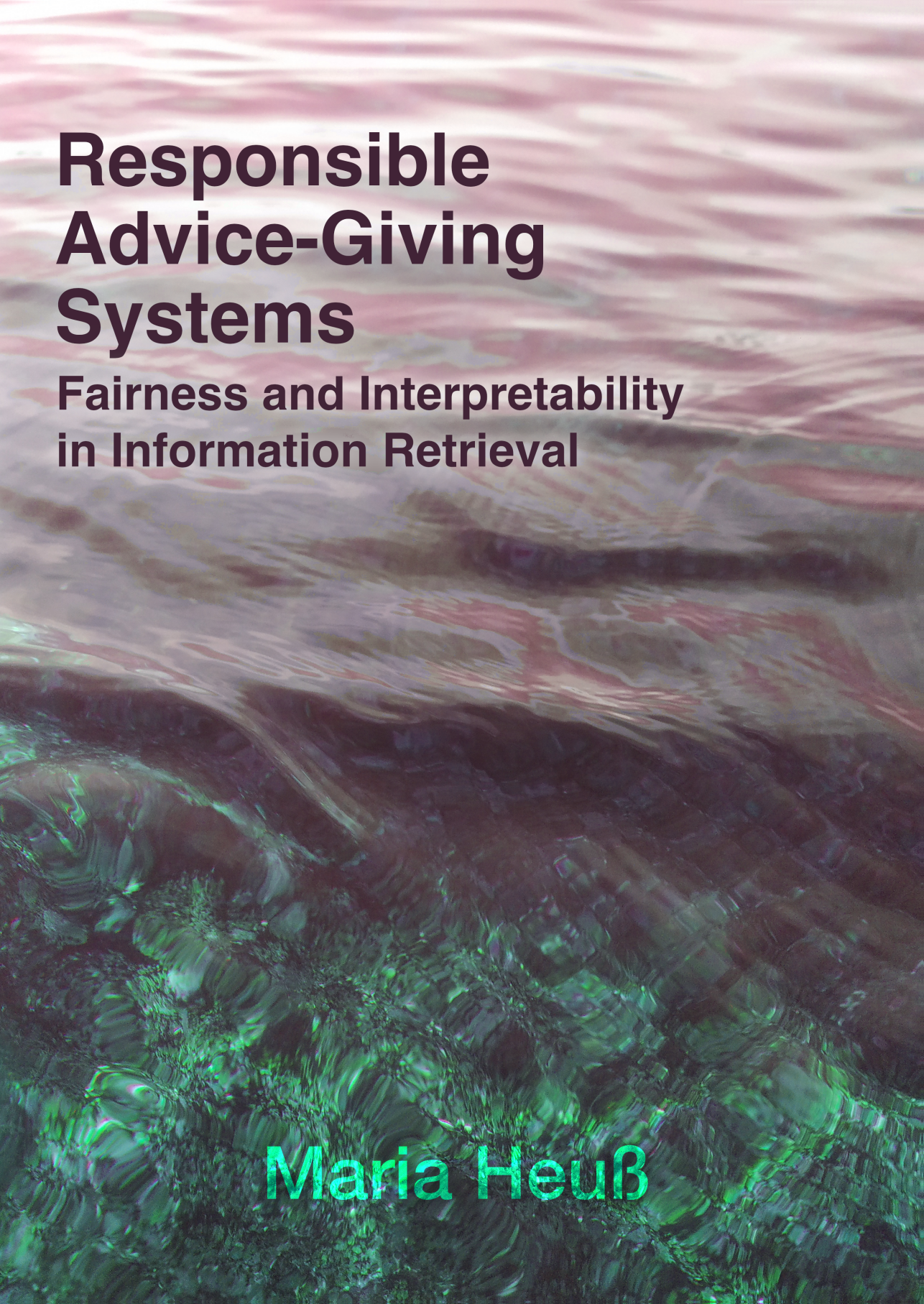

Responsible
Advice-Giving
Systems
Fairness and Interpretability
in Information Retrieval
Maria Heuß

Responsible Advice-Giving Systems

Fairness and Interpretability in Information Retrieval

Maria Heuss

Responsible Advice-Giving Systems

Fairness and Interpretability in Information Retrieval

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in
de Agnietenkapel
op maandag 27 oktober 2025, te 16:00 uur

door

Maria Clara Heuss

geboren te Nürnberg

Promotiecommissie

Promotor:	prof. dr. M. de Rijke	Universiteit van Amsterdam
Promotor:	dr. A. Anand	Technische Universiteit Delft
Overige leden:	prof. dr. C. Eickhoff	Eberhard Karls Universität Tübingen
	prof. dr. S. Ghebreab	Universiteit van Amsterdam
	dr. A. Lucic	Universiteit van Amsterdam
	prof. dr. S. Verberne	Universiteit Leiden
	dr. W.H. Zuidema	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The research was carried out at the Information Retrieval Lab at the University of Amsterdam, with support from the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, `https://hybrid-intelligence-centre.nl`, project nr. 024.004.022.

Cover by Maria Heuss and Mathijs Henquet
Printed by Proefschriftspecialist, Zaandam

ISBN: 978-94-93483-07-1

To Fatima, who remains with me as support, inspiration and role model.

Acknowledgments

I am not sure how long it had been my dream to become a researcher. Probably not since childhood, because what even is a researcher to a child? The desire to investigate deeply and discover new things about the world formed at a time when I did not envision myself in this position. In my mind, researchers were the smartest people on earth – they could speak intellectually (whatever that means) and they also had a certain look to them. During my studies, I almost did not dare to say that I was interested in pursuing research, always fearing I would encounter that skeptical individual who would point out the genius in class with higher grades, or who would ask me a question I couldn't answer. Fortunately, I now know better: good researchers neither need to look nor speak a certain way, nor do they necessarily need to be the smartest people. I was fortunate to meet a diverse set of people whom I all consider excellent researchers, each with different skill sets and perspectives. I can now proudly say that I am a researcher, yet I neither know everything nor am I the smartest person I know. Persistence, mathematical intuition, and the wish to change the world for the better have been great drivers of my academic career thus far. Nevertheless, I don't believe I would be where I am today without all the supportive people around me and the amazing role models I had the pleasure of getting to know. In what follows, I would like to thank some of my most important sources of support during these past five years, and before.

I cannot thank prof. dr. Maarten de Rijke enough, who has supported me in so many different ways during my PhD. Maarten, I cannot imagine being supervised better by any other person. I appreciate the never-ending support when it comes to accommodating personal circumstances, the high-level advice that always invited me to think further, and the constant availability to help with all kinds of academic challenges. I am proud to be a member of your large and colorful academic family and hope there will be many future opportunities for collaboration.

I also want to thank dr. Avishek Anand for the continued support, first as a mentor, then as my second supervisor. Your perspective really brought new inspiration to my work. All the brainstorming sessions, the invitation to randomly call you up for advice whenever I needed it, and our conversations about our shared experiences with our daughters, who were born on the exact same day, really brightened my time in the last 2-3 years. I hope that we can continue working together on the future of explainable IR.

I would also like to thank the members of my committee: Prof. dr. Carsten Eickhoff, prof. dr. Sennay Ghebreab, dr. Ana Lucic, prof. dr. Suzan Verberne, and dr. Jelle Zuidema. Thank you all for carefully reading my thesis, posing thoughtful questions, and for putting your trust in me by approving it. I will do my best to honor this approval in my future career as a researcher.

To my paranymphs Myrthe and Mariya: Thank you for being there throughout my PhD and for standing by my side during the defense. Thanks for the countless conversations, the support that I received from you, and the experiences that we shared together. Myrthe, thank you for brightening my day with fun stories about your life whenever we met up for coffee, at the playground, or for dinner, and for being such a great aunt to my kids. Mariya, thank you for accompanying me during all this time, for all your support and constant suggestions, and for all the fun dinners and sincere conversations that we had over the past years.

I want to thank all my colleagues from IRLab during those 5 years. Thank you Shashank for being my friend from the very beginning and for putting trust in me and allowing me to do the same. Thank you Philipp for becoming such a reliable friend and for offering help every time that things turn difficult. Thank you Sam for all the great chats about cooking and our lifes. Thank you Sami for being a friend, letting me stay in your place and for always being such fun to be around. Thank you Evangelos for putting trust in me and for always giving interesting input. Thank you Andrew and Mohammad for the perspectives that you have shared and the support that you provided when needed. Thank you Maarten for your support and for having me in the back of your mind when new opportunities arose. Thank you Ivana, Petra, Pablo and Kiki for having made life during my PhD easier and sometimes also more enjoyable and for always being there when help was needed. Thank you Svitlana and Ana for being such great examples of sharp and knowledgeable researchers that I aspire to be as well. Thanks to my FACT girls, Clara, Jasmin and Yuanna for the great discussions we had during our reading groups. Also, a special thanks to Yuanna – I always found your adventures to try out new things and your joyfulness very inspirational. Thank you Gabriel for being such a casual person to be around and for always allowing people to participate. Thank you Songgaojun for sharing your experience as an IRLab mom with me. Thank you Antonis for being the bright presence in our office and for distracting me from work from time to time. Thank you Kidist for all the kind words of support. Thank you Clemencia for inspiring me to sometimes set some boundaries. Thank you Roxana for the chats about interpretability but also about life. Thank you Vera for really doing your best to make me feel included during COVID. Thank you Gabrielle for always being there when support was needed. Thank you also to the other members of IRLab: Ali, Amin, Arezoo, Barrie, Chang, Chen, Chuan, Dan, Daniel, David, Dylan, Federico, Georgios, Harrie, Hongyu, Ilias, Ilya, Jiahuan, Jie, Jin, Jingfen, Jingwei, Julien, Justine, Lu, Maartje, Maryam, Maurits, Maxime, Ming, Mohanna, Mounia, Mozhdeh, Olivier, Panagiotis, Pooya, Romain, Ruben, Ruqing, Saedeh, Sebastian, Shaojie, Shubha, Simon, Sid, Siddharth, Teng, Thilina, Thong, Weijia, Xinyi, Yang, Yangjun, Yibin, Yifei, Yixing, Yongkang, Yougang, Yuanxing, Yubao, Yuyue, Zahra, Zhirui, Zihan, and Ziming – without you it would not have been the same.

I also want to thank the team of the Hybrid Intelligence Consortium for their extensive efforts in providing a place for us starting PhD students to connect to each other but also to connect to senior members of the team and to work on projects together. Frank and Rineke in particular, thank you for your leadership and advice.

I also want to say thank you to the senior researchers Suzan Verberne, Carsten Eickhoff, Claudia Hauff, Mounia Lalmas and Carlos Castillo who were there for advice at different stages of my PhD. Thank you Suzan, for being such a great role model for a successful woman in our field who manages to balance family with research. You have been such a reliable and valuable source of advice. Carsten, I am grateful for the weekly meetings that we had for some time, to just catch up and put our heads together. I am glad to have found a new shared interest in interpretability research that we can collaborate on further in the future. Claudia and Mounia, thank you for being the role models of female powerhouses that both of you are, both finding a balance between research and family, and for giving valuable advice that I am still remembering now. ChaTo, thank you for being there when I needed advice about my next career steps.

I also want to thank some of my closest collaborators: Daniel Cohen, Jonas Wallat, and Catherine Chen. Each of you has been a joy to work with. Dan, I really valued your mentorship and our collaboration on combining model uncertainty with fairness considerations. Jonas, we have been working together for just over a year and I really enjoy our close collaboration and brainstorming. We have so many project ideas waiting to be explored, and I am excited for all the different directions that our research can take us in the future. Catherine, you have been such a lovely and reliable collaborator and friend. You always have my back and I am looking forward to future collaborations.

Before I started my PhD, there were people during my school years and studies who contributed to my path that I want to thank now. I am grateful to my high school teachers Heinz Horwath and Herbert Bube. You sparked my interest in the mysteries of physics and, in the case of Herr Horwath, even got me thinking about some philosophical questions. I also want to thank Annette Huber-Klawitter. Thank you for supervising and supporting me with such patience during my mathematics studies. You were always an inspiration and one of my most important female role models.

