Measuring the Fairness Gap Between Retrieval and Generation in RAG Systems using a Cognitive Complexity Framework

Sandeep Avula
Amazon AWS AI/ML
Seattle, Washington, USA
sandeavu@amazon.com

Chia-Jung Lee
Amazon AWS AI/ML
Seattle, Washington, USA
cjlee1010@gmail.com

Rongting Zhang
Amazon AWS AI/ML
Seattle, Washington, USA
rongtz@amazon.com

Vanessa Murdock*
Amazon AWS AI/ML
Seattle, Washington, USA
vmurdock@acm.org

Abstract

In this paper, we investigate the problem of quantifying fairness in Retrieval-Augmented Generation (RAG) systems, particularly for complex cognitive tasks that go beyond factual question-answering. While RAG systems have demonstrated effectiveness in information extraction tasks, their fairness implications for cognitively complex tasks - including ideation, content creation, and analytical reasoning — remain under-explored. We propose a novel evaluation framework that extends IR fairness metrics by incorporating centrality-based measures to account for influence of retrieved documents on generated output beyond ranking. Our framework evaluates RAG systems across various cognitive dimensions using two ranking approaches: lexical (BM25) and dense (BGE), and language models of varying sizes. Our findings provide insights into: (1) the propagation of fairness disparities from retrieval to generation phases, and (2) the variation in system performance across different cognitive dimensions.

CCS Concepts

• **Information systems** → **Evaluation of retrieval results**.

Keywords

Retrieval Augmented Generation, RAG, Fairness, Information Retrieval

ACM Reference Format:
Sandeep Avula, Chia-Jung Lee, Rongting Zhang, and Vanessa Murdock. 2025. 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 onResearch and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3726302.3730230

*The author had minor contributions.

This work is licensed under a Creative Commons Attribution International 4.0 License.

SIGIR '25, July 13–18, 2025, Padua, Italy
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730230

1 Introduction

Retrieval-Augmented Generation (RAG) systems have become increasingly popular following the widespread deployment and use of Large Language Models (LLMs). Their appeal lies in grounding LLM responses within specific contexts, thereby delivering focused responses while also mitigating hallucinations [14]. However, the combination of search systems and LLMs brings forth a new critical challenge: the compounding of fairness issues inherent to both information retrieval systems and LLMs themselves. Recent research has examined bias propagation in RAG systems, focusing primarily on Q&A, classification, and regression-style tasks [13, 14, 23]. In these tasks, the evaluation metrics are well-bounded and clearly defined. However, as enterprises increasingly deploy RAGs as productivity tools for employees [21], users are positioned to employ these systems for complex cognitive tasks like ideation, content creation, analytical reasoning, and other activities that go beyond simple information extraction.

Traditional fairness metrics in information retrieval, such as Expected Exposure [5] and attention-based measures [18, 20], were designed with clear assumptions about user behavior and relevance (i.e., monotonically decreasing attention with rank position, clear relevance judgments, and position-based exposure models that directly correlate with user attention patterns). In these traditional ranking systems, a document's exposure weight is derived primarily from its position in the ranking, following a monotonically decreasing function that models how user attention naturally decays as they traverse down a ranked list.

RAG systems fundamentally challenge these assumptions. Unlike conventional ranking systems where documents lower in the list consistently receive less attention, RAGs exhibit more complex and non-monotonic attention patterns from their Generator component. Recent work has shown that the Generator can place significant attention on passages appearing later in the context [15, 17, 27], with attention distributions that deviate substantially from the predictable decay patterns observed in traditional ranking systems. We confirm this observation in our empirical analysis (Figure 1). These fundamental differences motivate a new approach to measure exposure and fairness in RAG systems.

