Measuring Group Fairness Differences in RAG Pipelines: A Pilot Study

Sijie Tao[1,*], Tetsuya Sakai[1,*]

[1]Waseda University, Tokyo, Japan

Abstract

In recent years, Retrieval-Augmented Generation (RAG) has attracted significant attention in both Natural Language Processing (NLP) and Information Retrieval (IR) communities. Although several studies have examined RAG evaluation from perspectives such as fairness, the shift in group fairness from retrieval to generation remains unexplored. Therefore, in this study, we propose an evaluation framework to measure the performance gap of the retriever and the generator in a RAG pipeline in terms of group fairness. Our method is straightforward and can be easily applied to standard RAG pipelines and quantifies how group fairness changes across stages. We built a simple pseudo-RAG scenario to demonstrate how our method works in practice, and our experiment showcased that our proposed framework yields informative results and complements existing evaluation metrics.

Keywords

Retrieval–Augmented Generation, Group Fairness, Evaluation, Information Retrieval

1. Introduction

Retrieval-Augmented Generation (RAG) [1] has recently become increasingly popular. A growing number of research has been conducted to investigate how RAG can be evaluated [2, 3, 4, 5]. Recently, some studies have examined fairness issues in RAG systems, emphasising that fairness is an important consideration [6, 7]. Moreover, Avula et al. proposed an evaluation framework to measure the fairness gap between the retrieval and generation stages in RAG Systems [8]. Understanding such gap is important for developers and researchers, as it enables them to gain a detailed view of the behaviour of each component in the RAG pipeline, and thus they can tune and improve the system more efficiently. However, the methodology of Avula et al. is an extension of existing group fairness metric, and therefore inherits its limitations. Additionally, to the best of our knowledge, except for Avula et al., little work has addressed the group fairness gap across stages of a RAG system.

In this study, we propose a new evaluation framework to measure the group fairness gap between the retrieval and generation stages in a RAG pipeline. Our framework is built based on Group Fairness and Relevance (GFR), a versatile framework for evaluating ranked lists in terms of group fairness and relevance [9], which overcomes limitation of existing group fairness metrics. We also demonstrate how our framework measure the group fairness gap in practice using data of the NTCIR-17 FairWeb-1 task [10] in a simple pseudo-RAG scenario. Our experiments reveal that our framework yield informative results. It provides an intuitive understanding of how group fairness shifts between different stages of the RAG pipeline. These insights can help researchers and developers better understand the behaviour of each component and design more specialised improvements.

The main contributions of this work are summarised as follows:

1. We propose a novel evaluation framework to measure the group fairness gap between the retrieval and generation stages in a RAG pipeline.
2. We utilised the data of the NTCIR-17 FairWeb-1 task, and built a RAG scenario to demonstrate how our evaluation metrics work.

This paper was reviewed and accepted by the program committee for BREV-RAG 2025: Beyond Relevance-based EValuation of RAG Systems, a SIGIR-AP 2025 workshop, held on December 10, 2025 in Xi'an, China.

*Corresponding authors.

tsjmailbox@ruri.waseda.jp (S. Tao); tetsuyasakai@acm.org (T. Sakai)

0000-0002-6751-5303 (S. Tao); 0000-0002-6720-963X (T. Sakai)

3. We quantitated and visualized the group fairness performance shift in a RAG pipeline.

2. Related Work

2.1. Measuring Fairness in RAG

Kim and Diaz [6] investigated the impact of fair rankings on the ranking and generation quality in RAG systems. Their work focuses on item-side fairness and the experiments evaluated twelve RAG models across seven tasks. The authors used expected exposure disparity (EE-D) [11, 12] to measure the fairness of rankings. Their experimental results reveal that incorporating fairness-aware retrieval does not necessarily harm performance, in many cases, it even improves both ranking and generation quality. The authors further demonstrated that fair retrieval leads to more balanced source attribution, ensuring that the generator fairly cites the materials it relies on. Their work challenges the "common sense" of a trade-off between fairness and relevance, and highlights the importance of fair attribution in RAG systems.

Wu et al. [7] explored group fairness issues in RAG systems. They proposed a group fairness evaluation framework for RAG, and examined group fairness issues in several RAG methods with their constructed benchmarks based on the data from the TREC 2022 Fair Ranking Track [13]. The authors used Group Disparity (GD) [14] and Equalized Odds (EO) [15] to evaluate group fairness. However, these metrics were not specifically designed for IR systems. Their experiments revealed fairness concerns across different stages through the RAG pipeline, but they did not focus on the fairness gap between these stages.

Avula et al. [8] proposed an evaluation framework to measure the group fairness gap between retrieval and generation in RAG systems. The authors built their fairness metric based on expected exposure [11] and Attention-Weighted Rank Fairness (AWRF) [16, 12] to measure the group fairness of the retrieval stage. For the generation stage, they adopted another expected exposure based method, which measures fairness by quantifying each retrieved passage's contribution to the generated text. Their experiments exhibited that fairness/unfairness can propagate from retrieval to generation in RAG pipelines. Moreover, the authors also found that as the context window size increases, the generator introduces its own fairness patterns: larger models tend to generate fairer results. Overall, utilising the proposed evaluation framework, this work highlights the interesting interaction between the retrieval and generation stages in RAG systems. However, since the evaluation metric in their work is built based on AWRF, it also inherits the limitations of the original metric. As discussed in detail by Sakai et al. [9], AWRF does not distinguish nominal and ordinal groups when handling intersectional group fairness. Specifically, AWRF computes the Cartesian product of the attribute sets, which undermines the nature of ordinal groups.

