Stochastic Retrieval for Fairness–Accuracy Trade-offs in Clinical RAG

Stochastic Retrieval for Fairness–Accuracy Trade-offs in Clinical RAG

Riccardo Lunardi
University of Udine
Udine, Italy
riccardo.lunardi@uniud.it

Vincenzo Della Mea
University of Udine
Udine, Italy
vincenzo.dellamea@uniud.it

Carsten Eickhoff
University of Tübingen
Tübingen, Germany
carsten.eickhoff@uni-tuebingen.de

Kevin Roitero
University of Udine
Udine, Italy
kevin.roitero@uniud.it

Abstract

Retrieval-Augmented Generation (RAG) systems are increasingly employed in clinical decision support, where patient cases are enriched with similar past records to guide downstream reasoning. However, clinical corpora often exhibit distributional imbalances, leading RAG systems to overlook underrepresented cases, thus limiting retrieval diversity and introducing potential bias in predictions. In this work, we study the trade-off between retrieval diversity (defined as capturing a broader range of patient contexts) and downstream task accuracy in a stochastic RAG pipeline applied to MIMIC-III discharge summaries. Given a clinical note, the model must infer the procedure a patient underwent based on top-k retrieved similar cases and the current one. We evaluate multiple task variants and fairness metrics to quantify representational fairness across demographic groups. Our results show that encouraging diversity in retrieval reduces demographic bias but that comes at the expense of accuracy. To address this, we explore the role of a fairness parameter that enables controllable trade-offs between fairness and accuracy, ultimately providing a more equitable and effective RAG system for clinical applications.

CCS Concepts

• **Information systems** → **Information retrieval diversity**; **Evaluation of retrieval results**; • **Computing methodologies** → **Natural language generation**.

Keywords

Retrieval-Augmented Generation, Fairness, Large Language Models, MIMIC-III

ACM Reference Format:
Riccardo Lunardi, Vincenzo Della Mea, Carsten Eickhoff, and Kevin Roitero. 2026. Stochastic Retrieval for Fairness–Accuracy Trade-offs in Clinical RAG. In *17th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '26), June 30–July 03, 2026, Rende (CS), Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3807503.3819374

1 Introduction

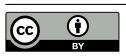
Retrieval-Augmented Generation (RAG) systems are increasingly used in clinical decision support systems, where Large Language Models (LLMs) retrieve relevant background documents such as previous patient notes or medical literature to improve downstream reasoning task [21, 26]. A common use case is inferring the set of primary medical procedures a patient underwent from their hospital discharge summary and a set of retrieved cases describing similar patients. This task is essential for administrative coding, clinical auditing, insurance reimbursement, and downstream decision support.

However, clinical corpora reflect real-world healthcare inequalities, with majority groups over-represented and minority populations under-documented. Such skewed distributions reduce the diversity of contexts retrieved by the RAG system. While in some cases local patterns may suffice for coding, over-reliance on majority-group evidence risks entrenching health disparities, and clinical decision support often benefits from exposure to a broader range of patient data [7, 10]. A model that consistently retrieves only cases from white, male, or well-insured patients may appear to perform well on average, while systematically failing on minority groups, missing rare procedures or atypical presentations that are more common in those populations [24]. Diversity in retrieved contexts is therefore crucial: it reduces the risk of missed diagnoses, captures heterogeneity across patient populations, and supports generalization across subgroups. [30].

Despite this, most existing RAG systems focus primarily on optimizing for accuracy, with little attention to fairness or diversity in the retrieved context. Recent work has started to explore these concerns, showing that stochastic retrieval techniques can introduce fairness-aware trade-offs [15, 36]. Without fairness guarantees, however, improvements in average accuracy may mask systematically uneven performance across groups. We therefore treat fairness as a prerequisite for reliable clinical RAG pipelines, not as a secondary concern.

In this work, we study the trade-off between accuracy and fairness in clinical procedure prediction. Given a discharge summary, the model predicts the patient's primary procedure using similar notes as context retrieved through a stochastic RAG pipeline. We then assess how retrieval diversity shapes fairness and performance across task variants, demographic groups, and fairness-oriented metrics. We address the following research questions:

- RQ1 In a RAG retrieval setting, how do accuracy and fairness in ranked exposure differ across demographic groups?
- RQ2 To what extent do fairness metrics capture disparities in deterministic versus stochastic RAG retrieval settings?
- RQ3 Can a tunable fairness parameter provide a controllable balance between accuracy and fairness in clinical RAG?



