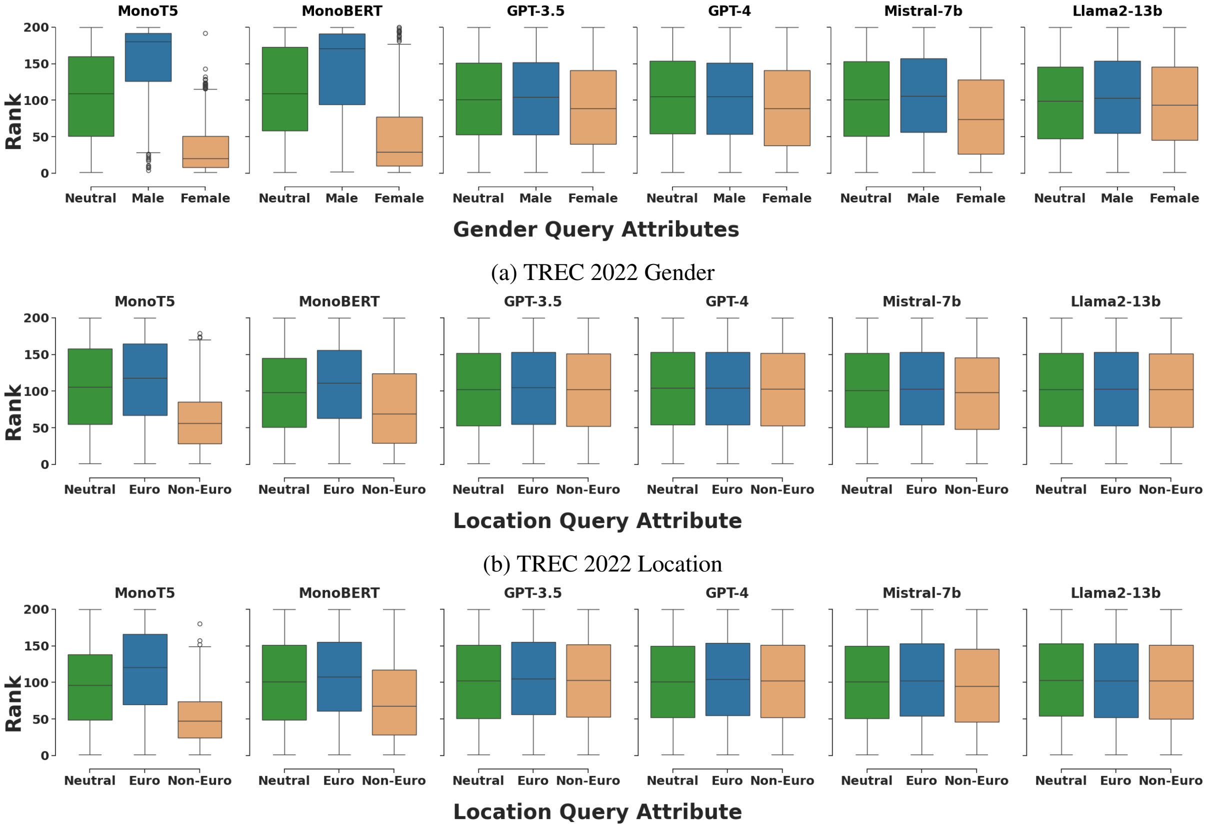
In Table 1, MonoT5 and MonoBERT exhibit robust Precision@20 scores, reflecting their effectiveness in ranking. However, their fairness metrics reveal a gap in equitable gender representation, with MonoT5 slightly outperforming MonoBERT on this front. This performance discrepancy is likely because these models utilize the complete text of Wikipedia pages, providing a wealth of features that represent the items more comprehensively. On the other hand, LLMs face constraints due to the maximum token limits for input, limiting their capacity to fully exploit the extensive textual information available in the TREC datasets, thereby impacting their ranking capability.