To address these challenges, we propose a novel evaluation framework that extends IR fairness metrics [6] by incorporating

centrality-based measures [16, 28] to account for influence of retrieved documents on generated output beyond ranking. Additionally, to evaluate the RAG systems for a variety of cognitive tasks, we use Anderson and Krathwahl's taxonomy of educational objectives [2], and focus on the cognitive process dimension. This particular framework has been used in prior IR studies to study user search behavior [22]. We experiment with different retrieval systems: lexical (BM25) and dense (BGE), different context sizes: {8, 16, 32}, and a generator using LLMs of different sizes and architectures. We investigate two research questions: **(RQ1)** *What is the fairness gap between retrieval and generation?* and **(RQ2)** *How is response fairness shaped by system components including (a) retriever architecture, (b) task complexity, and (c) generator characteristics?*

2 Related Work

Fairness in Information Retrieval: The measurement of fairness in information retrieval has been fundamentally shaped by empirical studies of how users interact with ranked results [12]. Traditional IR research has established that user attention broadly follows a monotonic decay as users traverse down a ranked list, making position a crucial factor in measuring fairness [6]. This understanding has influenced how we conceptualize and measure fairness across different stakeholders and definitions. Early work on fairness focused on group-level metrics, with approaches like statistical parity incorporating logarithmic position discounting [24] and extensions considering different prefixes of ranked lists [25]. The role of position and attention was further emphasized in individual fairness measures, such as amortized attention allocation [4] and pairwise exposure comparisons [3]. More recent frameworks have provided flexible approaches to modeling user attention patterns, such as the Attention-Weighted Ranking Fairness (AWRF) [20] and Expected Exposure of stochastic rankers [5]. These frameworks explicitly model how position affects exposure and consequently, fairness. These position-based fairness metrics, while effective for traditional search systems, need rethinking with the emergence of RAG systems where information consumption patterns by the Generator module of a RAG deviates significantly from user behavior.

Evaluation of RAG systems: The evaluation of RAG systems has primarily focused on end-to-end performance metrics [9, 11, 14], with recent work moving towards component-level effectiveness [8, 19, 23]. While these frameworks address system effectiveness, few studies have examined bias and fairness concerns in RAG systems beyond QA tasks [23]. Recent work has revealed several concerning patterns: Abolghasemi et al. [1] demonstrated systematic biases in how RAG systems attribute information to different types of authors, and Kim and Diaz [13] examined fairness issues in the ranking component and how it affects the quality of generation. Notably, recent studies show that the generator component in RAGs exhibits complex and non-monotonic attention patterns [26], differing significantly from how humans process ranked results [12]. This distinction challenges conventional IR metrics, and, in turn, fairness metrics that rely on position-based exposure models. Zhang et al. [26] further revealed that LLMs exhibit position bias when accessing information from retrieved documents. Building on these insights, we introduce a unified framework that accounts for the unique attention patterns of RAGs across different cognitive tasks.

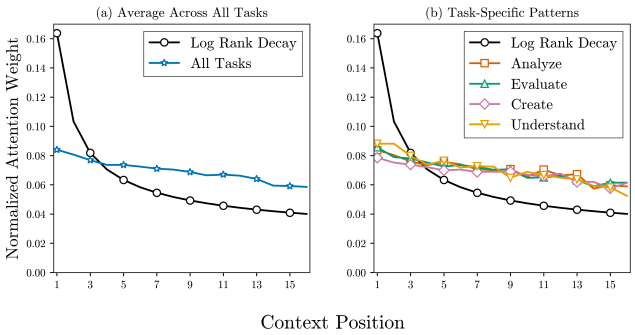

Figure 1: Comparison of attention weights: traditional rank-based decay versus our attribution-based weights for passages retrieved by the BGE ranker and processed by `mistral-large` (K=16).

3 Methods

Our approach to measuring fairness in RAG systems consists of three components: designing a benchmark for cognitive tasks, implementing varied RAG system configurations, and methodologies for defining fairness in RAGs.