BCB '26, Rende (CS), Italy

ACM ISBN 979-8-4007-2653-8/26/06
https://doi.org/10.1145/3807503.3819374

Our results show that increasing retrieval diversity improves group balance but slightly reduces accuracy. To address this trade-off, we introduce a tunable fairness parameter α, enabling practitioners to improve fairness with negligible loss in predictive performance. The code, data, and all supplementary material are available at https://osf.io/4cw3b

2 Background and Related Work

RAG has emerged as a powerful paradigm for enhancing LLMs with external knowledge, mitigating hallucinations and improving factual accuracy [18]. These advantages have driven adoption in high-stakes domains such as biomedicine and clinical decision support, where contextually reliable information is critical.

Despite their promises, RAG pipelines are not immune to bias, which may arise in multiple components: (i) retrieval bias, where ranking functions disproportionately favor majority-group contexts, leading to unequal exposure [16]; and (ii) generation bias, where the LLM may amplify disparities already present in the retrieved context [23].

Fairness concerns in retrieval and ranking have been extensively studied in Information Retrieval (IR) [28, 33]. Exposure bias highlights that higher-ranked items are more likely to be consumed and thus to influence downstream decisions. Approaches such as pooling [19], stochastic ranking [37], Gumbel-based perturbations [12], and Plackett-Luce models [25] have been proposed to mitigate disparities in exposure across subgroups. These ideas have recently been extended to RAG: Kim and Diaz [15] show that fairness-aware ranking substantially reduces subgroup disparities without significant utility loss, while Wu et al. [36] analyze how retrieval bias propagates through the pipeline and propose mitigation strategies.

A central challenge in this line of research is the fairness-accuracy trade-off: increasing retrieval diversity can improve subgroup coverage and mitigate disparities but often at the expense of relevance [22, 31]. While prior work has explored this tension in general-purpose question answering, its implications for clinical RAG remain underexplored. This gap motivates our study: by systematically varying retrieval depth (k) and introducing a stochastic sampling mechanism, we investigate how retrieval diversity shapes the fairness-accuracy trade-off in RAG systems applied to hospital discharge summaries.

3 Methodology

We describe our methodology in the following subsections, covering the task, data, RAG system, and evaluation metrics.

3.1 Task

The primary task in this work is to identify a patient's main procedure from their discharge summary, framed as a multiple-choice question answering problem. While this formulation simplifies real-world clinical settings, where candidate procedures are not explicitly provided, it enables controlled evaluation by reducing variability due to open-ended generation, allowing us to isolate the effect of retrieval on fairness and accuracy.

Given a discharge summary, the system is provided with a set of candidate procedure titles drawn from past patient cases, where only the primary ICD-9-CM procedure title associated with each case is used as a contextual cue. For example, consider a 68-year-old patient admitted with bronchopneumonia, who developed acute respiratory failure and required mechanical ventilation, along with ancillary procedures needed for treatment (e.g., radiographs, blood exams, etc.). In this scenario, the retrieved context might include procedures linked to similar respiratory cases. The LLM must identify the procedure most consistent with the current discharge summary (e.g., "Endotracheal Intubation"). The multiple-choice answer set always includes the main correct procedure alongside distractors randomly sampled from ICD-9-CM. The following prompt illustrates the interaction:

```
The patient was admitted to the hospital. Consider the
following patient note: [discharge_summary]. Similar
patients went through the following procedures:
[candidate_procedures].
What procedure did the patient undergo? Reply with the
number referring to the correct procedure the patient went
through:
1.[choice_1] | 2.[choice_2] | 3.[choice_3] | 4.[choice_4]
Answer:
```

We use the same prompt across all experiments for consistency and reproducibility [1, 29], along with a four-option, single-correct configuration, as it is widely used in existing benchmarks [6, 9] and provides a balance between task difficulty and interpretability of results.