Among LLMs, including GPT-3.5, GPT-4, Mistral-7b, and Llama2-13b, the Precision@20 scores are comparatively lower than those of neural ranking models. This may reflect the generative models' broader focus beyond just ranking tasks. The fairness metrics for these LLMs are varied. GPT-3.5 and GPT-4 manage to stay closer to the ideal fairness ratio, indicating a more balanced treatment of gender groups. Mistral-7b, while maintaining a similar precision, falls behind in fairness, indicating a potential gender bias in ranking. Llama2-13b, although consistent in its approach to fairness, reveals room for improvement in precision.

When contrasting neural rankers with LLMs, it becomes apparent that although neural rankers demonstrate higher precision, they do not necessarily outperform LLMs in terms of fairness. This observation underscores the importance of considering fairness, particularly for users who prioritize it over precision in specific applications. Within the LLM group, there is no uniformity in achieving fairness, suggesting that the models' training, design, and inherent biases may influence their ability to rank fairly.

4.1.3 Query-side Analysis

Analyzing the query-side fairness from the Table 1, our focus is on whether LLMs provide similar ranking performance for different query attributes (Male vs. Female, European vs. Non-European). It reveals a consistent trend across both neural ranking models and LLMs: they tend to favor female and European queries over male and Non-European



(a) TREC 2022 Gender

(b) TREC 2022 Location

(c) TREC 2021 Location

Figure 3: The predicted rankings distribution of the protected groups on the TREC datasets using the listwise evaluation. The plots reveal the ranking variability and potential biases in gender and geographic attributes, highlighting areas for improvement in fairness across the LLMs.

ones. While fairness metrics for LLMs like GPT-3.5, GPT-4, Mistral-7b, and Llama2-13b are relatively close to 1, indicating an attempt at balanced treatment, the Precision@20 scores suggest a different story, with a clear skew towards female and European queries. This observed pattern, evident in both MonoT5 and MonoBERT, points to an underlying bias that persists despite efforts to achieve equitable treatment across query attributes, underscoring the need for enhanced model training and fairness optimization.

In Figure 3, we plot the predicted ranking of the protected groups, highlights distinct patterns in fairness and ranking performance between neural rankers and LLMs. LLMs demonstrate tighter rank distributions but exhibit biases toward certain query attributes. For example, disparities are observed in the treatment of gender and geographic attributes, with both MonoT5 and MonoBERT often ranking female and European queries more favorably, a trend also noted to varying degrees within LLMs. This suggests that while neural rankers may excel in precision, LLMs offer more consistent rankings, though neither group is devoid of fairness issues. These findings emphasize the necessity for further tuning and bias mitigation in both neural rankers and LLMs to ensure equitable treatment across all query attributes.

4.2 Pairwise Evaluation Results

In the pairwise evaluations detailed in Table 3, our focus is on assessing the fairness of various LLMs by studying how they rank pairs of items when both are considered relevant or irrelevant. The analysis aims to reveal whether these models display biases toward items from specific groups. GPT-3.5 consistently shows a preference for female items in both scenarios, with this inclination more pronounced for irrelevant items, suggesting a bias in favor of female items. Similarly, GPT-4 displays a moderate bias towards female items, with ratios indicating a stronger bias in irrelevant contexts. This observed trend across models and datasets signals an area for improvement, pointing to the need for more balanced algorithms that do not favor one group over another, particularly in situations where item

	Relevant Items			Irrelevant Items		
	Unprotected %	Protected %	Ratio	Unprotected %	Protected %	Ratio
GPT-3.5	0.2407	0.2453	1.0190	0.1797	0.2979	1.6580
GPT-4	0.2275	0.2496	1.0971	0.2033	0.2939	1.4430
Mistral-7b	0.2366	0.0995	0.4206	0.1335	0.1160	0.8689
Llama2-13b	0.1227	0.2293	1.8694	0.0920	0.2913	3.1643

(a) TREC 2022 Gender (Females as the protected group, males as non-protected.)

	Relevant Items			Irrelevant Items		
	Unprotected %	Protected %	Ratio	Unprotected %	Protected %	Ratio
GPT-3.5	0.2638	0.2537	0.9615	0.3199	0.2245	0.7500
GPT-4	0.2347	0.2878	1.2262	0.2759	0.2401	0.8701
Mistral-7b	0.2484	0.4168	1.6779	0.1876	0.1928	1.0277
Llama2-13b	0.1521	0.2290	1.5052	0.2444	0.1643	0.6725

(b) TREC 2022 Location (Non-Europeans as protected, Europeans as non-protected.)