I really had to learn the importance of balance to stay healthy and enthusiastic about my work. There were periods in my life when the balance was off and I poured too much of my energy into work alone, which made me deeply unhappy. This is why I value so much all the great people who helped me maintain a balance between the things that I find interesting and rewarding and the activities and people that I love. Some of them I want to thank explicitly here, but there are many more who accompanied me during periods of my life and whom I want to implicitly thank as well.

Thank you Felix, for leaving home and exploring the world (well at least a small part of it) with me. Some of the important decisions that led me here, we made together and you were an invaluable source of support during the difficult times in my Bachelor. Thank you Alexander, for being the first person who talked to me in class here in the Netherlands and for being my first Dutch friend. You introduced me to so many amazing people, many of whom I now call my friends, and I really value your support during difficult times during my Master. Thank you Louis for being a friend to rely on and to pour my heart out to. You have been listening to countless rants over the past few years, you have been there whenever I needed help and I know I can always rely on you. I am looking forward to having you closer to us soon again. Thank you Mo for biking me home drunk and remaining my friend ever since. You were one of the reasons I started to feel at home in the Netherlands and I am looking forward to many more years of growing up together.

I wish I could also thank Fatima, who always believed in me more than I did, and who will always remain an inspiration to me.

I am endlessly grateful to have found my girl gang – Isabel, Katharina, Vivien, and Lisa – during my studies and am so happy that we have managed to stay in touch even though we all live in different places now. I have so many happy, crazy memories, and I feel so grateful that I found you gals to terrorize the math and physics faculty with during our studies. Each of you has inspired me through your different flavors of intellectual curiosity, your ability to push through difficult times, and through the fun that you all managed to have along the way. I hope for many more fun stories and sincere moments shared in the future.

Thank you, Kata, for always being there when I needed someone to talk to and to

share difficult experiences with. Thank you, Gabor, for sticking with us over all these years, for sometimes being rational together and sometimes less so, and for sharing so many experiences and moments. Thank you, Mehdi, Maarten, Hanne, Vincent, Anne, Fenna, Katinka, Rene, Sara, Matteo, Fabian and Veerle for the many dinners and hangouts and fun moments together. Thank you, Elisa, for figuring out that one dish that works with both of our food preferences and for cooking it at least once a week together during our Erasmus semester. Thank you, Urja and Sharvaree, for being my HI sisters and for going from confused to confident together.

Finally, I want to thank my family for being there for me at all times along the way.

Thank you Rens for supporting us so well with Ida and now also with Muriël and for the financial and emotional support over the years.

I want to thank my siblings David and Eva, who never took me too seriously and humbled me when needed, while at the same time remaining a constant source of support and providing interesting perspectives. Thanks for being there for me, providing me with your knowledge and expertise, and for making me consider things from a more social scientific point of view.

Thanks to both of my parents for being a constant source of support, for allowing us to move in during COVID, and for supporting us so well with our kids. There has always been great understanding of meetings that needed to happen during shared holidays or for sleeping in after a long night of working to meet a deadline, and I am so glad that my way of making my life-work balance work has been met with so much understanding and support. Liebe Mama und lieber Papa, ich weiss noch dass ich früher dachte, dass ihr alles wisst. Obwohl ich inzwischen vermute dass das nicht immer stimmt, kann ich mich immer noch zu jeder Zeit und an jedem Ort auf euren Rat und eure Hilfe verlassen. Vielen Dank für die Gelassenheit, all die gemeinsame Zeit und die Unterstützung, die ich von euch über die Jahre hinweg bekommen habe. Und danke Mama, dass du Wege gefunden hast, mich zu fördern und immer dazu bereit warst, neue Rätsel mit mir auszuprobieren.

I also want to thank my children Ida and Muriël for allowing me to follow two dreams at the same time: to be a mother and a researcher. Even though you cannot read this yet and Ida, you are having a hard time understanding what kind of doctor I will be who doesn't check up on people's health, thank you for being the biggest source of joy and love in my life and for inspiring me with observations and associations that adults just do not seem to be able to make anymore.

Lastly, I want to thank my most important source of support during the last 10 years of my life. Mathijs, thank you so much for always being there with me. Thank you for being my cheerleader and at times my IT support. Thank you for being a true partner that I can rely on and who has my back when work pulls me in. Thank you for giving me the stability at home that allows me to take some risks at work. Thank you for sharing so many interests with me, both in terms of scientific progress and things outside of science that we have explored together. Thank you for understanding my jokes and for giving me confidence in my abilities.

Maria Heuss
Amsterdam
September 2025

Contents

1 Introduction

Artificial Intelligence (AI) has been reshaping our world, affecting our daily routines, work lives, and social interactions [51, 65]. It is particularly transforming how we interact with information [86]. Whereas a few decades ago, people relied primarily on carefully curated newspapers, books, television, or university curricula, the emergence of information access systems like search tools and content recommendation platforms has both expanded individual possibilities and responsibilities in acquiring information [61]. While older generations might still remember spending hours in libraries searching for specific books or papers, meticulously copying pages to take home, today we can find answers to most questions with a simple keyword search.

Responsible information access. This shift toward instant information access has fundamentally altered our relationship with information [71]. The traditional requirement of investing a significant amount of time to answer even straightforward questions has been replaced by the ability to uncover vast amounts of information with just a vague search query. However, this convenience comes with new challenges.

While instant information access offers tremendous opportunities, it may also introduce significant drawbacks [188]. The responsibility for ensuring certain information quality standards has shifted from users to automated systems, raising critical questions about responsible deployment. Users now face reduced agency when presented with curated information, making it difficult to verify or control what they receive. Additionally, the lack of "friction" in information seeking may reduce active user participation, potentially compromising their ability to understand and question the process [188], and making them more vulnerable to misinformation.

These and similar concerns have given rise to the field of responsible AI [8], which addresses challenges at the interface between technology and society such as fair representation of population groups, accountability for algorithmic mistakes, and transparency in decision-making processes. This thesis examines these dimensions specifically within information access and advice-giving systems.

Advice-giving systems and their applications. With the increased ease of receiving information, the possibility of using information systems as direct advice-giving tools, rather than as research tools might seem tempting. Applications now range from ranking job candidates based on algorithmic fit assessments [23, 122, 130] to retrieving medical documents for diagnosis support [2, 203, 230], and increasingly include chat-based systems that provide natural language responses to user queries [3, 240].

The recent rise of large language models (LLMs) has accelerated this trend toward chat-based information systems, driven by their ability to generate plausible and fluent answers to almost any conceivable question. While traditional access to medical, legal, financial or mental health advice often involves significant costs and barriers [48, 78, 170], LLM-based agents offer affordable and convenient alternatives. Unlike previous static systems that output predefined rankings or summaries, chat-based systems enable interactive information exploration through follow-up questions, simple language, and real-time translation.

Risks and challenges of automated advice-giving. However, these developments underscore the importance of carefully considering associated risks. Data biases embedded in models can lead to unfair outcomes that disadvantage certain groups, as demonstrated by Amazon's hiring tool, which was scrapped in 2014 after exhibiting bias against women [50]. Representation issues, such as under-representation of women in occupation-related searches, can shift public perceptions about real-world distributions [109]. Technical aspects of systems, such as position bias in rankings and LLM information processing, can further amplify these societal biases [193, 234].

These challenges highlight the necessity of thorough model evaluation and providing users with tools to assess the reliability of model-generated advice. The field of explainable AI addresses these concerns by offering insights into model decision processes. However, increasing model complexity makes this both more challenging and more crucial. Explanations take various forms, from feature importance measures [171] that show input influence on outputs, to mechanistic interpretations [18] that identify key model components, each serving different purposes and user needs.

Overall, we should strive to create advice-giving systems that not only fulfill users' information needs and decision support requirements in most cases but that do so in a responsible way. The term "responsible" here encompasses many different aspects, including the previously mentioned explainability and fair-/unbiasedness, as well as diverse representation of viewpoints and information, accountability for mistakes, robustness against malicious attacks, privacy considerations, and more, as visualized in Figure 1.1.

Topics covered in this thesis. This thesis examines various components of modern advice-giving systems through the lens of responsible development, focusing particularly on fairness and model explainability/interpretability. Chapters 2 and 3 address the first pillar of Figure 1.1, the fairness of ranking systems, which are often used as intermediate or final components of advice-giving systems. These systems either provide users directly with ranked lists or pass them to final components such as user interfaces or LLM modules. Here, our research questions center around assumptions that are commonly made withing fair ranking frameworks and what happens when they do not hold.

Chapter 4 and 5 are concerned with the second pillar, the explainability of different system components, with Chapter 4 focusing on the explainability of ranking systems and Chapter 5 centering around the use of citations as explanations for RAG systems, which can be used as a final component and interface to the user. A more detailed overview of the topics covered in each chapter can be found in Section 1.1.

We anticipate that improvements in individual components will contribute to a better

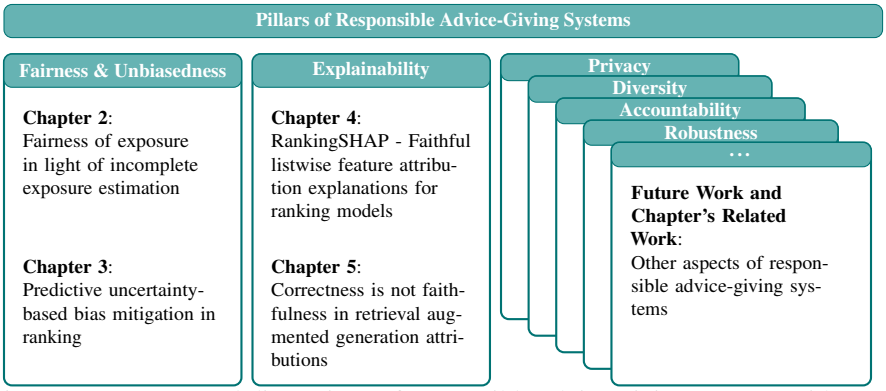


Figure 1.1: Conceptual dimensions of responsible advice-giving systems with cross-references to relevant thesis chapters.

understanding of how responsible systems should be designed and what form they should take overall.

1.1 Research Outline and Questions

Throughout this thesis, we investigate different aspects of responsible advice-giving. We will focus on different underlying mechanisms of information processing as well as different dimensions of responsibility. In particular, we will start by focusing on the fairness of different rankers (a feature-based ranker in Chapter 2 and a text-based ranker in Chapter 3). We will then move on to the explainability of algorithms underlying modern advice-giving systems, focusing on explaining ranking models in Chapter 4 and on RAG-based advice-giving through grounded generation in Chapter 5. Below, we describe the main research questions for each of those chapters.

1.1.1 Fairness in ranking systems

In the first part of this thesis, our focus is on fairness aspects of responsible advice-giving. There are different notions of fairness, depending on the task as well as different underlying assumptions about the bias contained in the data and its societal implications [149]. In the first technical chapter we consider fairness of exposure, which aims to provide each document group, or individual document with a fair share of user exposure, which is considered a finite resource that needs to be distributed among different documents presented in a ranked list. In past works, the underlying assumption of approaches to fair exposure is that the higher a document is ranked, the more exposure (or in other words user attention) it will be able to obtain, increasing its chances to be read, bought, considered for further investigation, etc. Yet, as this assumption does not always hold true, since for example, outliers in a ranked list can draw more exposure than would be usually assumed, and since these effects are not yet sufficiently researched to create good user models for these cases, our first research question aims to investigate how we can still provide a fair ranking policy even if the

exposure distribution for some possible ranking constellations is unknown. In particular, we are interested in situations where the exposure distribution over the documents is different from the usually assumed exposure-based exposure distribution, due to inter-document relationships, e.g., the case of outliers, where the contrast in presentation impacts the exposure of the several documents in the ranked list. We formulate our first research question as follows:

RQ A Can we define an exposure-fair ranking policy in situations where the expected exposure distribution is unknown for some rankings?