3.2 Data

We base our analysis on discharge summaries from MIMIC-III [14], a large publicly available critical care dataset. Although MIMIC-IV-Note is more recent and also accessible, it lacks demographic information and is therefore unsuitable for fairness-oriented analyses. A discharge summary is a semi-structured clinical note that documents a patient's hospital stay, including treatments and outcomes. In our setup, each summary is treated as the patient note of interest. The dataset contains 59,652 discharge summaries, each associated with one or more ICD-9-CM procedure codes. We conduct experiments on a subset of 4,485 summaries sampled to preserve the demographic distribution of the full dataset. Running stochastic RAG on the entire corpus would be prohibitively expensive: for each instance we retrieve up to 100 candidate documents and perform 100 stochastic sampling runs per combination of α and k, resulting in $8,970,000$ prompt evaluations.

We analyze fairness with respect to four demographic attributes, reported in Table 1. Splits follow majority/minority representation, with the largest demographic category designated as the non-protected group [35, 38]. All minority categories are aggregated under "All others" to ensure sufficient sample sizes for reliable statistical analysis, as several individual groups in MIMIC-III are sparsely represented. While this improves the stability of fairness estimates, it reduces granularity and may mask subgroup-specific disparities.

3.3 RAG System

We evaluate two variants of the RAG pipeline: one employing deterministic retrieval and one employing stochastic retrieval. In the deterministic setting, the retriever selects the top-k documents

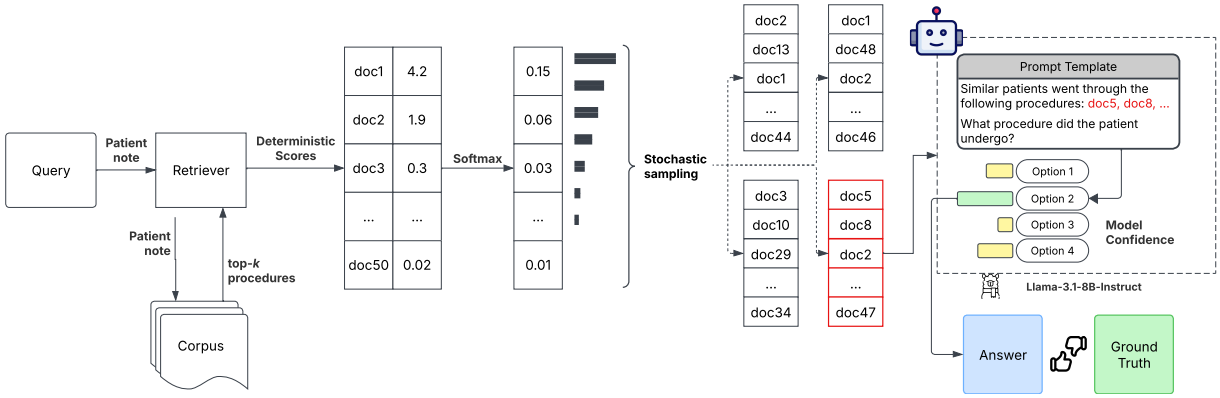


Figure 1: End-to-end RAG pipeline for clinical procedure prediction.

Table 1: Demographic groups considered for fairness.

Attribute	Non-Protected	Protected
Ethnicity	White (80%)	African American (10%), Asian (3%), others (7%)
Insurance	Medicare (54%)	Private (34%), Medicaid (9%), others (3%)
Religion	Catholic (38%)	Not specified (24%), Quaker (15%), others (23%)
Gender	Male (57%)	Female (43%)

strictly by relevance scores, passing this fixed set of contexts to the generator. In contrast, the stochastic setting introduces controlled randomness in document selection through Gumbel sampling [17], allowing alternative but still relevant contexts to be considered. Both systems adopt a BM25 retriever [32] to identify relevant passages from the MIMIC-III discharge summaries, and `Llama-3.1-8B-Instruct`[1] as the generator. An overview of the complete RAG pipeline, including both deterministic and stochastic retrieval variants, is illustrated in Figure 1. The figure highlights the flow from discharge input, through document retrieval and selection, to the final LLM-based procedure prediction.