2.2. Group Fairness Evaluation in IR

Attention-Weighted Rank Fairness (AWRF) is one of the commonly used group fairness evaluation metrics in Information Retrieval. AWRF was first introduced by Sapiezynski et al. [16], and then was named by Raj and Ekstrand [12]. Given a ranked list, AWRF first computes the cumulated exposure (i.e. the achieved distribution) using an nDCG-like attention weighting. It then measures the group fairness by the similarity between the achieved distribution and the target distribution. The similarity is defined as one minus the Jenson-Shannon divergence (JSD) [17] of the two distributions. Relevance is incorporated via nDCG * AWRF. TREC Fair Ranking Tracks applied AWRF for group fairness evaluation [18, 13]. AWRF has also been employed in other prior work [19, 20], while Abolghasemi et al. further extended it for specialized group fairness evaluation task [21].

Sakai et al. proposed Group Fairness and Relevance (GFR), a framework for evaluating ranked lists in terms of group fairness and relevance [9]. GFR is designed to be user-friendly and flexible. The framework embodies relevance and group fairness by simply taking sum of them. Therefore, users are allowed to select their preferred measures for relevance and group fairness to build their own evaluation

formulas. Sakai et al. demonstrated how GFR works in practice using data from previous shared tasks as well as real-world datasets. GFR was applied for the evaluation of the NTCIR FairWeb tasks [10, 22]. Moreover, Tao and Sakai examined the difference between GFR and AWRF in terms of rank correlation, discriminative power, and robustness to system bias [23]. Based on GFR, Sakai [24] and Sakai et al. [25] further proposed group fairness evaluation metrics for conversational search.

Other representative group fairness metrics in IR evaluation include Skew [26], Normalised Discounted KL divergence (NDKL) [27, 26], Attention Bias Ratio (ABR) [28], and Expected Cumulative Exposure (ECE) [16]. Sakai et al. [9] provide a detailed discussion of their limitations and differences.

3. Proposed Evaluation Framework

This section describes our proposed evaluation framework to measure the difference between the retrieval stage and the generation stage in terms of group fairness.

3.1. Group Fairness of the Retrieval Stage

Let the output of the retrieval stage (Re) be a ranked list L, following the GFR strategy, the group fairness of the retrieval stage can be computed by the following equation [9]:

$$\mathrm{GF}_{\mathrm{Re}}(L, m) = \sum_{k=1}^{L} \mathrm{Decay}(L, k)\mathrm{DistrSim}(L, k, m)\ , \tag{1}$$

where k is the rank, m stands for the attribute set (e.g. a set containing gender groups, country groups, etc.) to be considered.

The Decay function is based on Expected Reciprocal Rank (ERR) [29]. That is, for example, if the maximum relevance grade is set to two, the satisfaction probability at rank k ($p^{\mathrm{sat}}_{L,k}$) will be $3/4$ for L2 (highly relevant) documents, $1/4$ for L1 (relevant) documents, and 0 for L0 (non-relevant) documents. The Decay function is given by the following equation:

$$\mathrm{Decay}(L, k) = \begin{cases} p^{\mathrm{sat}}_{L,1}, & k = 1, \\ p^{\mathrm{sat}}_{L,k} \prod_{j=1}^{k-1} (1 - p^{\mathrm{sat}}_{L,j}), & k > 1. \end{cases} \tag{2}$$

The DistrSim function computes the similarity between the achieved distribution $d_{L,k}$ and the target distribution d^* at rank k:

$$\mathrm{DistrSim}(L, k, m) = \mathrm{DistrSim}(d_{L,k} \parallel d^*) = \begin{cases} 1 - \mathrm{JSD}(d_{L,k} \parallel d^*), & m \text{ is nominal}, \\ 1 - \mathrm{RNOD}(d_{L,k} \parallel d^*), & m \text{ is ordinal (option 1)}, \\ 1 - \mathrm{NMD}(d_{L,k} \parallel d^*), & m \text{ is ordinal (option 2)}, \end{cases} \tag{3}$$

where JSD is the Jensen-Shannon divergence [17], NMD is the Normalised Match Distance [30], and RNOD is the Root Normalised Order-aware Divergence [31][32]. The achieved distribution $d_{L,k}$ is computed by deriving page group membership [9]. The computation method of target distribution d^* is not explicitly defined, but as described in the next section, we used the data from the NTCIR-17 FairWeb-1 task, in which target distributions are already predefined.

JSD is used for attribute sets containing nominal groups (e.g. gender or country groups). NMD and RNOD are for attribute sets containing ordinal groups (e.g. researchers' age or h-index groups). It should be noted that JSD is not suitable for ordinal groups. The details are further discussed by Sakai et al. [9].