To answer RQ A, for the case where the exposure distribution is unknown due to inter-document relationships, we formalize the task of fair ranking under incomplete exposure estimation and design an algorithm that can generate a fair ranking policy, while avoiding to present the user with ranked lists of which we can not reliably estimate the exposure distribution. For this purpose we generalize a complex optimization approach to fair ranking of top-k rankings and define a re-sampling based approach that iteratively removes rankings with unknown exposure from the ranking policy without changing the position-based exposure that each document gets.

In our next chapter, we challenge another assumption frequently made when designing fair IR approaches, which is the assumption that the (relative) relevance of documents to the user can always be accurately predicted through some ranking model. In practice, any kind of prediction holds a certain level of uncertainty, in our case this would be uncertainty about the relative order of the documents in terms of actual relevance to the user. This uncertainty can be modeled [79], giving us an idea of where the model was certain about the relative order of the documents and for which documents it might be more likely to be wrong. We investigate whether we can make use of such uncertainty predictions, to reduce the exposure of documents that contain biases with lesser cost to user utility, leading us to the second research question:

RQ B Can we use the predictive uncertainty of the model prediction to improve ranking fairness?

To answer RQ B, we start by approximating the predictive model uncertainty of a text-based ranker through Laplace approximation. We use this uncertainty to design a simple approach that takes the ranking scores and adjusts them relative to their predicted uncertainty, depending on whether for the sake of fairness (or in this case to reduce the bias in the ranked list) we need to increase or decrease the exposure of said documents. We show experimentally that this simple and efficient approach beats all baselines, even the ones that are much more costly in terms of computation, opening up interesting questions for future predictive uncertainty-based fair ranking research.

1.1.2 Explaining advice-giving processes

The second part of this thesis tackles another challenge in responsible advice-giving systems, the task of explaining model decisions. Explanations can serve different purposes, dependent on the recipient of the explanation, the explained model process, and the task at hand. In our fourth chapter, we investigate feature attribution explanations

for listwise rankers. Feature attribution aims to explain the model decision by highlighting the importance of each input feature for the final model prediction [171]. While feature attribution for classification or regression tasks (i.e., pointwise tasks) has been thoroughly discussed in the past literature, explaining listwise ranking decisions, i.e., the order of documents rather than individual ranking scores, has received less attention. Hence, we formulate our third research question as follows:

RQ C How can we generate listwise ranking explanations for listwise ranking models?

To answer RQ C, we formally define listwise feature attribution. Since lots of different aspects of a ranking decision can be investigated, to get a complete image of the model decision, we introduce the "listwise explanation objective" that specifies which aspect of the ranking decision to explain, and provide some examples of such objectives and their use. One of the most frequently used pointwise feature attribution approaches called SHAP [136], which is based on the game theoretic concept of Shapley values [189], has not been defined for use in a listwise manner. We extend this approach by proposing RankingSHAP, as a direct instantiation of listwise feature attribution. Since the evaluation of listwise explanations has not been well established, we introduce two novel evaluation paradigms, one based on evaluation with a white box model and one based on the deletion and preservation check from the broader AI explainability literature. We show that the proposed RankingSHAP method performs competitively with other explanation frameworks.

In our fifth and last chapter we move away from explaining the process of ranking documents and towards explaining retrieval augmented generation (RAG) [126]. RAG can be seen as an additional step on top of the common retrieval and ranking pipeline that enables interactive question answering through the use of a large language model as interface between the retrieved information and the user. As a form of explanation for the source of the information that is presented in the generated answer, a citation can be generated which refers back to a certain document. Past work has mostly evaluated the citation's correctness, also called *answer faithfulness*, which aims to measure how well the information in the answer aligns with the information in the cited document [256]. We argue that this is not sufficient for ensuring responsible advice-giving, particularly in high-stakes domains where the consequences of misleading information can be severe. Rather than just looking at token matching or other correctness metrics, we should ensure that the cited information was actually used in the answer generation process. This distinction becomes critical when considering that citations can paradoxically increase user trust even when they are misleading, a phenomenon that is especially concerning in complex domains where users may lack the expertise to verify the relationship between citations and answers. Simply having a citation that contains correct information is insufficient if that information was not genuinely influential in generating the response, as this creates a false sense of transparency and accountability. We formulate our forth research question as follows:

RQ D Do RAG citations faithfully reflect the source of the information used in the answer generation process?

To answer RQ D, we define desiderata for good attribution and introduce the concept of *citation faithfulness*: in addition to the agreement in provided information, it requires

the document to be used during the answer generation process. We introduce the phenomenon of *post rationalization*, where the model first generates an answer without using a certain source document but still cites the document since it is required to do so through training or prompt. We show empirically that post-rationalization is a common phenomenon within a state-of-the-art RAG model that is being trained for attributed generation (i.e., including citations).

1.2 Main Contributions

In this section, we summarize the main contributions of the thesis.

1.2.1 Conceptual contributions

- We introduce the task of fairness of exposure under incomplete exposure estimation (Chapter 2).
- We introduce the notion of predictive uncertainty-based fair ranking (Chapter 3).
- We introduce citation faithfulness as opposed to citation correctness as a property of attributed generation (Chapter 5).
- We propose desiderata for citations that go beyond correctness and accuracy and are needed for trustworthy RAG systems (Chapter 5).

1.2.2 Algorithmic contributions

- We develop FELIX, a re-shuffling based algorithm that provides fair ranking policies while avoiding rankings with unknown exposure distributions (Chapter 2).
- We extend constraint optimization approaches for fairness of exposure to top-k ranking tasks, broadening the applicability of fair ranking methods (Chapter 2).
- We develop an efficient algorithm for the generalized Birkhoff-von Neumann decomposition, which is used in the constraint optimization approach to fairness (see the previous point), that achieves $\mathcal{O}(k^3n^2)$ complexity rather than $\mathcal{O}(n^4\sqrt{n})$, where k is the size of the ranked list and n is the total document count (Chapter 2).
- We apply Laplace approximation as a post-hoc method to approximate the predictive uncertainty of a ranking model (Chapter 3).
- We develop PUFR, a re-ranking approach that leverages the model uncertainty on the predicted ranking score distribution to produce less biased rankings (Chapter 3).
- We develop RankingSHAP, a Shapley-value based algorithm for listwise feature attribution that provides flexible investigation of ranking decisions through customizable explanation objectives (Chapter 4).

1.2.3 Theoretical contributions

- We prove the generalized Birkhoff-von Neumann theorem for non-square matrices, establishing the existence of solutions for top-k fairness optimization problems (Chapter 2).
- We prove that the time complexity of the proposed efficient implementation of the Birkhoff-von Neumann decomposition mentioned above is $\mathcal{O}(k^3n^2)$ (Chapter 2).
- We provide a rigorous mathematical formulation of listwise feature attribution for ranking models (Chapter 4).
- We define citation faithfulness, extending existing frameworks for citation correctness evaluation (Chapter 5).

1.2.4 Empirical contributions

- We evaluate our method FELIX for fair ranking under incomplete exposure estimation in a setup with unknown exposure distribution due to outliers in the presented ranked lists on two datasets of the TREC Fair Ranking track and compare with a recently introduced approach and several fair ranking approaches that do not consider incomplete exposure (Chapter 2).
- We evaluate our approach PUFR to several in- and post-processing bias mitigation approaches and show that it outperforms all baselines while being computationally more efficient. (Chapter 3).
- We establish multiple evaluation schemes for listwise feature attribution and conduct a comparative analysis of attribution methods on learning-to-rank models (Chapter 4)
- We provide empirical evidence of post-rationalized citations in a state-of-the-art RAG model, highlighting limitations in current attribution approaches (Chapter 5).

1.3 Thesis Overview

In this thesis we approach responsible advice-giving systems from several angles.

In Chapter 2, we focus on improving the fairness of ranking systems in scenarios where for some ranked lists the distribution of user exposure is unknown. We introduce a method that avoids such ranked lists and hence allows for a more accurate approximation of a fair ranking policy.

In Chapter 3, we take a slightly different perspective on fairness by looking at bias mitigation in ranked documents. We investigate how the predictive uncertainty of a ranking model about the order of the ranked documents can be used to improve fairness and mitigate biases without impacting the ranking performance too much. We introduce a simple but effective approach that utilizes uncertainty estimates of the predicted

relevance, outperforming all baseline approaches that use static predicted relevance scores in terms of both computational efficiency and on the utility-fairness frontier.

In Chapter 4 we move towards the field of explainable IR (XIR), in particular the task of explaining ranking models. We rigorously define listwise feature attribution and develop an algorithm that approximates listwise attribution values. We also provide two evaluation frameworks that are novel to the field of XIR that aim to test the faithfulness of the feature attribution explanations determined in this way.

In Chapter 5, we stay within the field of XIR, but move away from ranking models towards interpreting RAG models. We define desiderata for responsible attributed generation, in other words, properties that a good citation should have. We introduce the concept of citation faithfulness, building on insights from explainable AI by regarding citations as explanations of the generated answer. We also provide proof of unfaithful citation behavior through an experiment showing the existence of post-rationalization within RAG citations.

Each of these chapters is based on a single research paper (as described in Section 1.4) and can be read independently. Since each chapter approaches responsible advice-giving systems from a different angle, there is no background knowledge from any section that is required to follow any other chapter. Notation is kept consistent with the corresponding publications, leading to slight differences in notation between the chapters of this thesis. We highlight the most notable differences in notation between chapters at the beginning of each chapter.

Finally, Chapter 6 summarizes the findings of this thesis and provides a perspective on limitations and future work within the field of responsible advice-giving systems.

1.4 Origins

Below we list the publications that each of the chapters is based on.

Chapter 2 is based on M. Heuss, F. Sarvi, and M. de Rijke. Fairness of exposure in light of incomplete exposure estimation. In *SIGIR 2022: 45th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 759–769. ACM, July 2022.

- MH Conceptualization, Investigation, Validation, Software, Methodology, Writing - original draft, Writing - review & editing, Project administration.
- FS Software, Data curation, Writing - review & editing.
- MdR Supervision, Writing - review & editing.

Chapter 3 is based on M. Heuss, D. Cohen, M. Mansoury, M. de Rijke, and C. Eickhoff. Predictive uncertainty-based bias mitigation in ranking. In *CIKM 2023: 32nd ACM International Conference on Information and Knowledge Management*, pages 762–772. ACM, October 2023.

- MH Conceptualization, Investigation, Validation, Software, Methodology, Writing - original draft, Writing - review & editing, Project administration.
- DC Conceptualization, Software (uncertainty prediction), Methodology , Writing - original draft, Writing - review & editing.

MM Writing - review & editing.

MdR Supervision, Writing - review & editing.

CE Supervision, Conceptualization, Writing - review & editing.

Chapter 4 is based on M. Heuss, M. de Rijke, and A. Anand. RankingSHAP–Listwise feature attribution explanations for ranking models. In *SIGIR 2025: 48th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 381–391. ACM, July 2025.

MH Conceptualization, Investigation, Validation, Software, Methodology, Writing - original draft, Writing - review & editing, Project administration.

MdR Supervision, Writing - review & editing.

AA Supervision, Writing - review & editing.

Chapter 5 is based on J. Wallat, M. Heuss, M. de Rijke, and A. Anand. Correctness is not faithfulness in RAG attributions. In *ICTIR 2025: The 15th International Conference on the Theory of Information Retrieval*. ACM, July 2025.

MH and JW shared first authorship.

JW Conceptualization of the desiderata of attributed generation, Investigation and Software of the initial experiments, Methodology, Writing - original draft, Writing - review & editing, Visualization.

MH Conceptualization of citation faithfulness, Investigation and Software of the ablation study, Methodology, Writing - original draft, Writing - review & editing, Project administration.

MdR Supervision, Writing - review & editing.

AA Supervision, Conceptualization, Writing - review & editing.