We vary the retrieval depth $k \in \{5, 15, 25, 50, 100\}$. The degree of stochastic randomness is controlled by the fairness parameter α, which sharpens or flattens the sampling distribution to control the balance between relevance and diversity. Specifically, a low α flattens scores, increasing randomness and diversity in the sampled documents, while a high α accentuates score differences, making retrieval increasingly similar to the deterministic baseline. We vary $\alpha \in \{1, 2, 4, 8\}$, following prior work [15]. For each combination of α and k, we repeat the stochastic sampling 100 times, averaging results to obtain stable estimates of accuracy and fairness. This design disentangles the effects of retrieval depth (k) and stochasticity level (α): while k controls how much contextual information is available to the RAG system, α directly controls the balance between exploitation of high-scoring documents and exploration of lower-ranked but potentially diverse cases. We intentionally fix the pipeline (BM25 + Llama-3.1-8B-Instruct) to isolate the effect of stochastic retrieval; since the approach operates at the retrieval level, it is largely model-agnostic and can be combined with stronger retrievers or generators in future work.

[1]https://ai.meta.com/research/publications/the-llama-3-herd-of-models

3.4 Evaluation Metrics

We evaluate accuracy and fairness using complementary metrics. Accuracy Difference (AD) [5] quantifies disparities in accuracy across groups as the difference between the non-protected and protected group accuracy, with $AD = 0$ corresponding to *Overall Accuracy Equality* [4, 34]. This measure has been widely adopted to assess fairness in classification tasks [13, 27].

For fairness in ranked outputs, we adopt the Attention-Weighted Rank Fairness (AWRF) [28], which requires that the relative exposure of each group be proportional to its prevalence in the corpus. AWRF ranges from 0 and 1, with 0 denoting the ideal value of fairness.

In human-facing retrieval systems, attention is typically assumed to decay with rank position, often following exponential or logarithmic decay functions [2, 3]. In this context, what matters is not click-through probability, but rather which documents are injected into the generator's context window. Unlike human-facing retrieval systems, where attention decays with rank position, LLM generators can, in principle, process all retrieved documents simultaneously. Recent evidence suggests that this assumption may not hold perfectly in practice, as LLMs may fail to recall certain pieces of information even when they are present in the prompt, as demonstrated in "needle-in-a-haystack" evaluations [20]. Nonetheless, because the generator has access to the entire retrieved set simultaneously, and following recent efforts to reduce position bias in machine-oriented evaluation [8, 11], we consider uniform attention a more appropriate approximation than position-decay models designed for human browsing behavior.

4 Results

We analyze the three research questions in the following subsections.

(a) AD.

(b) AWRF.

Figure 2: Fairness metrics (AD, AWRF) comparing protected and non-protected groups under deterministic retrieval. Each plot represents a demographic group, each cell shows the metric value for a given retrieval depth k.

4.1 RQ1 Accuracy and fairness in deterministic RAG

We begin by examining the accuracy of deterministic RAG under different retrieval depths. Starting from a zero-shot baseline of 0.80, accuracy rises to 0.92 at $k = 5$, and then declines to 0.90 for larger retrieval depths. This suggests that a small number of highly relevant documents substantially benefits performance, while retrieving too many introduces noise that dilutes the most useful cases.

Figure 2a reports AD across demographic groups. Accuracy gaps are consistently small and generally not statistically significant ($p > 0.05$). The largest deviation occurs for Ethnicity at $k = 5$ (AD = 0.01), while for Gender the protected group (Female) marginally outperforms the non-protected group (Male) by 0.01. These results indicate that deterministic retrieval does not systematically disadvantage either group in terms of accuracy.

We next consider fairness in the rankings using AWRF. Figure 2b shows that unfairness is highest (least fair) at $k = 5$, and decreases as more documents are retrieved, as larger retrieval sets better reflect the underlying corpus distribution. Gender consistently shows the highest unfairness values even as k increases, highlighting that some attributes are more sensitive to deterministic retrieval than others. All differences are statistically significant ($p < 0.01$).

Together, these results confirm that deterministic RAG provides strong accuracy gains over zero-shot inference, but exposure disparities persist across groups even at large k, and the partial mitigation they offer comes at the cost of slightly lower accuracy.

4.2 RQ2 Fairness in stochastic vs. deterministic RAG

We compare deterministic and stochastic retrieval with respect to fairness, introducing the parameter α to control the relevance–fairness trade-off. Results are reported in Figures 3a and 3b. To improve readability, we show representative demographic attributes: Gender and Insurance for AD, and Ethnicity and Gender for AWRF.