When handling intersectional group fairness (i.e. there are more than one attribute set to consider), the overall group fairness of M attribute sets can be simply computed by the following equation:

$$\mathrm{GF}_{\mathrm{Re}}(L, M) = \sum_{k=1}^{L} \mathrm{Decay}(L, k) \sum_{m=1}^{M} w_m \mathrm{DistrSim}(L, k, m) \ , \tag{4}$$

where w_m is the weight for attribute set m.

3.2. Group Fairness of the Generation Stage

To measure group fairness of the generation stage (Gen) of a RAG pipeline, we treat the final answer A of Gen as a "ranked list" of length one (i.e. a single document). Therefore, group fairness of the generation stage can be simply computed by the following equation:

$$\mathrm{GF}_{\mathrm{Gen}}(A, m) = \mathrm{DistrSim}(A, m) = \mathrm{DistrSim}(d_A \parallel d^*) \ , \tag{5}$$

where d_A is the achieved distribution of the final answer A.

When handling multiple attribute sets, the overall group fairness of the generation stage is defined as follows:

$$\mathrm{GF}_{\mathrm{Gen}}(A, M) = \sum_{m=1}^{M} w_m \mathrm{GF}_{\mathrm{Gen}}(A, m), \tag{6}$$

3.3. Measuring Group Fairness Differences Between Stages

To quantify group fairness differences between the retrieval and generation stages, we propose Relative Improvement in Group Fairness (RIGF). Given $\mathrm{GF}_{\mathrm{Re}}$ and $\mathrm{GF}_{\mathrm{Gen}}$, RIGF can be simply computed by the following equation:

$$\mathrm{RIGF} = \frac{\Delta_{\mathrm{GF}}}{\mathrm{GF}_{\mathrm{Re}}} \times 100\% \ , \tag{7}$$

where $\Delta_{\mathrm{GF}} = \mathrm{GF}_{\mathrm{Gen}} - \mathrm{GF}_{\mathrm{Re}}$.

RIGF quantifies the relative change in group fairness at the generation stage: if RIGF=100%, it means that the group fairness score has doubled with respect to the retrieval stage. RIGF can also indicates whether the achieved distribution of the generation stage d_A is closer to the target distribution.

Additionally, we propose another metric named Achieved Distribution Similarity (ADSim) to measure how similar the achieved distributions are. ADsim is defined as follows:

$$\mathrm{ADsim} = \mathrm{DistrSim}(d_L \parallel d_A) \ , \tag{8}$$

where d_L is the achieved distribution of the ranked list L in the retrieval stage, and d_A is the achieved distribution in the generation stage (as defined earlier in Eq.(5)).

It should be noted that GFR does not compute an achieved distribution for the whole ranked list L. Therefore, we follow the strategy of AWRF to compute d_L for ASsim.

Given an attribute set m, we first compute a vector d'_L, which is the cumulated exposure that the ranked list L gives to each group in m:

$$d'_L = \sum_{k} aw@k * a_{L_k} \ , \tag{9}$$

where L_k is the document at rank k, and a_{L_k} is the group alignment vector of L_k (a group alignment vector of AWRF and page group membership of GFR are equivalent in effect). aw is the attention weight calculated by log discounting (i.e. an nDCG-like decay):

$$aw@k = \frac{1}{\log_2(k+1)} \tag{10}$$

The calculation of aw in this study is slightly different from that at TREC Fair Ranking Tracks. The reason is discussed in detail by Sakai et al. [9]. After obtaining d'_L, the achieved distribution d_L is simply computed by normalising d'_L:

$$d_L = \frac{d'_L}{|d'_L|} \tag{11}$$

When handling intersectional group fairness, we recommend reporting both the overall score and the per-attribute-set scores for RIGF, whereas for ADsim, only per-attribute-set scores should be reported.

4. Experiments and Results

4.1. Experimental Settings

We build a pseudo-RAG scenario to demonstrate how our evaluation framework functions in practice. We refer to it as a pseudo-RAG for two reasons. First, we use existing runs from the NTCIR-17 FairWeb-1 task as the output of the retrieval stage. Thus, the pipeline does not need a retriever component, but only includes some pre-processing of the retrieved documents together with the generator component. Another reason is that since the NTCIR-17 FairWeb-1 is a web search task where queries target entities such as researchers and movies, the retrieved documents may contain relevant entities, but they do not always provide enough evidence or context for the Large Language Model (LLM) to determine whether those entities satisfy the queries. Therefore, the LLM in this pipeline must combine the retrieved documents with its own knowledge to generate the final answer.

Next, we describe our pseudo-RAG pipeline in detail. As mentioned earlier, we use existing runs from the NTCIR-17 FairWeb-1 task as the result of the retrieval stage [10]. The runs and the official evaluation results are publicly available online. [1] We select two runs for our experiments: THUIR-QD-RR-3 from team THUIR [33] and RSLFW-Q-RR-4 from team RSLFW [34]. According to the official evaluation results of FairWeb-1, the THUIR run represents a top-performing system, while the RSLFW run is a mid-level system [10]. Each run contains results of three topic types: researchers, movies, and YouTube contents. We decide to conduct our experiments on the movie topics. There are two attribute sets for movie topics: ORIGIN and RATINGS. ORIGIN contains eight nominal groups derived from the "country of origin" field on a movie's IMDb page, whereas RATINGS consists of four ordinal groups based on the number of ratings. The IMDb metadata of the movies that appear in the generated answers is required to compute the achieved distributions. Such information can be obtained via an unofficial API named OMDb. [2] This enables us to automatically evaluate the generated answers without additional human annotation, which is not possible for researcher and YouTube contents topics. Therefore, in this study our experiments focus on the movie topics.