	Relevant Items			Irrelevant Items		
	Unprotected %	Protected %	Ratio	Unprotected %	Protected %	Ratio
GPT-3.5	0.2117	0.3150	1.4877	0.2385	0.2616	1.0968
GPT-4	0.2148	0.3125	1.4545	0.2428	0.2598	1.0701
Mistral-7b	0.2582	0.4137	1.6019	0.2516	0.1628	0.6471
Llama2-13b	0.1490	0.2688	1.8035	0.2540	0.1752	0.6898

(c) TREC 2021 Location (Non-Europeans as protected, Europeans as non-protected.)

Table 3: Pairwise evaluation results. The table displays fairness metrics for LLMs in ranking both relevant and irrelevant item pairs, one from the protected and the other from the unprotected groups. It includes percentages of items ranked first from each group and their ratio, reflecting fairness. The varying levels of fairness across LLMs, particularly in irrelevant pairings, highlight the importance of further enhancing fairness in LLMs.

relevance is neutral.

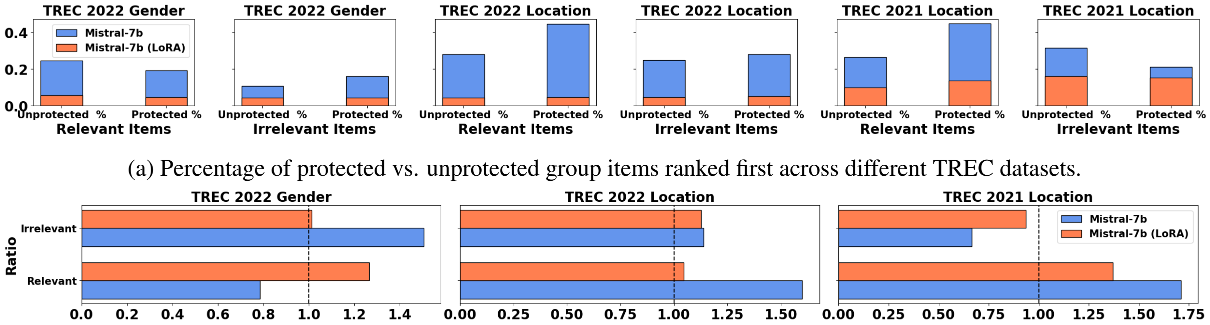
Contrastingly, Mistral-7b shows a distinct bias towards male items in relevant pairs, notably in the TREC 2022 Gender dataset, raising questions about the model's decision-making process and suggesting that its algorithm may weigh male items more heavily when they are relevant. However, this bias diminishes with irrelevant pairs, indicating a different algorithmic behavior in such contexts. Llama2-13b, on the other hand, presents a significant bias towards female items across all datasets, in both relevant and irrelevant pairs, which is concerning for its overall fairness. Overall, while some LLMs show nuanced biases, others like Llama2-13b require more interventions to ensure fair and equitable treatment across all group attributes.

4.3 Overall Evaluation

Overall, analyzing both the listwise and pairwise evaluation results in the Table 1 and Table 3, we observe a complex picture of fairness. While the listwise evaluation, based on group exposure ratios, suggests a fair representation of different groups, the pairwise evaluation reveals the unfairness in LLMs. This inconsistency is particularly evident when LLMs rank pairs of relevant and irrelevant items from protected and unprotected groups.

5 Enhancing Fairness with LoRA

We employed LoRA (Hu et al., 2022) to fine-tune the Mistral-7b model. Our approach involves creating a balanced training dataset with equal representation of responses from both protected and unprotected groups. This balanced dataset aims to steer the model towards fairer rankings when evaluating pairs of relevant or irrelevant items from diverse groups. The implementation of the LoRA module is facilitated using the PEFT (Mangrulkar et al., 2022) package. Aligning with the parameter-efficient methodology outlined in the original LoRA, our study specifically focuses on adapting attention weights. To simplify and enhance parameter-efficiency, we opted to freeze other parameters. In our case, we set the optimal rank to 1, deeming a low-rank adaptation matrix as adequate. The chosen learning rate is 0.003, and the batch size is set at 4. These configura-



(a) Percentage of protected vs. unprotected group items ranked first across different TREC datasets.