Figure 3a shows that AD values are consistently closer to zero, particularly for $\alpha = 8$, which approximates the deterministic regime. This indicates that stochastic retrieval reduces disparities in group-level accuracy while maintaining overall performance. The most notable improvements occur for Gender at $k = 25$ and Ethnicity at $k = 100$ with $\alpha \in \{1, 2\}$, where the accuracy gap between protected and non-protected groups decreases by 0.01 and 0.02 ($p < 0.01$).

Figure 3b shows that stochastic retrieval consistently yields AWRF values closer to zero across all groups, confirming more equitable exposure. The best fairness values (highlighted in blue) are concentrated in the upper-right region of the plots, corresponding to high k and low α: larger retrieval sets increase representational diversity, while lower α distributes exposure more evenly. Differences between $\alpha = 1$ and $\alpha = 8$ are statistically significant in nearly all cases ($p < 0.05$), as are differences across k values for all groups.

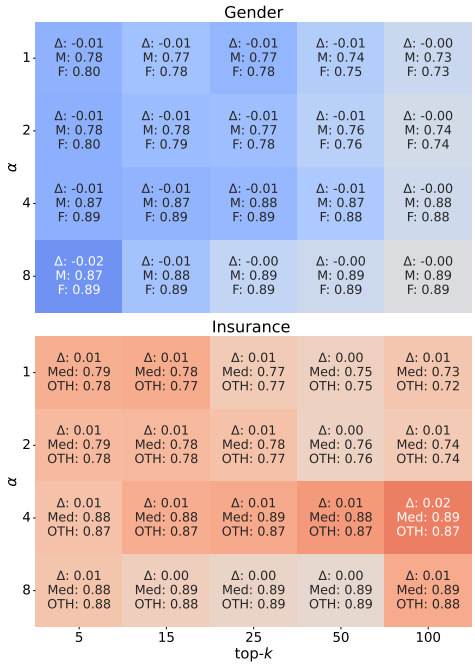
To examine whether disparities concentrate at specific rank positions, we extend the AWRF evaluation to sub-rank intervals of 15 documents using the $k = 100$ results, illustrated in Figure 4. For Insurance and Gender, top-ranked documents (positions 1–14) exhibit the highest AWRF, indicating that the most prominent results are less representative of protected groups, with fairness improving further down the ranking. This pattern disappears at $\alpha < 4$, confirming that lower values of the fairness parameter promote equitable exposure already at the very top of the ranking, without requiring large retrieval sets. For Ethnicity and Religion, AWRF remains low and stable across all sub-ranks and settings.

Overall, stochastic retrieval moderates the accuracy-fairness trade-off: while deterministic retrieval maximizes accuracy, it amplifies exposure disparities across groups. Stochastic retrieval narrows these disparities in both accuracy and exposure, including within the most visible portion of the rankings.

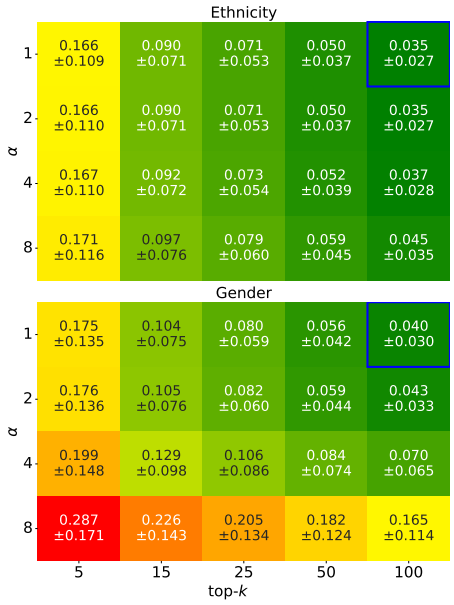
4.3 RQ3 Controlling the fairness–accuracy trade-off

Figure 5 summarizes the joint effect of α and k on accuracy and fairness. Increasing α consistently improves accuracy but worsens fairness, with the effect being most pronounced for smaller values of k. Larger k values mitigate the trade-off by broadening the retrieval space, though at the cost of slightly lower accuracy. Deterministic retrieval peaks at $k = 5$ (accuracy = 0.92), but this setting also corresponds to its least fair rankings (AWRF = 0.32).