The retrieved documents are all html documents from a corpus named Chuweb21D (not ClueWeb!) [35]. Before passing the retrieval results to the LLM for answer generation, we apply pre-processing to the documents. First, we use Python packages such as Trafilatura [36] [3] and Beautiful Soup [4] to extract plain text from the html documents. Next, the cleaned texts are divided into paragraphs by sentence-ending punctuation or newline characters. Paragraphs that are too short are dropped at this step. We then apply Maximal Marginal Relevance (MMR) [37] to the paragraphs, and select two representative paragraphs from each document. As the final pre-processing procedure, we employ MapReduce [38]. The selected representative paragraphs are first summarized by the LLM, and the summaries are then concatenated and finally fed back to the LLM for final answer generation.

For the LLM in our pseudo-RAG pipeline, we choose Meta-Llama-3.1-8B-Instruct. [5] All experiments were conducted on an Apple MacBook Pro with an M4 Max chipset.

[1] https://waseda.app.box.com/v/ntcir17fairweb1officialpublic
[2] https://www.omdbapi.com/
[3] https://trafilatura.readthedocs.io/en/latest/
[4] https://www.crummy.com/software/BeautifulSoup/
[5] https://huggingface.co/meta-llama/Llama-3.1-8B

4.2. Results and Analysis

Table 1 and 2 present the RIGF and ADsim results of our experiments. For ORIGIN, which consists of nominal groups, JSD is used to compute similarities between distributions, while RNOD is applied for RATINGS, which contains ordinal groups. The $\mathrm{GF}_{\mathrm{Re}}$ values are directly taken from the official evaluation results of the NTCIR-17 FairWeb-1 task [10]. The results show that despite having better retrieval performance, the top-performing THUIR run shows a slight drop in group fairness from the retrieval stage to the generation stage. In contrast, Table 2 indicates that the RSLFW run achieves an improvement in group fairness at the generation stage, suggesting that its achieved distribution tends to be closer to the target distribution than at the retrieval stage. Regarding ADsim, the mean values for the THUIR run are 0.7946 and 0.7520 for the two attribute sets, whereas the corresponding values for the RSLFW run are 0.8799 and 0.7664. These results suggest that the RSLFW run tends to have achieved distributions at the generation stage that are more similar to those at the retrieval stage than the THUIR run.

Table 3 reports the overall RIGF results, where the weights of the attribute sets are equally set to 0.5. A similar observation can be made from Table 3: the THUIR run shows a drop in terms of group fairness at the generation stage, while the RSLFW run achieves an improvement.

Note that the "N/A" cells occur when the value of $\mathrm{GF}_{\mathrm{Re}}$ is 0.0000, meaning that no relevant entity was found in the run. In a standard and fully-implemented RAG pipeline, the $\mathrm{GF}_{\mathrm{Gen}}$ values should also be 0.0000. However, the corresponding $\mathrm{GF}_{\mathrm{Gen}}$ values in the tables are not 0.0000 because the evaluation of this study is conducted automatically without human annotation. That is, when we evaluate the generated answers, as long as the movie is found in the OMDb API, it is treated as a relevant and correctly cited entity.

Figure 1 and 2 visualise the RIGF results. Both runs exhibit a drop in group fairness across most topics. Although the mean RIGF of the RSLFW run indicates an overall improvement at the generation stage, Figure 2 shows that this improvement is driven by substantial gains on several individual topics, while roughly half of the topics still show a drop at the generation stage.

Additionally, it should also be emphasized that we are not claiming that RSLFW performs better than THUIR. As mentioned before, this is a pseudo-RAG pipeline, and the evaluation is approximate and conducted without human annotation. Nevertheless, the proposed evaluation metrics clearly provide informative insights into the shift of group fairness between the retrieval and generation stages. Moreover, the computation of our metrics is straightforward, requiring only a few additional steps beyond existing group fairness evaluation metrics.

5. Conclusions and Future Work

In this study, we proposed an evaluation framework for measuring the performance shift between the retrieval and generation stages in a RAG pipeline in terms of group fairness. We proposed two metrics in our framework: RIGF for quantifying relative changes in group fairness from the retrieval stage to the generation stage, and ADsim for measuring the similarity between the achieved distributions of the two stages. We utilised data from the NTCIR-17 FairWeb-1 task, and built a pseudo-RAG scenario, to demonstrate how our evaluation framework functions in practice. The evaluation results indicate that our evaluation framework provides informative insights into group fairness across different stages of the RAG pipeline, topics, and systems.