Benchmark: In this study, we utilize and adapt the TREC 2022 Fair Ranking Track corpus [7]. This benchmark consists of 46 informational queries (topics), each of which is provided with a well-studied target distribution over individual and inter-sectional fairness dimensions such as gender or geolocation. Our choice of this dataset was motivated by two key factors: first, it provides rigorously established target distributions that are critical for developing our fairness metrics, and second, it enables evaluation of open-ended user inquiries across a broad spectrum of topics. This dataset is particularly relevant as RAG systems enable users to engage in complex tasks such as ideation, content creation, and analytical reasoning. We contrast this with QA and extractive tasks, as informational queries may lack definitive ground truth or clear boundaries for what constitutes sufficient information.

Table 1: Templates used for different cognitive dimensions

Dimension	Template
`Understand`	• What are the main characteristics and important aspects of {*topic*}? • How has {*topic*} developed and changed over time?
`Analyze`	• What factors have shaped or influenced {*topic*}? • What are the relationships between different aspects of {*topic*}?
`Evaluate`	• What are the major challenges and significant developments related to {*topic*}? • What aspects of {*topic*} have proven most valuable or significant, and why?
`Create`	• What possible future developments might we see in {*topic*}? • How could our understanding and study of {*topic*} be innovatively reimagined?

To systematically explore the diverse ways users interact with RAG systems, we adapt TREC topics following Anderson and Krathwohl's taxonomy of educational objectives [2]. We focus on four cognitive dimensions: `create` (generate new content by synthesizing information), `evaluate` (assess and compare information

to form conclusions), `analyze` (identify relationships across documents), and `understand` (interpret key concepts)[1]. For each dimension, we created two template query variants (Table 1), designed to be topic-agnostic while maintaining clear distinctions between cognitive dimensions. Using these 8 templates across our 46 topics, we generated 368 informational queries.

RAG Implementation: We implemented two ranking approaches: (1) `BM25` (from Pyserini[2]), a traditional lexical model, and (2) `BGE`[3], a dense retrieval model. Documents were split into 1024-character chunks with 50-character overlap. `BGE` was implemented as a reranker operating on the top 500 `BM25`-retrieved chunks. For generation, we used `Llama3` and `Mistral` models, chosen for their state-of-the-art performance and open-weight availability.

Fairness Metrics for Ranking: We quantify ranking fairness using Expected Exposure (EE-L), measuring the disparity between actual and target exposure distributions. Let A be an $n \times k$ binary matrix representing k group attributes for n retrieved passages. The actual exposure vector is:

$$\xi_A = A^T W, \quad \text{where } W = [w_1, \ldots, w_n]^T, \quad w_i = \frac{1}{\log(1+i)} \tag{1}$$

For the target exposure distribution, we define:

$$\xi^* = \frac{1}{2}(A^T r_q + a_{world}) \cdot \sum_{i=1}^{n} \frac{1}{\log(1+i)} \tag{2}$$

where r_q represents relevance judgments and a_{world} represents world population distribution for geographic attributes and equal distribution for gender attributes. The scaling factor ensures comparability with actual exposure. For intersectional fairness, we use the Cartesian product of attribute sets.

The EE-L metric for ranking is computed as:

$$\ell_{\texttt{ranker}}(\xi_A, \xi^*) = \|\xi_A - \xi^*\|_2^2 \tag{3}$$

Fairness Metrics for RAG Response: Unlike ranking systems where exposure is determined by rank position, RAG systems' exposure patterns depend on how the generator utilizes passages to construct responses. We adapt EE-L by measuring each passage's contribution to the generated text.

For a generated text $G = \{s_1, \ldots, s_m\}$ with retrieved passages $P = \{p_1, \ldots, p_n\}$, we compute attribution weights as follows:

1. Sentence-level centrality scores:

$$c_i = \frac{1}{m-1} \sum_{j \neq i} \text{BERTScore-F1}(s_i, s_j) \tag{4}$$

2. Passage-sentence entailment via RefChecker [10]:

$$E_{i,j} = \begin{cases} 1 & \text{if passage } p_j \text{ entails sentence } s_i \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

3. Attribution weights normalized to match ranking exposure scale:

$$w_j = \sum_i (c_i \times E_{i,j}) \cdot \frac{\sum_{i=1}^{n} \frac{1}{\log(1+i)}}{\sum_{j=1}^{n} \sum_i (c_i \times E_{i,j})} \tag{6}$$

The exposure vector $\xi_A = A^T W$ uses these weights, while target exposure ξ^* remains as defined in the ranking metric. The final EE-L metric is:

$$\ell_{\texttt{response}}(\xi_A, \xi^*) = \|\xi_A - \xi^*\|_2^2 \tag{7}$$

This approach captures both the importance of sentences (via centrality) and their information sources (via entailment), while ensuring fairness scores are comparable between ranking and generation through normalized weights.