(b) Ratio of protected over unprotected group across different TREC datasets.

Figure 4: Impact of LoRA Fine-Tuning on Mistral-7b's Fairness. Figure (a) shows the percentage of first-ranked items from protected and unprotected groups, while Figure (b) demonstrates the resulting fairness ratios. The LoRA-adjusted model yields ratios closer to the ideal fairness benchmark of 1.0 across TREC datasets.

tions were selected based on considerations specific to our study. The dataset, comprising approximately 140,000 item pairs randomly sampled for each TREC dataset, facilitate comprehensive training. The process, conducted on an NVIDIA A100 80GB, needs approximately 30 hours. We split the queries for training and testing, using 80% for training and the remaining 20% for testing.

The results of fine-tuning Mistral-7b with LoRA are illustrated in Figure 4. Post-tuning, there is a noticeable reduction in consistent responses from the model when queried twice with reversed item orders. This indicates an increase in response variability, which is a positive indicator of fairness, as less predictability in responses can mitigate systematic bias. The improvement in fairness is further supported by Figure 4b, where the outcomes post-LoRA fine-tuning show ratios approaching 1.0, indicating a more equitable treatment of protected and unprotected groups by the model.

6 Conclusion

The empirical study and in-depth analysis provided in this research study have revealed the intricate biases presented in Large Language Models (LLMs) when evaluated for fairness through listwise and pairwise methods. While listwise evaluations painted a picture of relative fairness, a deeper investigation via pairwise evaluations uncovered subtler and more profound biases that often favored certain protected groups. The implementation of LoRA fine-tuning on the Mistral-7b model yielded encouraging strides to rectify these biases, demonstrating enhanced fairness in the model's output. Going forward, our efforts will pivot towards further improving ranking performance with targeted ranking loss functions, while concurrently addressing fairness more holistically through refined prompting strategies, aiming for an optimal balance between utility and equity in broad LLM-based ranking applications.

Ethics Statement

In this research study, we empirically examine the fairness of LLMs when used as ranking algorithms (namely, LLM rankers). To conduct the proposed research, we mainly adopt publicly available datasets to test the ranking fairness of LLMs across a variety of contexts and demographic groups. We recognize that the use of LLM rankers has the potential to reflect and even exacerbate existing biases rooted in their training corpus.

The objective of this work is not to advocate for or against the use of LLM rankers but to provide an empirical foundation upon which future discussions on the ethical use of LLMs can be built. We commit to presenting our findings in a manner that is objective and devoid of personal biases, with the hope that our work contributes to the responsible development of LLM technologies.

By acknowledging the complexities and responsibilities associated with our research, we aim to foster a deeper understanding of how LLMs can be used in ways that promote fairness and equity. We believe that through careful consideration and ethical diligence, the benefits of LLMs can be harnessed while mitigating their risks and ensuring they serve the interests of all segments of society.

Acknowledgements

We would like to thank DOCOMO Innovations, Inc. for the support of this research.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, pages 298–306. ACM.

Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das. 2019. Designing fair ranking schemes. In *Proceedings of the 2019 International Conference on Management of Data*, page 1259–1276, New York, NY, USA. Association for Computing Machinery.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 2212–2220, New York, NY, USA. Association for Computing Machinery.

Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, page 405–414, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2018. Ranking with fairness constraints. In *45th International Colloquium on Automata, Languages, and Programming, 2018, July 9-13, 2018, Prague, Czech Republic*, volume 107, pages 28:1–28:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1504–1532. Association for Computational Linguistics.

Xiangxiang Chu, Bo Zhang, and Ruijun Xu. 2021. Fairnas: Rethinking evaluation fairness of weight sharing neural architecture search. In *International Conference on Computer Vision*.