Stochastic retrieval offers a balanced compromise. For $\alpha \geq 4$, accuracy exceeds the zero-shot baseline of 0.80 across all k, confirming the effectiveness of controlled stochasticity. The most balanced



(a) AD.



(b) AWRF.

Figure 3: Fairness metrics (AD, AWRF) comparing protected and non-protected groups under stochastic retrieval. Each plot represents a demographic group, each cell shows the metric value under a given top-k and fairness parameter α.

configuration is $\alpha = 4$ with $k > 5$, which approaches deterministic accuracy while substantially reducing unfairness. At $\alpha = 8$, accuracy is strong across all k values but fairness degrades for small k, reflecting a convergence toward accuracy-dominated retrieval.

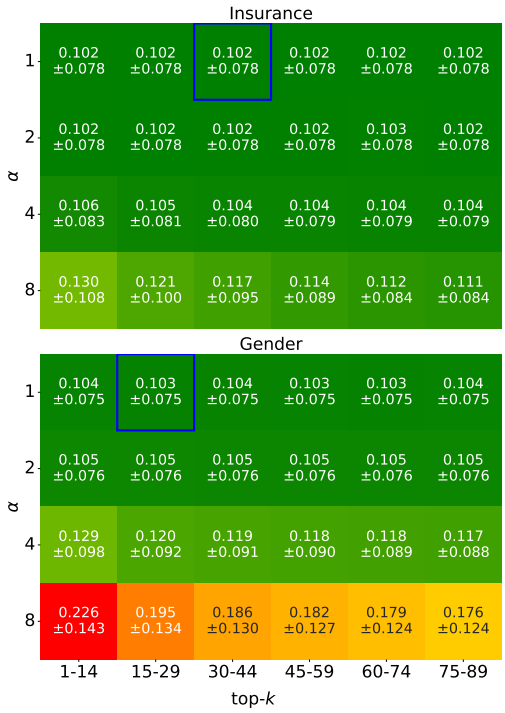


Figure 4: Comparison of fairness (AWRF) between protected and non-protected groups for stochastic retrieval across successive sub-ranks of 15 documents.

Group-specific markers in the plot reveal disparities in fairness across demographic groups, with some consistently positioned at higher AWRF values. This suggests that representational bias persists even when global accuracy is high, underlining the importance of tunable parameters for balancing fairness and accuracy.

5 Conclusions and Future Work

In this work, we study the trade-off between fairness and accuracy in RAG for clinical procedure prediction. Introducing stochasticity into retrieval reduces disparities in both accuracy and exposure, yielding fairer outcomes across demographic groups, while the tunable parameter α provides practitioners with a direct mechanism to balance representation equity and predictive performance. Our results underscore that fairness must be treated as a key priority in clinical RAG pipelines. Even when global accuracy improves, systematic under-representation of minority groups can cause failures on at-risk populations, amplifying existing health inequities.

The study has three main limitations: (i) experiments rely on a subset of MIMIC-III; (ii) minority categories are aggregated into a single "All others" group, potentially masking subgroup-specific effects; and (iii) we evaluate a single retriever–generator pipeline (BM25 + Llama-3.1-8B-Instruct) to isolate the effect of stochastic retrieval, which may limit generalizability to other architectures.

Addressing these constraints, extending the evaluation to the full dataset, and testing with stronger models are left to future work.