The thesis also benefited from work on the following publications:

- A. Vardasbi, G. Bénédict, S. Gupta, M. Heuss, P. Khandel, M. Li, and F. Sarvi. The University of Amsterdam at the TREC 2021 fair ranking track. In *TREC*, 2021.
- F. Sarvi, M. Heuss, M. Aliannejadi, S. Schelter, and M. de Rijke. Understanding and mitigating the effect of outliers in fair ranking. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 861–869, February 2022.
- C. Rus, J. Kareem, C. Xu, Y. Liu, Z. Deng, and M. Heuss. AMS42 at the NTCIR-18 FairWeb-2 task. *Proceedings of NTCIR-18*, May 2025.
- M. de Rijke, B. van den Hurk, F. Salim, A. Al Khourdajie, N. Bai, R. Calzone, D. Curran, G. Demil, L. Frew, N. Gießing, M. K. Gupta, M. Heuss, S. Hobeichi, D. Huard, J. Kang, A. Lucic, T. Mallick, S. Nath, A. Okem, B. Pernici, T. Rajapakse, H. Saleem, H. Scells, N. Schneider, D. Spina, Y. Tian, E. Totin, A. Trotman, R. Valavandan, D. Workneh, and Y. Xie. Information retrieval for climate impact: Report on the MANILA24 workshop. *SIGIR Forum*, 59(2), June 2025.

- M. Heuss, C. Chen, A. Anand, C. Eickhoff, and S. Verberne. Workshop on explainability in information retrieval. In *SIGIR 2025: 48th international ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, July 2025.
- A. Dotsinski, U. Thakur, M. Ivanov, M. H. Khan, and M. Heuss. On the generalizability of "Competition of mechanisms: Tracing how language models handle facts and counterfactuals". *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.
- T. Wiegman, L. Perotti, V. Pravdová, O. Brand, and M. Heuss. Reproducibility study of: "Competition of mechanisms: Tracing how language models handle facts and counterfactuals". *Transactions on Machine Learning Research*, 2025.
- D. Campregher, Y. Chen, S. Hoffman, and M. Heuss. Tracing facts or just copies? A critical investigation of the competitions of mechanisms in large language models. *Transactions on Machine Learning Research*, 2025.

Part I

Fairness in Ranking Systems

2

Fairness of Exposure in Light of Incomplete Exposure Estimation

As part of the first half of this thesis, which examines fairness in ranking systems, this chapter focuses on exposure fairness in scenarios where the assumption of accurate exposure estimation does not hold. Exposure fairness operates on the principle that user attention, which can be thought of as a finite resource collected by documents across different sessions, should be distributed equitably. Typically, fair exposure allocation is determined based on merit, such as predicted user utility.

We investigate whether we can guarantee fairness when, due to inter-document relationships that impact how the user looks at the list of documents as a whole, the distribution of user exposure can not be accurately estimated for a certain type of ranked list. As an example for such inter-document relationships consider visual outliers, that attract more user attention than would usually be the case for a document placed at the same position. Including such ranked lists in the selection of lists that are presented to the users would prevent us from making any meaningful guarantees for exposure fairness. With this we address the following research question:

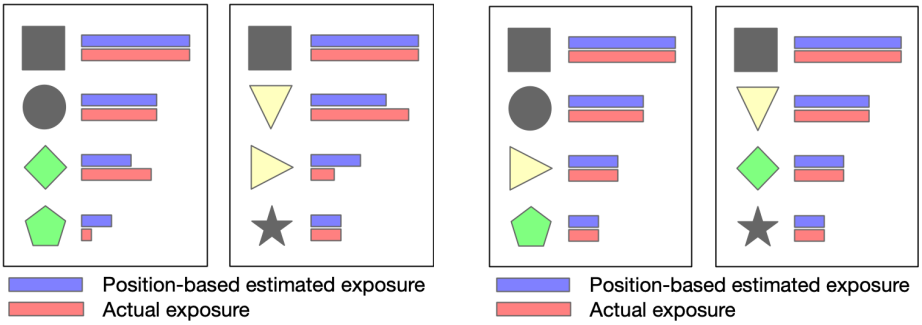
RQ A Can we define an exposure-fair ranking policy in situations where the expected exposure distribution is unknown for some rankings?

Throughout this chapter, we use the term "item" rather than "document" to align with the specific domain of our investigation, though the principles apply broadly to document ranking systems.

2.1 Introduction

There has been increased interest in fair ranking systems, as witnessed by the number of publications [62, 249], the topic's attention during keynotes at leading conferences [31, 103], and challenges such as the TREC Fair Ranking track [63]. Several particularities about rankings make this task especially challenging.

This chapter was published as M. Heuss, F. Sarvi, and M. de Rijke. Fairness of exposure in light of incomplete exposure estimation. In *SIGIR 2022: 45th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 759–769. ACM, July 2022.



(a) (b)

Figure 2.1: Rankings with unknown exposure distribution which are due to inter-item dependencies between items marked by the same color and similar shapes (a). By shuffling some items between rankings in the stochastic ranking policy these dependencies can be reduced such that the estimated exposure agrees with the actual exposure that each item gets (b).

First, often ranking systems act as a tool for two-sided marketplaces, such as job markets [82] or music recommender systems [145]. On one side, users want relevant item recommendations. On the other side, items or their providers are interested in being exposed to as many users as possible. Second, biases like position bias can cause a traditional deterministic ranking to amplify small differences in predicted scores into vast differences in user attention [20, 193].

An important line of research on fairness in ranking deals with *fairness of exposure*. Given a ranking, we can estimate how much exposure each item gets in expectation during inference. We call this the *exposure distribution* of the ranking. Singh and Joachims [193] define several notions of fairness of exposure for rankings, among them *disparate treatment*. This notion defines a stochastic ranking policy to be fair if each item or item-group gets expected exposure proportional to its merit. We will mostly focus on individual fairness, where we want to provide each individual item with exposure relative to its merit.

Incomplete exposure estimation. Previous methods for fairness of exposure assume that we can estimate the exposure distribution of any ranking in the set of all possible rankings. For this, a user model like the position-based model [20, 193, 223, 237], or the ERR-based model [56] can be used. However, there are cases where, due to inter-item dependencies that are not accounted for by any of the existing user models, for certain rankings, user-behaviour does not follow the user model; for such rankings we cannot estimate the exposure distribution accurately. See Figure 2.1a for an illustration. E.g., Sarvi et al. [185] show that visual outliers can have a great impact on the exposure distribution within a ranking, since such outliers attract more user attention. This phenomenon is an example of inter-item dependencies where one item can be perceived as an outlier in the context of items it is presented together with. It can cause the exposure distribution to diverge from the distribution assumed by the user model.

Simply ignoring the incomplete knowledge about the exposure of some of the

rankings would imply that we cannot guarantee fairness. Also, by ignoring potentially incomplete exposure estimation, we might introduce a new kind of bias into the collected click data, since items that got more exposure than estimated will have propensity values that are too high, leading to overestimation of their relevance. One solution would be to obtain a more accurate user browsing model by estimating the exposure distribution of rankings that do not follow the user model, through a large-scale user study. To the best of our knowledge no such studies have been conducted. It is also not clear whether one can always reliably estimate the exposure distribution for all possible rankings.

Instead, we propose to avoid showing rankings with unknown exposure distribution to the user by reducing their weight in the probability distribution of the stochastic ranking policy.

Fair top-k ranking. So far, the literature on fairness of exposure has mostly focused on full-length rankings. Top-k rankings are well studied in the general information retrieval (IR) literature [37, 55, 243, 248]; many real-world ranking applications require us to expose just a short list of items. Often there are more relevant items than can be shown to the user, hence it is important to consider fairness of exposure for this set-up as well. Although there have been few approaches to *fair top-k ranking* [248, 250], most are concerned with demographic parity, rather than merit-based fairness of exposure.

Our contributions. In this chapter we develop a method to find ranking policies that avoid presenting rankings with unknown exposure distribution, while still optimizing for user utility and fairness. Under the assumption that inter-item dependencies are the reason for the shift in exposure, our method works by shuffling items between different rankings to avoid presenting them in a context where they disturb the position-based exposure distribution, as illustrated in Figure 2.1b.

We also present what we believe to be the first approach towards fairness of exposure in the top-k setting for the convex optimization approach towards fairness. We generalize the Birkhoff-von Neumann theorem and use this to extend [193] to the top-k setting.

To summarize, our main contributions in this chapter are as follows:

- We introduce the task of fairness of exposure in light of incomplete exposure estimation and define a novel method FELIX that provides us with a fair ranking policy that avoids rankings with unknown exposure distribution.
- To make FELIX applicable to a broader range of use cases, we extend the constrained optimization approach to fairness of exposure to the top-k case.
- We test and compare FELIX on the outlier use case introduced in [185] and show big improvements over other top-k fair ranking methods in terms of effectiveness in avoiding rankings containing outliers, while staying within the fairness constraints.

2.2 Related Work

Fairness in ranking. For a detailed overview of fair ranking we refer to [62, 249]. Yang and Stoyanovich [238] seem to have been the first to formalize fairness for rankings in a rank-aware manner, by calculating parity for different top-k cut-offs

and summing over these values with a rank-based discount. Zehlike et al. [248, 250] discuss representational fairness for top-k rankings and define a re-ranking algorithm that ensures a share of items from the protected groups in every prefix of the top-k, while Celis et al. [32] formulate the problem as a constrained optimization problem. These papers look for a deterministic ranker, not a stochastic ranking policy, and emphasize on representational fairness and demographic parity.

Singh and Joachims [192] introduce the notion of expected exposure and define fairness of exposure with respect to demographic parity and equal opportunity, where the expected exposure is calculated w.r.t. position bias. Later work [193] defines different types of fairness of exposure w.r.t. disparate impact and disparate treatment, and address the task as a constrained optimization problem. Biega et al. [20] define equity of attention as an alternative notion of fairness for rankings that is also based on exposure; they also address the task as a constrained optimization problem. Wang and Joachims [223] also consider fairness of exposure combined with diversity in rankings. We build on [193] and use the non-uniqueness property of the Birkhoff-von Neumann decomposition that is also used in [223] to produce more diverse rankings. Importantly, we reduce the probability that the user is shown a ranking with unknown exposure distribution rather than providing the user with more diverse rankings as in [223].

Another line of research aims to include fairness in the learning process by including a fairness objective in the objective function [56, 194, 217, 247]. Since inter-item relationships are hard to model within the in-processing set-up, in our work we focus on a post-processing method for avoiding rankings with unknown exposure distribution and leave work on in-processing methods for the future.

Another work that looks into the the topic of uncertainty within fair ranking is [195], which explores fairness of exposure when there is uncertainty about the merit. In contrast to this work, we are considering uncertainty about the exposure of certain rankings.

Exposure estimation in ranking. In counterfactual learning to rank (CLTR) true estimation of exposure plays a central role [104]. Early work on CLTR corrects for position bias using exposure, estimated by a click model [45], as the propensity to inversely weight the importance of clicks [104, 225]. More recent work focuses on estimating examination probabilities [6, 10, 69, 214, 216, 226], which also correlates with exposure, correcting for more types of bias. Recent work on learning fair rankings from implicit feedback [237] simultaneously corrects for position bias and implicit biases in the data. There is no prior work on how to adapt these models for the case where certain rankings do not follow the general user model.

Prior work has shown that exposure might be impacted by other factors than just position and the relevance of other items. Yue et al. [246] observe that visual attractiveness can impact the exposure that items get; Sapiezynski et al. [184] acknowledge that the attention that users give to items in a ranking depends on context; and Wang et al. [224] address the impact of click bait items on exposure distribution. Sarvi et al. [185] show that the existence of visual outliers in rankings can skew the exposure distribution amongst the items, causing outliers to draw more attention than estimated by the position-based user model that non-outlier rankings seem to follow.

In this chapter, we focus on similar but more general use cases, where due to inter-

item relationships the exposure distribution for some rankings differs from the generally assumed distribution, that can be described through existing user models.