Data Analysis: For RQ1, we used the Wilcoxon signed-rank test with a one-sided alternative hypothesis to compare retrieval and generation fairness scores. We analyzed both absolute scores and relative differences to understand fairness degradation. For RQ2, we employed mixed-effects regression models to analyze how $\ell_{\texttt{response}}$ varies with ranker type, $\ell_{\texttt{ranker}}$, task type, context size, and LLM choice, using `topic_id` as the random effect.

4 Results

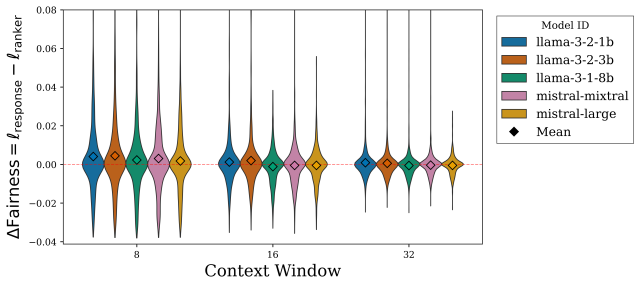

Figure 2: Comparison of ranking ($\ell_{ranking}$) and generation fairness ($\ell_{response}$) across different model sizes and context windows.

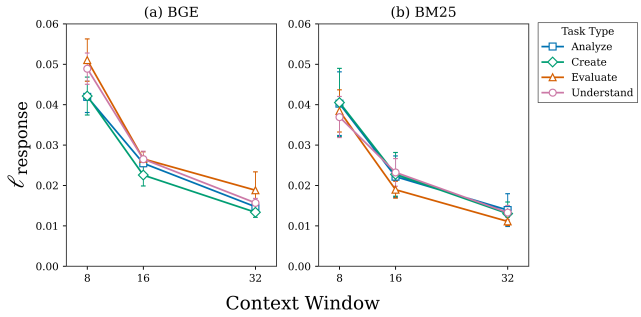

Figure 3: $\ell_{response}$ across cognitive dimensions and context sizes.

RQ1: What is the fairness gap between retrieval and generation? Our analysis suggests a consistent fairness degradation from retrieval to generation across all context sizes. For small contexts (K=8), the generated responses were significantly more unfair (i.e., higher EE-$L_{\texttt{response}}$) compared to retrieved results ($p < 0.001$).

[1]Two dimensions were excluded: "remember" (fact recall) and "apply" (procedural execution), as they are less relevant for our investigation.
[2]https://github.com/castorini/pyserini
[3]https://huggingface.co/BAAI/bge-large-en

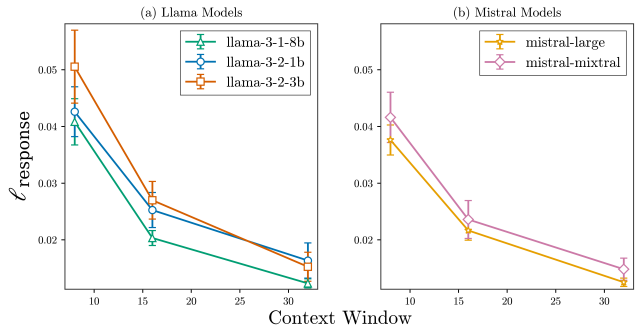

Figure 4: Fairness patterns across model sizes and context windows. Larger models achieve better fairness in both `Llama` (left) and `Mistral` (right) families.