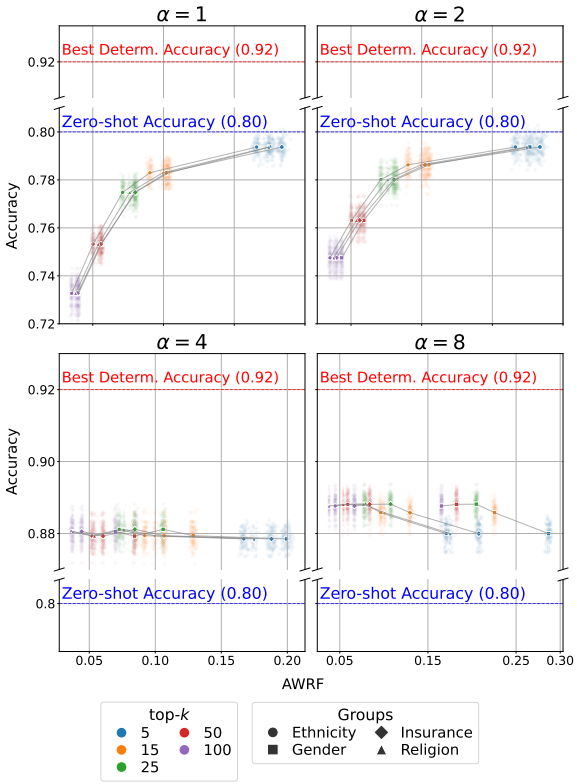


Figure 5: Joint fairness–accuracy trade-off across values of α and retrieval depth k.

Acknowledgments

This work was supported in part by the ESF+ 2021/2027 Regional Program of the Autonomous Region Friuli Venezia Giulia (PPO 2023, Program No. 22/23 – LINE A: PhD programmes).