2.3 Background

We introduce preliminaries in fair ranking that form the basis for a new method for ranking under fairness constraints, while avoiding to present rankings with unknown exposure distribution.

2.3.1 Stochastic ranking policies

Depending on the definition of fairness being used, often a single deterministic ranking cannot achieve fairness [20, 56]. Instead, probabilistic rankers can be used to provide a fair distribution of exposure among items. Given a query q and set of candidate items, $\mathcal{D}_q = \{d_i\}_{i=1,\dots,n}$, to be ranked, we define a *stochastic ranking policy* π_q as a probability distribution over all possible rankings $\mathcal{R}_{\mathcal{D}_q}$. That is, π_q assigns each ranking $\sigma_j \in \mathcal{R}_{\mathcal{D}_q}$ a probability $\pi_q(\sigma_j)$ that it will be shown to the user.

To evaluate the fairness of a ranking policy we determine the *expected exposure* $\epsilon(d_i \mid \pi_q)$ that each item d_i obtains when enough rankings have been presented to users. To compute this, we need to assume a browsing model that explains the probability of a user visiting an item. Diaz et al. [56] adopt user models corresponding to the ranked-based precision (RBP) and expected-reciprocal rank (ERR), while Singh and Joachims [193] use the position-based user model (PBM). We follow the latter, as it is commonly used in the fairness literature [20, 193, 223, 237]. Assuming that the exposure of an item in a ranking, $\epsilon(d_i \mid \sigma)$, is purely based on its position, the *expected exposure* $\epsilon(d_i \mid \pi_q)$ of document d_i for policy π_q can be calculated as:

$$\begin{aligned}\epsilon(d_i \mid \pi_q) &= \mathbb{E}_{\sigma\sim\pi_q}\epsilon(d_i \mid \sigma) \\ &= \sum_{\sigma\in\mathcal{R}_{D_q}} \pi_q(\sigma)\cdot\epsilon(d_i \mid \sigma) \\ &= \sum_{\sigma\in\mathcal{R}_{D_q}} \pi_q(\sigma)\cdot\frac{1}{\log(1+\text{rank}(d_i \mid \sigma))},\end{aligned} \tag{2.1}$$

where we assume that the exposure can be calculated based on the rank: $\epsilon(d_i \mid \sigma) = v(rank(d_i \mid \sigma))$ with exposure at rank j given by $v(j) = \frac{1}{\log(1+j)}$.

2.3.2 Fairness of exposure

The definition of what constitutes a fair ranking may vary between application scenarios and types of biases being addressed [249]. We focus on individual fairness, but our approach can easily be extended for group fairness. Our goal is to make sure that similar items receive a similar amount of exposure that is proportional to their merit. The *merit* $u(d \mid q)$ of an item, $d \in \mathcal{D}$, indicates how much exposure it deserves to get from users with respect to query q. We define the merit of an item as its relevance to the query.

The idea of *fairness of exposure* [193] is to provide each item with exposure ϵ that is proportional to its merit:

$$\frac{\epsilon(d_i \mid \pi_q)}{u(d_i \mid q)} = \frac{\epsilon(d_j \mid \pi_q)}{u(d_j \mid q)} \quad \forall d_i, d_j \in \mathcal{D}. \tag{2.2}$$

2.3.3 Finding a stochastic policy under fairness constraints

To be able to satisfy certain fairness constraints, we need to find a stochastic ranking policy (Section 2.3.1). Singh and Joachims [193] approach the problem by optimizing for user utility under fairness constraints via linear programming. As our method is based on theirs, we introduce it in more detail. For each query q and item $d \in \mathcal{D}$, let $u(d \mid q)$ be its relevance to the user. We define the *utility* U of a ranking policy π_q as the expected utility to the user, when shown a ranking sampled from π_q:

$$\begin{aligned} U(\pi_q) &= \sum_{d \in \mathcal{D}} \epsilon(d \mid \pi_q) \cdot u(d \mid q) \\ &= \mathbb{E}_{\sigma \sim \pi_q} \sum_{d \in \mathcal{D}} \epsilon(d \mid \sigma) \cdot u(d \mid q). \end{aligned} \tag{2.3}$$

As we assume a position-based user model, $\epsilon(d \mid \sigma)$ is purely dependent on the position of d in the ranking. Therefore, the expected utility U can be calculated based on the probabilities $P_{i,j} = P(d_i \text{ is placed at rank } j)$:

$$\begin{aligned} U(\pi_q) &= \sum_{d_i \in \mathcal{D}} \sum_{j \in \{1,\dots,n\}} P_{i.j} \cdot v(j) \cdot u(d_i \mid q) \\ &= \mathbf{u}^T \mathbf{P} \mathbf{v}, \end{aligned} \tag{2.4}$$

where $n = |\mathcal{D}|$ is the number of items in the ranking, $\mathbf{u}$ the vector containing the merit of each item, $\mathbf{v}$ the vector containing the position bias at each position, and $\mathbf{P} = \{P_{i,j}\}_{i,j=1,\dots,n}$. Singh and Joachims [193] show that the disparate treatment constraint from Eq. (2.2) can be formulated as a linear constraint in $\mathbf{P}$, which yields a convex optimization problem of the form:

$$\begin{aligned} \mathbf{P} = \operatorname{argmax}_{\mathbf{P}} \ & \mathbf{u}^T \mathbf{P} \mathbf{v} \\ \text{such that } & \mathbb{1}^T \mathbf{P} = \mathbb{1} \\ & \mathbf{P}\mathbb{1} = \mathbb{1} \\ & 0 \leq P_{i,j} \leq 1 \\ & \mathbf{P} \text{ is fair.} \end{aligned} \tag{2.5}$$

A solution $\mathbf{P}$ to this optimization problem is a doubly stochastic matrix, called the *marginal rank probability* (MRP) *matrix*. The solution $\mathbf{P}$ needs to be transformed into an executable stochastic ranking policy. The Birkhoff-von Neumann theorem [22] gives us a constructive proof that such a matrix can be decomposed into a convex sum of $M \leq n^2 - n + 1$ permutation matrices:

$$\mathbf{P} = \sum_{m=1,\dots,M} \alpha_m P_{\sigma_m} \text{ such that } \sum_{m=1,\dots,M} \alpha_m = 1 \, (0 \leq \alpha_m \leq 1). \tag{2.6}$$

Since each permutation matrix corresponds to some ranking, we denote the permutation matrix corresponding to σ by P_σ.

With this we have found a stochastic policy π with $\pi(\sigma_m) = \alpha_m$ and $\pi(\sigma) = 0$ for all σ not contained in this convex sum. Note that this decomposition is not necessarily unique; in Section 2.4.3 below we will make use of this fact.

2.3.4 The impact of outliers on the exposure in rankings

Sarvi et al. [185] provide evidence that commonly made assumptions on the user-behaviour might not hold when the presented ranking contains visible outliers that might attract the attention of the user. Since outliers are an example where inter-item dependencies between documents can change the exposure distribution among the items in a ranked list, we work with this example for our experiments in Section 2.5. We follow the set-up of [185], where the authors assume that outliers can be determined through outlier detection on a specific visual item feature $g(d)$ that might impact the user's perception of an item. In the case of scholarly search, which is used as an example in the experiments, such a feature could be the number of citations that each document has.

Outliers are considered in a context $C \subset \mathcal{D}$ of items that are presented together, which could for instance be the top-k that is presented in a single search engine result page (SERP). Given such a context $C = \{d_1, \dots, d_k\} \subset \mathcal{D}$, we use the features, $g(d_1), \dots, g(d_k)$, as input for the outlier detection. Sarvi et al. [185] find that the performance of their method for removing outliers from the rankings is not very sensitive to the outlier detection method. For simplicity, we will therefore use the Z-score:

$$z(g_i) = \frac{g_i - \mu}{s}, \tag{2.7}$$

where $g_i = g(d_i)$, and $\mu = \frac{1}{k}\sum_{i=1}^{k} g_i$ and $s = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(g_i - \mu)^2}$ denote the mean and standard deviation of the scores in that context. Given these Z-scores, we define an item d_i to be an outlier if $|z(g_i)| > \lambda$, where λ can be chosen dependent on the sensitivity towards outlier items. Here, we diverge slightly from [185], who use a more complex outlier detection method.

Next, we introduce an extension to the convex optimization approach to fairness of exposure from Section 2.3.3 for top-k rankings. We use the definition of fairness of exposure with respect to disparate treatment from Section 2.3.2 and work with stochastic policies from Section 2.3.1. We also develop a method that avoids displaying rankings with unknown exposure distribution, using the outlier use case from Section 2.3.4 for our experiments in Section 2.5.

2.4 Fairness of Exposure under Incomplete Exposure Estimation

As discussed in Section 2.3.1, previous work on fair ranking assumes that we can estimate the exposure distribution for all rankings in a policy with one user model. Often,

the position-based user model is used. But there are cases where these assumptions do not hold up. Sarvi et al. [185] show that the existence of outliers in a displayed ranking can strongly impact the exposure distribution of the ranking. To the best of our knowledge, there is no prior work on estimating the exposure distribution of such rankings. If such *rankings with unknown exposure distribution* are part of a stochastic ranking policy (i.e., if such a ranking has a non-zero probability of being presented to the user), we cannot determine whether the policy is fair. Therefore, for attaining fair stochastic policies we should avoid using such rankings. This introduces the task of fair ranking under incomplete exposure estimation.

In this section we develop a method for the task of **F**airness of **E**xposure in **L**ight of **I**ncomplete e**X**posure estimation, FELIX, that provides a ranking policy that avoids rankings with unknown exposure distribution without damaging fairness or utility. FELIX is based on the assumption that the shift in the exposure distribution is caused by inter-item relationships between the items that are ranked together. Hence, depending on the context an item is presented in, it could either follow the position-based exposure distribution or it could draw more or less exposure than assumed. In the example, an outlier in a ranking might draw more attention than a non-outlier item at the same position, as demonstrated in [185]. When presented in a more diverse ranking, the same item might not be considered an outlier any more and follow the assumed position-based exposure distribution. Compared to the method for removing outliers from the top-k in [185], FELIX is more generally applicable to any use case where, due to inter-item dependencies, some rankings have unknown exposure distribution. Also, FELIX allows us to consider outliers in the local context that they are presented in, while Sarvi et al.'s approach can only remove outliers with respect to the global context of all items in the list.

Since the context in which items are presented in plays a central role for our task, naturally we are interested in our method to work in the top-k setting. Therefore, we first generalize the constrained optimization approach towards fairness of exposure, introduced in [193], to the top-k setting and present an efficient way to determine a fair policy. Then we present our method FELIX that uses iterative re-sampling to determine a stochastic policy that avoids presenting rankings with unknown exposure distribution to the user, while staying within the fairness constraints.

2.4.1 Fair ranking in the top-k setting

We will now extend the convex optimization approach to fairness to the top-k setting. Let n be the number of candidate items to be ranked and $k \leq n$ be the number of ranks of the desired rankings. As explained in Section 2.3.3, searching for a stochastic policy under fairness constraints can be done by first searching for a marginal rank probability matrix $\mathbf{P}$ that satisfies the fairness constraints, and then decomposing this matrix. Since we are interested in the top-k case, $\mathbf{P} = \{P_{i,j}\}_{i=1,\dots n, j=1,\dots k}$ is now a $n \times k$ matrix, where $P_{i,j}$ is the probability that item i is placed at rank j. With $\mathbf{u}$ the n-dimensional utility vector and $\mathbf{v}$ the k-dimensional vector containing the examination probability at

each of the top-k positions we can solve the following linear program:

$$\begin{aligned}\mathbf{P} = \operatorname{argmax}_{\mathbf{P}}\ & \mathbf{u}^T\mathbf{P}\mathbf{v} \\ \text{such that } & \mathbb{1}_n^T\mathbf{P} = \mathbb{1}_k \\ & \mathbf{P}\mathbb{1}_k \leq \mathbb{1}_n \text{ (element-wise inequality)} \\ & 0 \leq P_{i,j} \leq 1 \\ & \mathbf{P} \text{ is fair.}\end{aligned} \tag{2.8}$$

Given the marginal rank probability matrix $\mathbf{P}$, we want to determine a stochastic policy given by a distribution over actual rankings. In the $n \times n$ setting, the Birkhoff-von Neumann (BvN) decomposition provides us with an algorithm to determine such a distribution. The following result generalizes the BvN theorem to the $n \times k$ setting where n is not necessarily equal to k.