While this gap narrowed with larger contexts (K=16, K=32), our analysis confirms that generation consistently produces less fair results than retrieval across all context sizes ($p < 0.01$). This demonstrates the existence of a persistent fairness gap between retrieval and generation components, though its severity can be moderated by increasing context size.

RQ2a: How is response fairness shaped by the retriever architecture? Our analysis reveals two distinct aspects of how retrievers influence fairness in RAG systems. First, the initial retrieval fairness (EE-L_{ranker}) strongly determines response fairness, though this relationship linearly reduces as context size increases (K=8, β=0.969, $p < 0.001$); (K=16, β=0.800, $p < 0.001$); (K=32, β=0.770, $p < 0.001$). This suggests that while generated responses largely inherit the fairness characteristics (bias) of retrieved passages, larger contexts provide the generator more flexibility to deviate from these initial bias patterns.

The architectural impact of retrievers (BM25 vs BGE) presents an interesting dynamic, particularly evident in small contexts. At K=8, if both retrievers achieved the same level of retrieval fairness, BGE would lead to significantly fairer responses than BM25 (β=0.011, p=0.002). For instance, with a retrieval fairness score of 0.02, BGE would yield responses with fairness around 0.019, while BM25 would yield less fair responses around 0.030. This architectural advantage of BGE diminishes with larger contexts, becoming non-significant at K=16 (β=0.003, p=0.238) and K=32.

RQ2b: How is response fairness shaped by task complexity? Our analysis suggests that the task types' impact on response fairness happens only with larger context windows (K=32), with no significant differences observed at smaller contexts. At larger contexts, both `create` (β=−0.004, p=0.039) and `understand` tasks (β=−0.004, p=0.031) show higher fairness in comparison to `evaluate` tasks. This suggests that for `create` and `understand` tasks, the generator module is able to source information from passages in a manner that better aligns with the target distribution.

In terms of interactions between task type and retriever choice, we observe significant effects at large context windows (K=32). We find that BGE helps maintain more consistent fairness levels across different task types, while BM25 shows more variation in fairness depending on the task type. Specifically, when using BM25, `analyze` (β=0.006, p=0.025) and `create` tasks (β=0.005, p=0.038) show significantly higher EE-$L_{response}$ scores (less fair) compared to `evaluate` tasks.

RQ2c: How is response fairness shaped by generator characteristics? For model size effects, we find a consistent pattern across both the `Llama` and `Mistral` model families: larger models tend to produce fairer responses. In the `Llama` family (`3-1-8b` > `3-2-3b` > `3-2-1b`), we observe that the largest model `3-1-8b` consistently generates responses that better align with the target distribution. At context K=8, the middle-sized model `3-2-3b` shows significantly higher unfairness compared to both the largest model (β=0.011, p=0.001) and the smallest model (β=0.008, p=0.011). At K=16, both smaller models show significantly higher unfairness compared to `3-1-8b` (`3-2-3b`: β=0.007, $p < 0.001$; `3-2-1b`: β=0.006, p=0.002). The trend continues at context K=32, though with smaller effect sizes (`3-2-1b` vs `3-1-8b`: β=0.004, p=0.010). Similarly, in the `Mistral` family, we observe that the larger model `mistral-large` produces fairer responses compared to `mixtral` at context K=32 (β=0.002, p=0.012). These findings suggest that larger models are better at sourcing information from the passages in a manner that better aligns with the target distribution.

5 Discussion and Conclusion

Our results highlight an interesting interplay between the retrieval and generation components in RAG systems' fairness outcomes. The strong and consistent effect of EE-L_{ranker} across all contexts and model sizes demonstrates significant propagation of retrieval biases to the final response. However, the influence of the ranker's biases linearly decreases with an increase in the context window, allowing the generator module to introduce its own patterns of bias, potentially through selective attention to certain passages or emphasis on particular aspects of the retrieved content. Interestingly, these generator-introduced biases vary systematically with model size. Larger models consistently produce fairer responses despite receiving the same retrieved passages as their smaller counterparts, suggesting that increased model capacity helps in better utilizing retrieved information while minimizing the introduction of additional biases.