References

[1] Negar Arabzadeh and Charles L.A. Clarke. 2025. A Human-AI Comparative Analysis of Prompt Sensitivity in LLM-Based Relevance Judgment. In *Proc. SIGIR* (Padua, Italy) *(SIGIR '25)*. ACM, 5 pages.
[2] Leif Azzopardi. 2021. Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval. In *Proc. CHIIR* (Canberra ACT, Australia) *(CHIIR '21)*. ACM, 11 pages.
[3] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (May 2018).
[4] Richard Berk, Hoda Heidari, Shahin Jabbari, et al. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 50, 1 (2021).
[5] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. 2021. Bias in machine learning software: why? how? what to do?. In *Proc. ESEC/FSE* (Athens, Greece) *(ESEC/FSE 2021)*. ACM, 12 pages.
[6] Peter Clark, Isaac Cowhey, Oren Etzioni, et al. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. arXiv:1803.05457 [cs.AI]
[7] Luis Emilio Gomez and Patrick Bernet. 2019. Diversity improves performance and outcomes. *Journal of the National Medical Association* 111, 4 (2019).
[8] Junqing He, Kunhao Pan, Xiaoqun Dong, et al. 2024. Never Lost in the Middle: Mastering Long-Context Question Answering with Position-Agnostic Decompositional Training. In *Proc. ACL*. ACL, Bangkok, Thailand.
[9] Dan Hendrycks, Collin Burns, Steven Basart, et al. 2021. Measuring Massive Multitask Language Understanding. *Proc. ICLR* (2021).
[10] Tina Hernandez-Boussard, Shazia Mehmood Siddique, Arlene S Bierman, et al. 2023. Promoting equity in clinical decision making: dismantling race-based medicine: commentary examines promoting equity in clinical decision-making. *Health Affairs* 42, 10 (2023).
[11] Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, et al. 2024. Found in the middle: Calibrating Positional Attention Bias Improves Long Context Utilization. In *Findings of ACL*. ACL, Bangkok, Thailand.
[12] Siyuan Huang, Zhiyuan Ma, Jintao Du, et al. 2025. Gumbel Reranking: Differentiable End-to-End Reranker Optimization. In *Proc. ACL*. ACL, Vienna, Austria.
[13] Disi Ji, Padhraic Smyth, and Mark Steyvers. 2020. Can I Trust My Fairness Metric? Assessing Fairness with Unlabeled Data and Bayesian Inference. In *Adv. NeurIPS*, Vol. 33. Curran Associates, Inc.
[14] Alistair E W Johnson, Tom J Pollard, Lu Shen, et al. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3, 1 (May 2016).
[15] To Eun Kim and Fernando Diaz. 2025. Towards Fair RAG: On the Impact of Fair Ranking in Retrieval-Augmented Generation. In *Proc. ICTIR* (Padua, Italy) *(ICTIR '25)*. ACM, 11 pages.
[16] Anja Klasnja, Negar Arabzadeh, Mahbod Mehrvarz, et al. 2022. On the Characteristics of Ranking-based Gender Bias Measures. In *Proc. WWW* (Barcelona, Spain) *(WebSci '22)*. ACM, 5 pages.
[17] Wouter Kool, Herke Van Hoof, and Max Welling. 2020. Ancestral Gumbel-top-k sampling for sampling without replacement. *J. Mach. Learn. Res.* 21, 1, Article 47 (Jan. 2020), 36 pages.
[18] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Adv. NeurIPS*, Vol. 33. Curran Associates, Inc.
[19] Aldo Lipani. 2016. Fairness in Information Retrieval. In *Proc. SIGIR* (Pisa, Italy) *(SIGIR '16)*. ACM, 1 pages.
[20] Nelson F. Liu, Kevin Lin, John Hewitt, et al. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Trans. Assoc. Comput. Linguist.* 12 (2024).
[21] Yuxing Lu, Xukai Zhao, and Jinzhuo Wang. 2024. ClinicalRAG: Enhancing Clinical Decision Support through Heterogeneous Knowledge Retrieval. In *Proce. KnowLLM*. ACL, Bangkok, Thailand.
[22] Rishabh Mehrotra, James McInerney, Hugues Bouchard, et al. 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. In *Proc. CIKM* (Torino, Italy) *(CIKM '18)*. ACM, 9 pages.
[23] Sai Munikoti, Anurag Acharya, Sridevi Wagle, et al. 2024. Evaluating the Effectiveness of Retrieval-Augmented Large Language Models in Scientific Document Reasoning. In *Proc. SDP*. ACL, Bangkok, Thailand.
[24] Ziad Obermeyer, Brian Powers, Christine Vogeli, et al. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019).
[25] Harrie Oosterhuis. 2021. Computationally Efficient Optimization of Plackett-Luce Ranking Models for Relevance and Fairness. In *Proc. SIGIR* (Virtual Event, Canada) *(SIGIR '21)*. ACM, 10 pages.
[26] Berk B Ozmen and Piyush Mathur. 2025. Evidence-based artificial intelligence: Implementing retrieval-augmented generation models to enhance clinical decision support in plastic surgery. *JPRAS* 104 (2025).
[27] George Raftopoulos, Nikos Fazakis, Gregory Davrazos, et al. 2025. A Comprehensive Review and Benchmarking of Fairness-Aware Variants of Machine Learning Models. *Algorithms* 18, 7 (2025).
[28] Amifa Raj and Michael D. Ekstrand. 2022. Measuring Fairness in Ranked Results: An Analytical and Empirical Comparison. In *Proc. SIGIR* (Madrid, Spain) *(SIGIR '22)*. ACM, 11 pages.
[29] Amirhossein Razavi, Mina Soltangheis, Negar Arabzadeh, et al. 2025. Benchmarking Prompt Sensitivity in Large Language Models. In *Proc. ECIR* (Lucca, Italy). Springer-Verlag, Berlin, Heidelberg, 11 pages.
[30] Felix G. Rebitschek, Josef F. Krems, and Georg Jahn. 2016. The diversity effect in diagnostic reasoning. *Memory & Cognition* 44, 5 (01 July 2016).
[31] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation of BERT Rankers. In *Proc. SIGIR* (Virtual Event, Canada) *(SIGIR '21)*. ACM, 11 pages.
[32] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009).
[33] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proc. SIGKDD* (London, United Kingdom) *(KDD '18)*. ACM, 10 pages.
[34] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proc. FairWare* (Gothenburg, Sweden) *(FairWare '18)*. ACM, 7 pages.
[35] Yuan Wang, Zhiqiang Tao, and Yi Fang. 2022. A Meta-learning Approach to Fair Ranking. In *Proc. SIGIR* (Madrid, Spain) *(SIGIR '22)*. ACM, 6 pages.
[36] Xuyang Wu, Shuowei Li, Hsin-Tai Wu, et al. 2025. Does RAG Introduce Unfairness in LLMs? Evaluating Fairness in Retrieval-Augmented Generation Systems. In *Proc. COLING*. ACL, Abu Dhabi, UAE.
[37] Hamed Zamani and Michael Bendersky. 2024. Stochastic RAG: End-to-End Retrieval-Augmented Generation through Expected Utility Maximization. arXiv:2405.02816 [cs.CL]
[38] Xuan Zhao, Simone Fabbrizzi, Paula Reyero Lobo, et al. 2024. Adversarial Reweighting Guided by Wasserstein Distance to Achieve Demographic Parity. In *2024 IEEE International Conference on Big Data (BigData)*.