For future work, we plan to extend our experiment beyond the pseudo-RAG setting by applying our evaluation framework to fully implemented RAG pipelines and involve human annotations to ensure a more precise evaluation. In addition, expanding the experiments to more systems and larger datasets will help validate the robustness of the proposed evaluation metrics. Finally, future research could further investigate extensions or modifications of the framework, as well as the potential trade-offs between item-based fairness, group fairness, retrieval performance, and generation quality.

Table 1
RIGF and ADsim results of THUIR-QD-RR-3.

ORIGIN					RATINGS				
qid	GF_{Re}	GF_{Gen}	RIGF	ADsim	qid	GF_{Re}	GF_{Gen}	RIGF	ADsim
M001	0.7887	0.6333	-19.71	0.9318	M001	0.7177	0.4796	-33.18	1.0000
M002	0.5444	0.0000	-100.00	N/A	M002	0.6629	0.0000	-100.00	N/A
M003	0.7847	0.5187	-33.89	0.9268	M003	0.7784	0.4796	-38.39	0.5875
M004	0.2274	0.6310	177.48	0.6295	M004	0.2314	0.6181	167.12	0.8557
M005	0.6513	0.4859	-25.40	0.8243	M005	0.6063	0.6497	7.16	0.9938
M006	0.4578	0.5123	11.92	0.6640	M006	0.6198	0.6773	9.27	0.7988
M007	0.6301	0.6196	-1.67	0.8597	M007	0.6639	0.4796	-27.76	0.8689
M008	0.6283	0.3970	-36.82	0.6736	M008	0.6538	0.4796	-26.65	0.7172
M009	0.6806	0.5983	-12.09	0.9522	M009	0.6667	0.6632	-0.52	0.7548
M010	0.7269	0.5768	-20.65	0.8374	M010	0.7309	0.6773	-7.34	0.5484
M011	0.5421	0.6170	13.82	0.7132	M011	0.5239	0.5472	4.46	0.9415
M012	0.2987	0.6452	116.01	0.8517	M012	0.5468	0.6230	13.93	0.4470
M013	0.7218	0.2617	-63.74	0.7229	M013	0.7029	0.4796	-31.77	0.8401
M014	0.4030	0.6056	50.27	0.7430	M014	0.4191	0.4796	14.43	0.4227
M015	0.0000	0.6781	N/A	N/A	M015	0.0000	0.6320	N/A	N/A
mean	0.5391	0.5187	3.97	0.7946	mean	0.5683	0.5310	-3.52	0.7520

Table 2
RIGF and ADsim results of RSLFW-Q-RR-4

ORIGIN					RATINGS				
qid	GF_{Re}	GF_{Gen}	RIGF	ADsim	qid	GF_{Re}	GF_{Gen}	RIGF	ADsim
M001	0.5291	0.6806	28.63	0.8089	M001	0.5288	0.4796	-9.31	1.0000
M002	0.2299	0.5879	155.73	0.7810	M002	0.2365	0.6773	186.36	0.6536
M003	0.8268	0.5145	-37.77	0.9379	M003	0.7868	0.4796	-39.05	0.6547
M004	0.0000	0.4859	N/A	N/A	M004	0.0000	0.6773	N/A	N/A
M005	0.7098	0.6436	-9.32	0.9044	M005	0.6961	0.5948	-14.55	0.7796
M006	0.4061	0.5769	42.07	0.9067	M006	0.4233	0.4796	13.30	0.7009
M007	0.7428	0.2617	-64.77	0.7842	M007	0.6139	0.4796	-21.88	0.9179
M008	0.6870	0.5417	-21.16	0.9927	M008	0.6884	0.5948	-13.59	0.6630
M009	0.2260	0.2617	15.81	1.0000	M009	0.2283	0.4796	110.07	1.0000
M010	0.5622	0.4631	-17.62	0.8503	M010	0.5570	0.6497	16.65	0.7668
M011	0.5406	0.4859	-10.12	0.8888	M011	0.6133	0.6773	10.43	0.9277
M012	0.8202	0.5768	-29.68	0.8458	M012	0.8447	0.4796	-43.22	0.6475
M013	0.6353	0.2617	-58.80	0.8808	M013	0.6305	0.5948	-5.66	0.6843
M014	0.2355	0.6268	166.16	0.8568	M014	0.2901	0.4796	65.32	0.5670
M015	0.0000	0.6764	N/A	N/A	M015	0.0000	0.6320	N/A	N/A
mean	0.4768	0.5097	12.24	0.8799	mean	0.4758	0.5637	19.60	0.7664

Declaration on Generative AI

The authors used no generative AI tools for content creation beyond grammar or spelling checks.