Task complexity's impact on fairness emerges only with larger contexts (K=32), while at smaller contexts, response unfairness primarily stems from the retrieval component. At K=32, BGE maintains consistent fairness across task types, while BM25 shows variable performance - particularly for `analyze` (β=0.006, p=0.025) and `create` tasks (β=0.005, p=0.038) which show higher unfairness compared to `evaluate` tasks. This suggests dense retrievers may maintain stable fairness across cognitive tasks with larger context windows.

References

[1] Amin Abolghasemi, Leif Azzopardi, Seyyed Hadi Hashemi, Maarten de Rijke, and Suzan Verberne. 2024. Evaluation of Attribution Bias in Retrieval-Augmented Large Language Models. *arXiv preprint arXiv:2410.12380* (2024).

[2] Lorin W Anderson and David R Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition.* Addison Wesley Longman, Inc.

[3] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2212–2220.

[4] 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. 405–414.

[5] Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. 2020. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 275–284.

[6] Michael D Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz, et al. 2022. Fairness in information access systems. *Foundations and Trends extregistered in Information Retrieval* 16, 1-2 (2022), 1–177.

[7] Michael D Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. 2023. Overview of the TREC 2022 fair ranking track. *arXiv preprint arXiv:2302.05558* (2023).

[8] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. 150–158.

[9] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*. PMLR, 3929–3938.

[10] Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. RefChecker: Reference-based Fine-grained Hallucination Checker and Benchmark for Large Language Models. (2024). arXiv:2405.14486 [cs.CL]

[11] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research* 24, 251 (2023), 1–43.

[12] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *Acm Sigir Forum*, Vol. 51. Acm New York, NY, USA, 4–11.

[13] To Eun Kim and Fernando Diaz. 2024. Towards Fair RAG: On the Impact of Fair Ranking in Retrieval-Augmented Generation. *arXiv preprint arXiv:2409.11598* (2024).

[14] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[15] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173.

[16] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*. 404–411.

[17] Guanghui Qin, Yukun Feng, and Benjamin Van Durme. 2022. The NLP task effectiveness of long-range transformers. *arXiv preprint arXiv:2202.07856* (2022).

[18] Amifa Raj and Michael D Ekstrand. 2020. Comparing fair ranking metrics. *arXiv preprint arXiv:2009.01311* (2020).

[19] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 338–354.

[20] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the impact of user attentionon fair group representation in ranked lists. In *Companion proceedings of the 2019 world wide web conference*. 553–562.

[21] Olivia Shone. 2025. *5 key features and benefits of retrieval augmented generation (RAG)*. Microsoft. https://www.microsoft.com/en-us/microsoft-cloud/blog/2025/02/13/5-key-features-and-benefits-of-retrieval-augmented-generation-rag/ Accessed: 2025-02-17.

[22] Kelsey Urgo, Jaime Arguello, and Robert Capra. 2020. The effects of learning objectives on searchers' perceptions and behaviors. In *Proceedings of the 2020 acm sigir on international conference on theory of information retrieval*. 77–84.

[23] Xuyang Wu, Shuowei Li, Hsin-Tai Wu, Zhiqiang Tao, and Yi Fang. 2025. Does RAG Introduce Unfairness in LLMs? Evaluating Fairness in Retrieval-Augmented Generation Systems. In *Proceedings of the 31st International Conference on Computational Linguistics*. 10021–10036.

[24] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th international conference on scientific and statistical database management*. 1–6.

[25] 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*. 1569–1578.

[26] Meiru Zhang, Zaiqiao Meng, and Nigel Collier. 2024. Can We Instruct LLMs to Compensate for Position Bias?. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 12545–12556.

[27] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

[28] Hao Zheng and Mirella Lapata. 2019. Sentence Centrality Revisited for Unsupervised Summarization. arXiv:1906.03508 [cs.CL] https://arxiv.org/abs/1906.03508