Theorem 2.4.1. *Any matrix $P = \{a_{i,j}\}_{i\leq n, j\leq k}$ with $\forall i,j : 0 \leq a_{i,j} \leq 1$, $\forall j : \sum_{i=1}^{n} a_{i,j} = 1$ and $\forall i : \sum_{j=1}^{k} a_{i,j} \leq 1$ can be written as the convex sum $P = \sum_{l=1}^{m} \alpha_l \cdot P_l$ of permutation matrices P_l with coefficients $\alpha_l \in [0,1]$ such that $\sum_{l=1}^{m} \alpha_l = 1$.*

Proof. In Lemma 2.4.2 below, we show that P can be extended to a doubly stochastic matrix P'. We can use the BvN decomposition for doubly stochastic matrices to find a decomposition for P', which will induce a decomposition for P. For a more detailed proof, see the Section 2.9.1. □

Here we say that $P' \in \mathbb{R}^{n'\times k'}$ is an *extension* of $P \in \mathbb{R}^{n\times k}$ if $n' \geq n, k' \geq k$, and $P_{i,j} = P'_{i,j}$ for all (i,j) with $i \leq n$ and $j \leq k$. We will denote this by $P'|_{i\leq n, j\leq k} = P$.

Lemma 2.4.2. *Let $P = \{a_{i,j}\}_{i\leq n, j\leq k}$ be a matrix with the same properties as described in Theorem 2.4.1 with $k \leq n$. Then there is a matrix $P' = \{a'_{i,j}\}_{i\leq n, j\leq n}$ with $\forall i,j : 0 \leq a'_{i,j} \leq 1$ such that $P = P'|_{i\leq n, j\leq k}$, and $\forall i : \sum_{j=1}^{n} a'_{i,j} = 1$ and $\forall j : \sum_{i=1}^{n} a'_{i,j} = 1$.*

Proof. Define $P' = \{a'_{i,j}\}_{i\leq n, j\leq n}$ as

$$a'_{i,j} = \begin{cases} a_{i,j} & \text{if } j \leq k \\ \frac{1-\sum_{j'=1}^{k} a_{i,j'}}{n-k} & \text{if } j > k. \end{cases} \tag{2.9}$$

Then $P'|_{i\leq n, j\leq k} = P$ by definition. P' satisfies all the requirements from the lemma. A proof of this can be found in Section 2.9.1. □

By transposing P we can show that the Lemma also holds if $k > n$.

2.4.2 An efficient implementation of the generalized Birkhoff-von Neumann decomposition

For an implementation of the generalized Birkhoff-von Neumann theorem, one can in theory use the proof of Theorem 2.4.1 and extend the MRP-matrix, that we obtained by solving the convex optimization problem from Eq. 2.8, to a full $n \times n$-matrix. This

Algorithm 1 Algorithm for the generalized Birkhoff-von Neumann decomposition.

Require: $P \in \text{Mat}_{n\times k}$ with properties as in Theorem 2.4.1

1: Initialize $\mathcal{P} = \{\}$ empty decomposition
2: Extend P to $\widetilde{P}$ by adding a column $\{c_i\}_{i=1,\dots,n}$ with values $c_i = 1 - \sum_{j=1}^{k} P_{i,j}$
3: **while** $\widetilde{P} \neq 0$ **do**
4: Translate $\widetilde{P}$ to a bipartite graph with n resp. $k+1$ vertices on each side with edges between the i-th and j-th vertex if $P_{i,j} \neq 0$
5: Find a perfect matching m (with multiplicity of $n-k$ for the last vertex) with the adjusted Hopcroft-Karp algorithm
6: Translate m to a matrix P^m, where $P^m|_{i\leq n, j\leq k}$ forms a permutation matrix.
7: $\alpha = \min_{\{i,j|P^m_{i,j}\neq 0\}}(\widetilde{P}_{i,j})$
8: $\mathcal{P} \leftarrow \mathcal{P} + (\alpha, P^m|_{i\leq n, j\leq k})$
9: $\widetilde{P} \leftarrow \widetilde{P} - \alpha P^m$
10: **end while**
11: Return $\mathcal{P}$

matrix can then be decomposed into the convex sum of permutation matrices with help of the BvN theorem for doubly stochastic matrices after which we can restrict the matrices again to the first k columns. Since the complexity of the BvN decomposition for square matrices is $\mathcal{O}(n^4\sqrt{n})$ [98, 105] and hence infeasible for large n, we propose an alternative implementation for $n \times k$ or $k \times n$ matrices with $k < n$, that can be implemented with time complexity $\mathcal{O}(k^3n^2)$.

Algorithm 1 gives a structured overview of our algorithm for the generalized BvN decomposition. We start off by noting that the way in which we extended the doubly stochastic matrix from P in the proof of Lemma 2.4.2 is not unique. For any index pair $(i,j),(i',j')$ with $j,j' > k$ we can subtract some value β from $a'_{i,j}$ and $a'_{i',j'}$, while adding the same value to $a'_{i',j}$ and $a'_{i,j'}$. The resulting matrix will have the same properties as P' and will also be an extension of P. Therefore, instead of extending P to a full doubly stochastic matrix, we can extend it to an $n \times (k+1)$ matrix $\widetilde{P}$, where the last column contains the entries that make the values of each row sum to 1. In the decomposition we split off matrices that are permutation matrices on the first k columns and have $n-k$ non-zero entries on the last column; see line 2 in Algorithm 1.

We can use this realization to extend the implementation of the BvN algorithm [21], which translates the marginal rank probability matrix into a bipartite graph and uses the Hopcroft-Karp algorithm [98] to find a perfect matching m, which in turn can be translated back into a permutation matrix, P^m; see line 4, 5 and 6.[1]

In the next step, line 7, we calculate the biggest coefficient α, such that subtracting the scaled permutation matrix αP^m, still results in a matrix with only non-negative coefficients. We add the coefficient-matrix pair to the decomposition and subtract the scaled permutation matrix from $\widetilde{P}$; see line 8 and 9. By translating the matrix $\widetilde{P}$ into a bipartite graph, where the node corresponding to the $(k+1)$-th column has multiplicity $n-k$, and adjusting the Hopcroft-Karp algorithm (line 5) slightly to allow for certain

[1] For the implementation we used `https://networkx.org` and `https://github.com/jfinkels/birkhoff`

vertices to be matched with higher multiplicity, we can significantly speed up this part of the algorithm from $n^2\sqrt{n}$ to k^2n. Since the upper bound of matrices in the decomposition decreases from order n^2 to kn the complexity changes as stated in the following Theorem. A proof of this statement can be found in Section 2.9.2.

Theorem 2.4.3. *Using the modified top-k algorithm for the generalized Birkhoff-von Neumann theorem, Algorithm 1, a decomposition as described in Theorem 2.4.1 can be obtained with time complexity $\mathcal{O}(k^3n^2)$.*

2.4.3 Determining a stochastic policy that avoids rankings with unknown exposure distribution

As explained in Section 2.3.4, certain types of rankings can have a non-typical exposure distribution. Allowing such rankings invalidates the approach by Singh and Joachims [193], since a position-based exposure vector $\mathbf{v}$ is used in both the utility calculation and the fairness constraint in their approach. In this section our goal is to find a stochastic policy that avoids rankings for which the exposure distribution is unknown. We will use a re-sampling strategy, which, after the decomposition step in Eq. 2.6, rejects rankings with unknown exposure distribution. The core idea we present below is based on the assumption that the inter-item dependencies between some of the items is the cause of the shift in exposure and that by shuffling the items between different rankings, rankings with unknown exposure distribution might be changed into rankings with known exposure distribution.

Algorithm 2 gives a step-by-step overview of the algorithm used by FELIX. Similarly to Wang and Joachims [223], we make use of the fact that the Birkhoff-von Neumann decomposition is not unique. For most doubly stochastic matrices there is a large number of possible decompositions [60], which makes it possible for us to search for a decomposition that does not have a lot of weight on rankings with unknown exposure distribution. After determining the MRP matrix $\mathbf{P}$ (line 1), we decompose it into the sum $\mathbf{P} = \sum_{i=1}^{M} \alpha_i P_{\sigma_i}$. In the top-$k$ setting this can be done by using the generalized Birkhoff-von Neumann algorithm (Algorithm 1); see Algorithm 2 line 4. We write $\mathcal{P} = \{(\alpha_i, P_{\sigma_i})\}_{i=1,\dots,M}$ for the set of coefficient, matrix pairs in this convex sum. Once the matrix is fully decomposed, we divide the resulting coefficient, permutation matrix pairs (α_i, P_{σ_i}) into two groups, one containing all the permutations where the corresponding ranking has a known exposure distribution amongst its items and the other one containing pairs corresponding to rankings with unknown exposure distribution:

$$\begin{aligned} \mathcal{P}_{known} &= \{(\alpha_i, P_{\sigma_i}) \in \mathcal{P} | \sigma_i \text{ has known exposure distribution}\} \\ \mathcal{P}_{unknown} &= \mathcal{P} - \mathcal{P}_{known}. \end{aligned}$$

We use the elements of $\mathcal{P}_{known}$ directly as a part of the final decomposition; see lines 5–7. The elements of $\mathcal{P}_{unknown}$ are aggregated, weighted by their coefficient; see line 8.

$$\widetilde{\mathbf{P}} = \sum_{(\alpha_i, P_i) \in \mathcal{P}_{unknown}} \alpha_i \cdot P_i. \tag{2.10}$$

Algorithm 2 Fairness of Exposure in Light of Incomplete Exposure Estimation (FELIX)

Require: $\mathcal{D}_q$, k, merit vector $\mathbf{u}$, position bias vector $\mathbf{v}$, number of iterations $iter$

1: Determine MRP matrix $\mathbf{P}$ as in Eq. 2.8 with $\mathbf{u}$ and $\mathbf{v}$
2: Initialize $\pi(\sigma) = 0, \forall \sigma \in \mathcal{R}_{\mathcal{D}}$
3: **while** $iter \neq 0$ **do**
4: $\quad \mathcal{P} \leftarrow$ Decompose $\mathbf{P}$ with Algorithm 1
5: $\quad$ **for all** $(\alpha, P_\sigma) \in \mathcal{P}_{known}$ **do**
6: $\quad\quad \pi(\sigma) \leftarrow \pi(\sigma) + \alpha$
7: $\quad$ **end for**
8: $\quad \mathbf{P} \leftarrow \sum_{(\alpha, P_\sigma) \in \mathcal{P}_{unknown}} \alpha \cdot P_\sigma$
9: $\quad iter \leftarrow iter - 1$
10: **end while**
11: **for all** $(\alpha, P_\sigma) \in \mathcal{P}_{unknown}$ **do**
12: $\quad \pi(\sigma) \leftarrow \pi(\sigma) + \alpha$
13: **end for**
14: Return π

Up to scalar multiplication, the resulting matrix $\widetilde{\mathbf{P}}$ satisfies the required characteristics of Theorem 2.4.1 and hence can be decomposed again with the generalized BvN decomposition (Algorithm 1).

This decomposition-aggregation process repeats for a number of iterations, $iter$ (line 3–10). In each iteration, the recombination of rankings with unknown exposure distribution makes it possible for the algorithm to group items together that previously have not been together in one ranking. Through this re-sampling, the context in which items are presented changes, which often also means that the exposure distribution of these newly ranked list is known. Note that this approach does not remove items from the rankings, but rather shuffles the items among different rankings within the decomposition. After $iter$ iterations the remaining rankings with unknown exposure distribution are being added to the policy (line 11–13) to ensure the fairness and utility, that was optimized for.