Table 3
Overall RIGF results ($w_{\mathrm{ORIGIN}} = w_{\mathrm{RATINGS}} = 0.5$)

THUIR-QD-RR-3				RSLFW-Q-RR-4			
qid	$\mathrm{GF}_{\mathrm{Re}}$	$\mathrm{GF}_{\mathrm{Gen}}$	RIGF	qid	$\mathrm{GF}_{\mathrm{Re}}$	$\mathrm{GF}_{\mathrm{Gen}}$	RIGF
M001	0.7532	0.5564	-26.12	M001	0.5290	0.5801	9.67
M002	0.6037	0.0000	-100.00	M002	0.2332	0.6326	171.27
M003	0.7816	0.4992	-36.13	M003	0.8068	0.4971	-38.39
M004	0.2294	0.6245	172.25	M004	0.0000	0.5816	N/A
M005	0.6288	0.5678	-9.70	M005	0.7030	0.6192	-11.91
M006	0.5388	0.5948	10.39	M006	0.4147	0.5283	27.38
M007	0.6470	0.5496	-15.06	M007	0.6784	0.3707	-45.36
M008	0.6411	0.4383	-31.63	M008	0.6877	0.5682	-17.37
M009	0.6737	0.6308	-6.37	M009	0.2272	0.3707	63.17
M010	0.7289	0.6270	-13.98	M010	0.5596	0.5564	-0.57
M011	0.5330	0.5821	9.22	M011	0.5770	0.5816	0.80
M012	0.4228	0.6341	49.99	M012	0.8325	0.5282	-36.55
M013	0.7124	0.3707	-47.97	M013	0.6329	0.4283	-32.33
M014	0.4111	0.5426	32.00	M014	0.2628	0.5532	110.50
M015	0.0000	0.6550	N/A	M015	0.0000	0.6542	N/A
mean	0.5537	0.5249	-0.94	mean	0.4763	0.5367	15.41

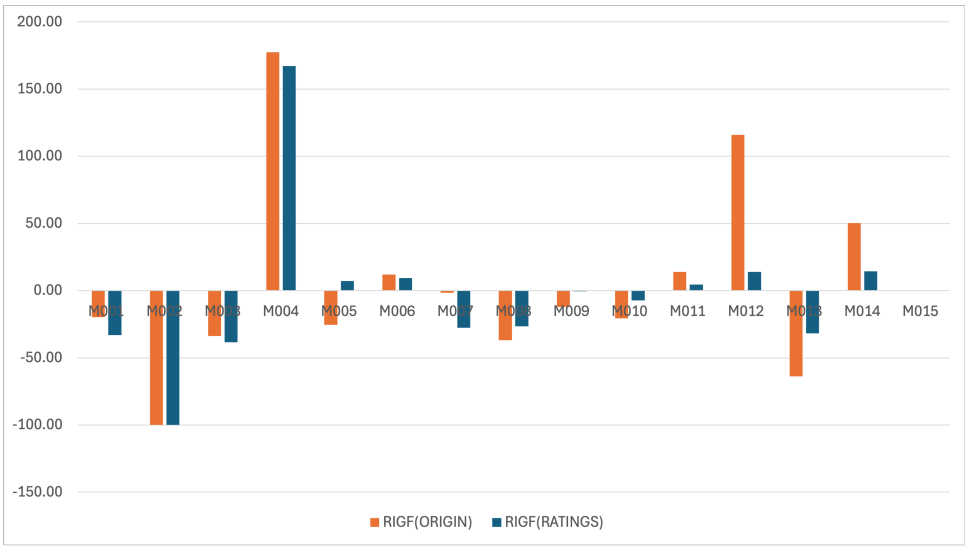

Figure 1: Visualisation of RIGF results (THUIR-QD-RR-3)

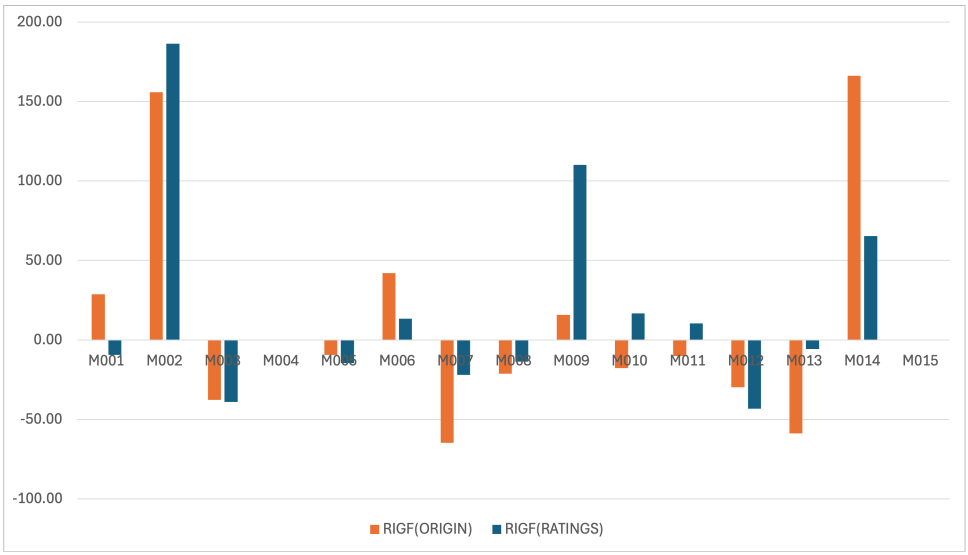

Figure 2: Visualisation of RIGF results (RSLFW-Q-RR-4)

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[2] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, Z. Liu, Evaluation of Retrieval-Augmented Generation: A Survey, Springer Nature Singapore, 2025, p. 102–120. URL: http://dx.doi.org/10.1007/978-981-96-1024-2_8. doi:10.1007/978-981-96-1024-2_8.