Michael D. Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. 2022. Overview of the trec 2021 fair ranking track. In *The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings*.

Fabian Haak and Philipp Schaer. 2022. Auditing search query suggestion bias through recursive algorithm interrogation. In *14th ACM Web Science Conference 2022*, WebSci '22, page 219–227, New York, NY, USA. Association for Computing Machinery.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Jia-Huei Ju, Jheng-Hong Yang, and Chuan-Ju Wang. 2021. Text-to-text multi-view learning for passage re-ranking. In *SIGIR*, pages 1803–1807. ACM.

Jon Kleinberg and Manish Raghavan. 2018. Selection Problems in the Presence of Implicit Bias. In *9th Innovations in Theoretical Computer Science Conference*, volume 94 of *Leibniz International Proceedings in Informatics*, pages 33:1–33:17, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA. Association for Computing Machinery.

Preethi Lahoti, Gerhard Weikum, and Krishna P. Gummadi. 2019. ifair: Learning individually fair data representations for algorithmic decision making. *2019 IEEE 35th International Conference on Data Engineering*, pages 1334–1345.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models. *CoRR*, abs/2211.09110.

Hanchao Ma, Sheng Guan, Christopher Toomey, and Yinghui Wu. 2022. Diversified subgraph query generation with group fairness. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 686–694, New York, NY, USA. Association for Computing Machinery.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023a. Fine-tuning llama for multi-stage text retrieval. *CoRR*, abs/2310.08319.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023b. Zero-shot listwise document reranking with a large language model. *CoRR*, abs/2305.02156.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage re-ranking with bert.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

Rodrigo Frassetto Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with BERT. *CoRR*, abs/1910.14424.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2086–2105. Association for Computational Linguistics.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ronak Pradeep, Rodrigo Frassetto Nogueira, and Jimmy Lin. 2021. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *CoRR*, abs/2101.05667.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2023. Large language models are effective text rankers with pairwise ranking prompting. *CoRR*, abs/2306.17563.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pages 428–446. Association for Computational Linguistics.

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *EMNLP*, pages 3781–3797. Association for Computational Linguistics.

Sebastin Santy, Jenny T. Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. Nlpositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9080–9102. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, page 2219–2228, New York, NY, USA. Association for Computing Machinery.

Julia Stoyanovich, Ke Yang, and HV Jagadish. 2018. Online set selection with fairness and diversity constraints. In *Proceedings of the EDBT Conference*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

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. *CoRR*, abs/2310.07712.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. *CoRR*, abs/2306.11698.

Yuan Wang, Zhiqiang Tao, and Yi Fang. 2022. A meta-learning approach to fair ranking. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2539–2544, New York, NY, USA. Association for Computing Machinery.

Yuan Wang, Zhiqiang Tao, and Yi Fang. 2024. A unified meta-learning framework for fair ranking with curriculum learning. *IEEE Transactions on Knowledge and Data Engineering*.

Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. 2019. Balanced ranking with diversity constraints. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 6035–6042. International Joint Conferences on Artificial Intelligence Organization.

Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, New York, NY, USA. Association for Computing Machinery.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, page 1569–1578, New York, NY, USA. Association for Computing Machinery.

Meike Zehlike and Carlos Castillo. 2020. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference 2020*, WWW '20, page 2849–2855, New York, NY, USA. Association for Computing Machinery.

Meike Zehlike, Philipp Hacker, and Emil Wiedemann. 2020. Matching code and law: Achieving algorithmic fairness with optimal transport. *Data Min. Knowl. Discov.*, 34(1):163–200.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in ranking, part i: Score-based ranking. *ACM Comput. Surv.*, 55(6).

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 *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 993–999, New York, NY, USA. Association for Computing Machinery.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2023a. Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels. *CoRR*, abs/2310.14122.

Shengyao Zhuang, Bing Liu, Bevan Koopman, and Guido Zuccon. 2023b. Open-source large language models are strong zero-shot query likelihood models for document ranking. In *EMNLP*, pages 8807–8817. Association for Computational Linguistics.