2.4.4 Upshot

To summarize Section 2.4, we extended the continuous optimization approach to fairness for the top-k setting in Section 2.4.1 by proving that the Birkhoff-von Neumann theorem, which is used to decompose the matrix that was attained through the convex optimization, can be extended to a more general setting. In Section 2.4.2 we gave an algorithm for the decomposition in the top-k case and discussed an efficient implementation. This extends the space of use cases to which this approach to fair ranking can be applied. We will use this in our experiments, which will partly be conducted in the top-k setting. In Section 2.4.3 FELIX is introduced, which, by iteratively rejecting rankings with unknown exposure distribution, reduces the probability that such rankings are shown to the user.

Next, we test the performance of the proposed method for top-k fairness. Furthermore, we investigate how well FELIX is able to avoid rankings with unknown exposure

Table 2.1: Descriptive statistics of the original and pre-processed TREC Fair Ranking track 2019 and 2020 data.

	2019		2020	
	Train	Test	Train	Test
Avg. list size (original)	4.1	4.1	23.5	23.4
Avg. list size (pre-proc.)	4.1	13.0	23.5	31.9
Avg. # rel. items/list (original)	2.0	2.0	3.7	3.4
Avg. # rel. items/list (pre-proc.)	2.0	4.4	3.7	4.5

distribution and how this impacts the performance w.r.t. fairness and user utility.

2.5 Experimental Set-up

We experiment with two variants of our model: to evaluate our top-k approach to fair ranking we use FELIX without re-sampling i.e., with only one iteration, denoted by **FELIX**$_{iter=1}$; to evaluate our method for reducing the probability of generating rankings with unknown exposure we use 20 iterations (**FELIX**$_{iter=20}$).

Our experiments aim to answer the following research questions: **RQ2.1** Can **FELIX**$_{iter=1}$ provide fair top-k rankings while maintaining the user utility compared to the baselines? **RQ2.2** Can **FELIX**$_{iter=20}$ reduce the probability of showing rankings with unknown exposure distribution to the user without compromising fairness or utility, compared to other methods? We use the case of rankings with outliers as an example for rankings with unknown exposure distribution. As Sarvi et al. [185] show, outliers can change the exposure distribution that items collect in expectation; we broadly follow their experimental set-up to be able to compare to prior work that is, for this specific use case, closest to our approach.

Datasets. Our experiments in Section 2.6 use two academic search datasets provided by the TREC19 and TREC20 Fair Ranking track.[2] These datasets come with queries, relevance judgements, and information about the authors and academic articles extracted from the Semantic Scholar Open Corpus.[3] See Table 2.1 for descriptive statistics of the datasets. Since we experiment on the task of removing outliers from the top-k, which only makes sense for queries with enough items, for testing we only use rankings with at least 20 items. The 2020 dataset comes with 200 queries for training and 200 for testing; keeping only the lists with at least 20 papers leaves us with 112 test queries. Similarly, the 2019 dataset comes with 631 queries for training and 631 for testing. However the test set contains only 3 queries with more than 20 items, which is not acceptable. As a pragmatic solution, we keep lists with at least 10 items, which leaves us with 69 test queries, but up-sample each of these queries to 50 items by using the feature vectors of non-relevant items from other random lists as negative samples.

Experiments. We consider approaches where correcting for fairness is a post-processing

[2] https://fair-trec.github.io/

[3] http://api.semanticscholar.org/corpus/

step. We use ListNet [30] as our learning to rank (LTR) model for the ranking step, with a maximum of 30 epochs, the Adam optimizer with learning rate of 0.02, and early stopping. As input to the LTR model we use the same data as OMIT[4] with 25 features based on term frequencies, BM25 [174], and language models [209, 254].[5]

To be able to treat the output of the LTR model as the relevance probabilities we normalize the predicted scores to be within the range $[\epsilon, 1]$ with $\epsilon = 10^{-4}$. Choosing $\epsilon > 0$ ensures that each item has a non-zero probability of being placed in a ranking.

As mentioned earlier in this section, we use rankings that contain visible outliers as example for rankings with unknown exposure distribution. Following [185] we use the number of citations of a paper as a visible feature that may be subject to outliers. For the context in which outliers are perceived we use the top-k items. We use the Z-score with threshold value 2.5 to determine whether an item can be considered an outlier; see Section 2.3.4.

We conduct two types of experiments. The first experiment imitates the experimental set-up of Sarvi et al. [185], where full rankings are formed but the presence of outliers is only measured in the top-k of each ranking. The second experiment looks at top-k ranking. We use $k = 10$ in our experiments and aim for individual fairness as opposed to [185, 193], where group fairness is used.

Baselines. To answer research questions RQ2.1 and RQ2.2, we compare $\textbf{FELIX}_{iter=1}$ and $\textbf{FELIX}_{iter=20}$ with the following baselines:

PL As suggested in [56], we use a Plackett-Luce (PL) ranker initialized with the predicted, normalized scores of the LTR model.

PL-random We use a PL ranker over a uniform score distribution as a baseline for a random ranker.

Vanilla We use the method introduced by Singh and Joachims [193] with only fairness constraints as the vanilla baseline. This is the model we build upon.

Deterministic This baseline is ListNet, our traditional LTR model.

OMIT The method was introduced in [185], where a similar optimization problem is solved as for Vanilla, but with an additional regularizing objective that punishes rankings with a global outlier in the top-k.

For the experiments on the top-k, we only sample $k = 10$ items from the PL models, **PL@10** and **PL-random@10**. Since $\textbf{FELIX}_{iter=1}$ is a novel extension of the Vanilla convex optimization approach for the top-k setting, we do not have the Vanilla baseline in this setting. For OMIT we use our top-k convex optimization approach with the additional outlier objective, **OMIT@10**, to be able to compare the outlier reduction of $\text{FELIX}_{iter=20}$ and OMIT in the top-k setting.

Evaluation. To evaluate fairness we use the EE-L metric [56]. The target exposure of item d_i is calculated as $\epsilon^*(d_i) = \epsilon_{total} \cdot u(d_i) / \sum_j u(d_j)$, where ϵ_{total} is the total amount of exposure that users spend in expectation on the ranking, and $u(d_i)$ is the

[4] `https://github.com/arezooSarvi/OMIT_Fair_ranking`

[5] Our experimental code is based on `https://github.com/MilkaLichtblau/BA_Laura`.

merit, i.e. relevance, of item d_i. Given the expected exposure of all items as a vector ϵ, the *expected exposure loss*, *EE-L* can be calculated as:

$$\text{EE-L} = \ell\left(\epsilon, \epsilon^*\right) = \left\|\epsilon - \epsilon^*\right\|_2^2. \tag{2.11}$$

Ranking utility performance is measured with NDCG.

For a given query, to evaluate how well a policy π performs in avoiding rankings with unknown exposure distribution, we measure the probability that such a ranking is displayed by the policy. In our experiments this translates to measuring the probability that a randomly sampled ranking, σ contains an outlier:

$$\begin{aligned} P(u \mid \pi) &= P(\sigma \text{ has unknown exposure distribution} \mid \sigma \sim \pi) \\ &=_{\text{here}} P(\#\text{ outliers in } \sigma \geq 1 \mid \sigma \sim \pi). \end{aligned}$$

Additionally, for comparability with [185], we measure:

$$\text{Outlierness}\,@k(\pi) = \mathbb{E}_{\sigma\sim\pi} \sum_{d_i \in \text{top-}k(\sigma)} \mathbb{1}(d_i \text{ is outlier}) z(d_i).$$

For each metric we report the average value taken over all queries. Each experiment was conducted 5 times with different train/validation split and different random seed. Each split uses 80% of the train-data for training and 20% of the train-data for validation. In our result tables we report the mean results. We test for significance with a two tailed paired students t-test, using the metric values over all queries as input and comparing each method with FELIX$_{iter=20}$.

2.6 Results

Table 2.2 and 2.3 contain the results for our experiments on the top-k and full ranking set-up, respectively.

RQ2.1: Can FELIX$_{iter=1}$ provide fair top-k rankings while maintaining the user utility compared to the baselines? To answer this research question we first compare the performance of FELIX$_{iter=1}$ with PL@10, since this is the only baseline that has as its objective to create fair top-k ranking policies. For both utility and fairness FELIX$_{iter=1}$ performs marginally better on TREC20 data. In the case of TREC19 data, FELIX$_{iter=1}$ still has slightly better user utility; the fairness scores are close to identical. Overall none of these differences are significant.

As a sanity check, looking at our other baselines, we see that w.r.t. user utility (NDCG), in Table 2.2 the deterministic ranker outperforms all probabilistic rankers, which is expected since it is purely optimized for utility. This is reflected in the fairness score, where the deterministic ranker scores significantly worse than FELIX$_{iter=20}$. W.r.t. utility, the random ranker is outperformed by all other probabilistic ranking methods, showing that these methods present users with better results than a uniform ranking policy would.

To summarize, we find no significant differences in terms of utility or fairness between FELIX$_{iter=1}$ on the one hand and the PL-ranker on the one hand. This makes

Table 2.2: Top-k rankings. Significance is measured with a two-tailed paired t-test; all comparisons are against FELIX$_{iter=20}$.

	Method	Optimizing Fairness	NDCG↑ @5	NDCG↑ @10	Fairness↓ EE-L	$P(u \mid \pi)$↓ @10	Outlierness↓ @10
TREC20	**FELIX**$_{iter=20}$	Yes	0.203	0.279	**6.22**	**0.20**	**0.115**
	FELIX$_{iter=1}$	Yes	0.203	0.279	6.23	0.39*	0.151*
	PL@10	Yes	0.197	0.275	6.24	0.47*	0.174*
	PL-random@10	No	0.177*	0.249*	6.29	0.47*	0.175*
	Deterministic	No	**0.287***	**0.370***	7.22*	0.41*	0.154*
	OMIT@10	Yes	0.198	0.273	6.34	0.33*	0.132*
TREC19	**FELIX**$_{iter=20}$	Yes	0.12	0.16	5.9	**0.12**	**0.08**
	FELIX$_{iter=1}$	Yes	0.12	0.16	5.9	0.30*	0.12*
	PL@10	Yes	0.11	0.16	**5.8**	0.35*	0.14*
	PL-random@10	No	0.10	0.15	**5.8**	0.41*	0.16*
	Deterministic	No	**0.15**	**0.21***	7.5*	0.25*	0.12*
	OMIT@10	Yes	0.11	0.15	6.0	0.23*	0.10

our approach suitable for top-k ranking under fairness constraints and hence allows us to extend FELIX for this setting. In the rest of this section, we will see other advantages of FELIX over the PL baseline.

RQ2.2: Can FELIX$_{iter=20}$ reduce the probability of showing rankings with unknown exposure distribution to the user, without having to compromise fairness or utility, compared to other methods? We are interested in the trade-offs between user utility, fairness and the probability of showing rankings with unknown exposure, which is indicated by $P(u \mid \pi)$, in Tables 2.2 and 2.3. For the TREC20 data, in both settings FELIX$_{iter=20}$ successfully improves $P(u \mid \pi)$ while maintaining the NDCG@10 and EE-L scores compared to all baselines. Our main baseline to compare with for this research question is OMIT, as it is the only model that optimizes for presenting fewer outliers in the top-k positions. Compared to OMIT, FELIX$_{iter=20}$ achieves significantly better results in terms of $P(u \mid \pi)$ for both settings, while keeping the same (or better) scores for other metrics. For the top-k experiment, we also see a significant improvement w.r.t. $P(u \mid \pi)$, compared to FELIX$_{iter=1}$: iteratively re-sampling successfully reduces the number of rankings with unknown exposure distribution in the policy. For the TREC19 data we can still observe that FELIX$_{iter=20}$ offers the best trade-off between the three objectives in the top-k setting. However, the improvements w.r.t. the outlier removal are less significant in the full length experiments. Since for this dataset we used an up-sampling strategy that adds varying negative samples, the variation within these experiments is much higher, which makes the results less reliable and causes the observed differences to be less significant. Still, since the results broadly agree with the results for the more reliable TREC20 dataset, we take this as confirmation for the conclusions drawn there.