[3] J. Saad-Falcon, O. Khattab, C. Potts, M. Zaharia, ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 338–354. URL: https://aclanthology.org/2024.naacl-long.20/. doi:10.18653/v1/2024.naacl-long.20.

[4] S. Es, J. James, L. Espinosa Anke, S. Schockaert, RAGAs: Automated Evaluation of Retrieval Augmented Generation, in: N. Aletras, O. De Clercq (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 150–158. URL: https://aclanthology.org/2024.eacl-demo.16/. doi:10.18653/v1/2024.eacl-demo.16.

[5] J. Chen, H. Lin, X. Han, L. Sun, Benchmarking Large Language Models in Retrieval-Augmented Generation, in: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'24/IAAI'24/EAAI'24, AAAI Press, 2024. URL: https://doi.org/10.1609/aaai.v38i16.29728. doi:10.1609/aaai.v38i16.29728.

[6] T. E. Kim, F. Diaz, Towards Fair RAG: On the Impact of Fair Ranking in Retrieval-Augmented Generation, in: Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR), ICTIR '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 33–43. URL: https://doi.org/10.1145/3731120.3744599. doi:10.1145/3731120.3744599.

[7] X. Wu, S. Li, H.-T. Wu, Z. Tao, Y. Fang, Does RAG Introduce Unfairness in LLMs? Evaluating Fairness in Retrieval-Augmented Generation Systems, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 10021–10036. URL: https://aclanthology.org/2025.coling-main.669/.

[8] S. Avula, C.-J. Lee, R. Zhang, V. Murdock, Measuring the Fairness Gap Between Retrieval and Generation in RAG Systems using a Cognitive Complexity Framework, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 2994–2998. URL: https://doi.org/10.1145/3726302.3730230. doi:10.1145/3726302.3730230.

[9] T. Sakai, J. Y. Kim, I. Kang, A Versatile Framework for Evaluating Ranked Lists in Terms of Group Fairness and Relevance, ACM Trans. Inf. Syst. 42 (2023). URL: https://doi.org/10.1145/3589763. doi:10.1145/3589763.

[10] S. Tao, N. Chen, T. Sakai, Z. Chu, H. Arai, I. Soboroff, N. Ferro, M. Maistro, Overview of the NTCIR-17 FairWeb-1 Task, in: Proceedings of the 17th NTCIR (NII Testbeds and Community for Informtion Access Research) Conference, Tokyo, JP, 2024. URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=956867. doi:https://doi.org/10.20736/0002001318.

[11] F. Diaz, B. Mitra, M. D. Ekstrand, A. J. Biega, B. Carterette, Evaluating Stochastic Rankings with Expected Exposure, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 275–284. URL: https://doi.org/10.1145/3340531.3411962. doi:10.1145/3340531.3411962.

[12] A. Raj, M. D. Ekstrand, Comparing Fair Ranking Metrics, arXiv preprint arXiv:2009.01311 (2020).
[13] M. D. Ekstrand, G. McDonald, A. Raj, I. Johnson, Overview of the TREC 2022 Fair Ranking Track, in: The Thirty-First Text REtrieval Conference (TREC 2022) Proceedings, volume 500-338 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 2023. URL: https://trec.nist.gov/pubs/trec31/papers/Overview-FR.pdf, held online, November 15–19, 2022.
[14] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, D. Roth, A Comparative Study of Fairness-enhancing Interventions in Machine Learning, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 329–338. URL: https://doi.org/10.1145/3287560.3287589. doi:10.1145/3287560.3287589.
[15] M. Hardt, E. Price, N. Srebro, Equality of Opportunity in Supervised Learning, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, Curran Associates Inc., Red Hook, NY, USA, 2016, p. 3323–3331.
[16] P. Sapiezynski, W. Zeng, R. E Robertson, A. Mislove, C. Wilson, Quantifying the Impact of User Attention Fair Group Representation in Ranked Lists, in: Companion Proceedings of The 2019 World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 553–562. URL: https://doi.org/10.1145/3308560.3317595. doi:10.1145/3308560.3317595.
[17] M. Menéndez, J. Pardo, L. Pardo, M. Pardo, The Jensen-Shannon Divergence, Journal of the Franklin Institute 334 (1997) 307–318. URL: https://www.sciencedirect.com/science/article/pii/S0016003296000634. doi:https://doi.org/10.1016/S0016-0032(96)00063-4.
[18] M. D. Ekstrand, A. Raj, G. McDonald, I. Johnson, Overview of the TREC 2021 Fair Ranking Track, in: The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings, volume 500-335 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 2022. URL: https://trec.nist.gov/pubs/trec30/papers/Overview-FR.pdf, held online, November 15–19, 2021.
[19] E. Yang, T. Jänich, J. Mayfield, D. Lawrie, Language Fairness in Multilingual Information Retrieval, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 2487–2491. URL: https://doi.org/10.1145/3626772.3657943. doi:10.1145/3626772.3657943.
[20] T. Jaenich, G. McDonald, I. Ounis, Improving Exposure Allocation in Rankings by Query Generation, in: Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part V, Springer-Verlag, Berlin, Heidelberg, 2024, p. 121–129. URL: https://doi.org/10.1007/978-3-031-56069-9_9. doi:10.1007/978-3-031-56069-9_9.
[21] A. Abolghasemi, L. Azzopardi, A. Askari, M. de Rijke, S. Verberne, Measuring Bias in a Ranked List Using Term-Based Representations, in: Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part V, Springer-Verlag, Berlin, Heidelberg, 2024, p. 3–19. URL: https://doi.org/10.1007/978-3-031-56069-9_1. doi:10.1007/978-3-031-56069-9_1.
[22] S. Tao, T. Sakai, N. Chen, H. Li, Y. Tu, J. Wang, H. Fang, Y. Zhang, M. Maistro, Overview of the NTCIR-18 FairWeb-2 Task, in: Proceedings of the 18th NTCIR (NII Testbeds and Community for Informtion Access Research) Conference, Tokyo, JP, 2025. doi:https://doi.org/10.20736/0002002030.
[23] S. Tao, T. Sakai, Measuring Group Fairness in Web Search: AWRF or GFR?, in: Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP 2025), ACM, Xi'an, China, 2025, pp. 1–7. URL: https://doi.org/10.1145/3767695.3769504. doi:10.1145/3767695.3769504.
[24] T. Sakai, Fairness-based Evaluation of Conversational Search: A Pilot Study, Procedings of EVIA 2023 (2023) 5–13. URL: https://repository.nii.ac.jp/records/2001350. doi:10.20736/0002001350.
[25] T. Sakai, S. Tao, Y.-I. Song, Evaluating Group Fairness and Relevance in Conversational Search:

An Alternative Formulation, Procedings of EVIA 2025 (2025) 15–22. URL: https://cir.nii.ac.jp/crid/1390304471623016320. doi:10.20736/0002002107.

[26] S. C. Geyik, S. Ambler, K. Kenthapadi, Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 2221–2231. URL: https://doi.org/10.1145/3292500.3330691. doi:10.1145/3292500.3330691.

[27] K. Yang, J. Stoyanovich, Measuring Fairness in Ranked Outputs, in: Proceedings of the 29th International Conference on Scientific and Statistical Database Management, SSDBM '17, Association for Computing Machinery, New York, NY, USA, 2017. URL: https://doi.org/10.1145/3085504.3085526. doi:10.1145/3085504.3085526.

[28] A. Ghosh, R. Dutt, C. Wilson, When Fair Ranking Meets Uncertain Inference, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 1033–1043. URL: https://doi.org/10.1145/3404835.3462850. doi:10.1145/3404835.3462850.

[29] O. Chapelle, D. Metzler, Y. Zhang, P. Grinspan, Expected Reciprocal Rank for Graded Relevance, in: Proceedings of ACM CIKM 2009, 2009, pp. 621–630.

[30] M. Werman, S. Peleg, A. Rosenfeld, A Distance Metric for Multidimensional Histograms, Computer Vision, Graphics, and Image Processing 32 (1985) 328–336.

[31] T. Sakai, Evaluating Evaluation Measures for Ordinal Classification and Ordinal Quantification, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 2759–2769.

[32] T. Sakai, A Closer Look at Evaluation Measures for Ordinal Quantification, in: Proceedings of the CIKM 2021 Workshops, 2021.

[33] Y. Tu, H. Li, Z. Chu, Q. Ai, Y. Liu, THUIR at the NTCIR-17 FairWeb-1 Task: An Initial Exploration of the Relationship Between Relevance and Fairness, in: Proceedings of the 17th NTCIR (NII Testbeds and Community for Informtion Access Research) Conference, Tokyo, JP, 2023. doi:https://doi.org/10.20736/0002001317.

[34] F. Li, K. Shi, N. C. Kenta Inaba, Sijie Tao, T. Sakai, RSLFW at the NTCIR-17 FairWeb-1 Task, in: Proceedings of the 17th NTCIR (NII Testbeds and Community for Informtion Access Research) Conference, Tokyo, JP, 2023. doi:https://doi.org/10.20736/0002001303.

[35] Z. Chu, T. Sakai, Q. Ai, Y. Liu, Chuweb21D: A Deduped English Document Collection for Web Search Tasks, in: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 63–72. URL: https://doi.org/10.1145/3624918.3625317. doi:10.1145/3624918.3625317.

[36] A. Barbaresi, Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction, in: H. Ji, J. C. Park, R. Xia (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2021, pp. 122–131. URL: https://aclanthology.org/2021.acl-demo.15/. doi:10.18653/v1/2021.acl-demo.15.

[37] J. Carbonell, J. Goldstein, The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries, in: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98, Association for Computing Machinery, New York, NY, USA, 1998, p. 335–336. URL: https://doi.org/10.1145/290941.291025. doi:10.1145/290941.291025.

[38] J. Dean, S. Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, Commun. ACM 51 (2008) 107–113. URL: https://doi.org/10.1145/1327452.1327492. doi:10.1145/1327452.1327492.