We also report the Outlierness metric, as introduced in [185], to show that the

Table 2.3: Full length rankings, remove outliers from the top-k. Significance is reported in the same way as in Table 2.2.

	Method	Optimizing Fairness	NDCG↑ @5	NDCG↑ @10	Fairness↓ EE-L	$P(u \mid \pi)$↓ @10	Outlierness↓ @10
TREC20	**FELIX**$_{iter=20}$	Yes	0.221	0.302	**24.5**	**0.24**	**0.126**
	Vanilla	Yes	0.221	0.302	**24.5**	0.40*	0.163*
	PL	Yes	0.192*	0.269*	24.7	0.45*	0.169*
	PL-random	No	0.178*	0.249*	24.9	0.47*	0.175*
	Deterministic	No	**0.267**	**0.348**	24.7	0.40*	0.152
	OMIT	Yes	0.221	0.302	**24.5**	0.34*	0.139
TREC19	**FELIX**$_{iter=20}$	Yes	0.15	0.22	**46.4**	**0.11**	**0.06**
	Vanilla	Yes	0.16	0.22	**46.4**	0.14	0.07
	PL	Yes	0.12	0.17	**46.4**	0.32*	0.13*
	PL-random	No	0.10*	0.15*	46.5	0.41*	0.16*
	Deterministic	No	**0.17**	**0.23**	46.6	0.12	0.07
	OMIT	Yes	0.13	0.18	46.5	0.15	**0.06**

improvement of FELIX$_{iter=20}$ is not just due to the evaluation metric introduced in this chapter but that there is an actual improvement w.r.t. the outlier use case.

We conclude that in our experiments, FELIX$_{iter=20}$ is able to effectively reduce the probability that a ranking with unknown exposure distribution is shown to the user, without a drop in utility or fairness, compared to other fair ranking methods and OMIT.

Discussion. If we compare our results to those in [185], OMIT does not perform as well as expected w.r.t. $P(u \mid \pi)$ and Outlierness. We see two reasons for this. First, OMIT considers outliers in the context of the whole list, while we consider outliers in the context of the top-k that they are presented in; their approach is able to remove outliers defined in the global context from the rankings but does not consider the outliers in the local context they are presented in, which is what we are evaluating for.

Second, in this chapter we consider individual fairness, while Sarvi et al. [185] report results on group fairness. For individual fairness the number of constraints is much higher, therefore the space we are optimizing over is smaller, making it challenging for OMIT to find a good solution that is optimized for both utility and reducing outliers while satisfying all the fairness constraints. FELIX$_{iter=20}$ does not suffer from this, since, instead of adding an additional objective term to the optimization, it intervenes at the decomposition step, making it independent from the constraints used in the optimization.

This comparison shows that FELIX is very general in terms of use cases that it can be applied to. The condition that determines whether a ranking has a known exposure distribution can be focused on each individual ranking without having to rely on global assumptions. This allows us to really consider inter-item dependencies, while OMIT needs to work with the heuristic of global outliers instead. This also highlights the advantages of FELIX over the aperL-ranker method. While for most experiments

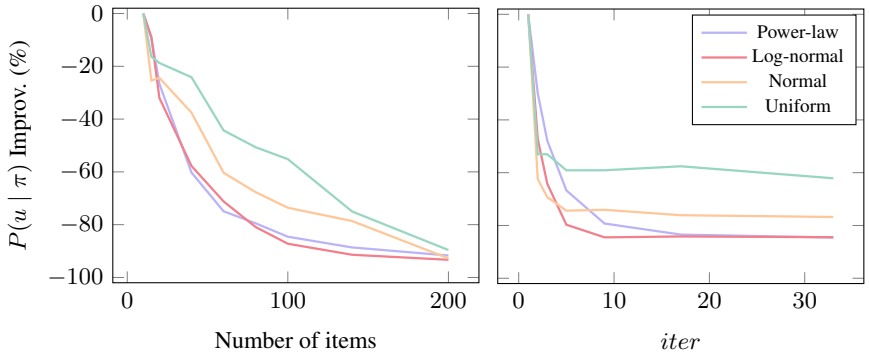


Figure 2.2: Sensitivity analysis. Relative reduction in $P(u \mid \pi)$ in $\%$ on the y-axis for different numbers of available candidate items (left) and different numbers of iterations (right).

there was no significant difference in utility and fairness between those two methods, considering inter-item dependencies within the rankings is not possible for the PL approach to fair ranking.

2.7 Sensitivity Analysis of FELIX

Given the results obtained in the previous section, we now analyze the ability of FELIX to reduce the number of rankings with unknown exposure distribution along two important dimensions: (D1) the number of available item candidates; and (D2) the number of re-sampling iterations, $iter$ (see line 3 in Algorithm 2).

For the TREC datasets most queries have less than 40 items, hence, we use a simulated set-up. This gives us more control, allowing us to observe FELIX's behaviour for different distributions and numbers of candidate items. Each analysis is conducted with a series of $m = 100$ simulated sets of n items (one can think of these item-sets as corresponding to m imaginary queries). Since we want to focus on the effectiveness of FELIX, rather than the quality of the predicted labels, we assume that for each item we know the correct probability that an item is relevant to users. For our analysis we sample these scores uniformly in the interval $[0, 1]$. The feature that is used for the outlier detection is sampled from a different probability distribution. We conduct experiments on the uniform, normal, log-normal, and power-law distribution to see how dependent the results are on the underlying data distribution. Each of the different distributions has a different base probability for a list of a given length to contain an outlier, and hence can be seen as different levels of difficulty for removing the rankings with unknown exposure distribution. With the definition of outliers used in this chapter and a list length of 10, the probability that such a list contains an outlier is 0.6% for the uniform, 2.7% for the normal, 36.3% for the log normal and 60.5% for the power-law distribution.

(D1) Candidate items. The left plot in Figure 2.2 shows the relative reduction of rankings with outliers with a varying number of candidate items. We use 20 re-sampling iterations. We see that for all distributions, FELIX performs increasingly better as the

number of items increases. Having more items to shuffle between various rankings gives the method more flexibility in putting outlier items into different contexts, in which they do not appear as outliers.

(D2) Re-sampling parameter. The right plot of Figure 2.2 shows how well FELIX is able to remove outliers from the rankings based on the number of re-sampling iterations, which is the only new hyper-parameter introduced by our method. We use 100 candidate items per query. We find that with an increasing number of re-samples, FELIX can remove more outliers. Nevertheless, the gains seem to be diminishing, depending on the distribution after 5–20 iterations.

Broader implications. Ranking systems often work in two stages, where in the first stage a certain number of documents are retrieved and in the second stage they are re-ranked with help of a learning to rank method. Our analysis of the number of candidate items (D1) can help deciding on how many items to retrieve in the first stage. Moreover, the analysis of the re-sampling parameter (D2) can help with deciding on a good performance/computation time trade-off when choosing the number of allowed re-sampling iterations.

2.8 Conclusion

Motivated by recent work on the impact of outliers on the exposure distribution within a ranking, we introduced the task of fair ranking under incomplete exposure estimation. We defined a new method, FELIX, that avoids showing rankings to the user which, due to inter-item dependencies, have unknown exposure distribution. We extended the convex optimization approach to fairness to the top-k setting and gave an efficient implementation of the algorithm that makes it feasible, even for a large number of items. We showed empirically that FELIX is able to significantly reduce the probability of generating rankings with unknown exposure, without hurting user utility or fairness compared to previous fair ranking methods.

FELIX is a first step towards fair ranking in cases where due to inter-item dependencies there is uncertainty about the exposure distribution of some rankings. By defining an efficient algorithm for the top-k setting, we enable the usage of the convex optimization approach towards fairness for use cases with a large number of items, which previously had been infeasible. We discussed that this approach gives more flexibility than other methods and allows, for example, to consider the relationship between items.

One limitation of our work in this chapter is that, since the policy achieved by the convex optimization is only fair in expectation, this approach is most useful for head queries with a large number of repetitions. Use cases where this might be applied include job search, where next to the individual fairness criterion a correction for historical biases should be considered, or item search for items that are frequently bought. Second, our results are based on the assumption that the unknown exposure comes from inter-item dependencies and that the same items that cause one ranking to have unknown exposure distribution, when placed in another context will result in a ranking with known exposure distribution. This assumption holds for rankings with visible outliers, however, to prove the generalizability of this approach, experiments with other use cases are needed. Lastly, to have enough flexibility within the Birkhoff-von Neumann

decomposition algorithm, enough entries of this matrix need to be non-zero. Using group fairness with only two groups, results in a marginal rank probability matrix that is a linear combination of just two permutation matrices [193]. More groups introduce more stochasticity, therefore this method is particularly interesting when working with individual fairness or a larger number of groups.

A potential direction for future work is to investigate whether FELIX can be extended for different user models. In this chapter we assume that most rankings follow a position-based exposure distribution. For other user-models like the cascade model a different approach might be necessary. Also, more research needs to be done on inter-item dependencies between items in a ranking and their impact on the exposure for different use cases. Phenomena like outliers or click bait have been explored to some extent but other types of cognitive bias that impact how we perceive items in relation to others have been broadly unexplored in the context of ranking systems. Lastly, extending user models to include inter-item dependencies such as outliers might allow for a more direct approach to fair ranking in cases where the exposure distribution is unknown.

Data and Code

To facilitate reproducibility of our work, all code and parameters are shared at `https://github.com/MariaHeuss/2022-SIGIR-FOE-Incomplete-Exposure`.

2.9 Proofs

2.9.1 Extended proof for the generalized Birkhoff-von Neumann

We give a more detailed proof of Lemma 2.4.2 and Theorem 2.4.1. Recall that we say that $P' \in \mathbb{R}^{n' \times k'}$ is an *extension* of $P \in \mathbb{R}^{n \times k}$ if $n' \geq n, k' \geq k$, and $P_{i,j} = P'_{i,j}$ for all (i,j) with $i \leq n$ and $j \leq k$. We denote this by $P'|_{i \leq n, j \leq k} = P$.

Lemma 2.9.1. *Let $P = \{a_{i,j}\}_{i \leq n, j \leq k}$ be a matrix with the same properties as described in Theorem 2.4.1 with $k \leq n$. Then there is a matrix $P' = \{a'_{i,j}\}_{i \leq n, j \leq n}$ with $\forall i,j : 0 \leq a'_{i,j} \leq 1$ such that $P = P'|_{i \leq n, j \leq k}$, and $\forall i : \sum_{j=1}^{n} a'_{i,j} = 1$ and $\forall j : \sum_{i=1}^{n} a'_{i,j} = 1$.*

Proof. Define $P' = \{a'_{i,j}\}_{i \leq n, j \leq n}$ as

$$a'_{i,j} = \begin{cases} a_{i,j} & \text{if } j \leq k \\ \frac{1 - \sum_{j'=1}^{k} a_{i,j'}}{n-k} & \text{if } j > k. \end{cases} \tag{2.12}$$

Then $P'|_{i \leq n, j \leq k} = P$ by definition. Since for all i, $0 \leq \sum_{j=1}^{k} a_{i,j} \leq 1$ we also have $0 \leq \frac{1 - \sum_{j'=1}^{k} a_{i,j'}}{n-k} \leq 1$. Moreover, for all $i \leq n$:

$$\sum_{j=1}^{n} a'_{i,j} = \sum_{j=1}^{k} a_{i,j} + \sum_{j=k+1}^{n} \frac{1 - \sum_{j'=1} a_{i,j'}}{n-k}$$

$$
\begin{aligned}
&= \sum_{j=1}^{k} a_{i,j} + (n-k) \cdot \frac{1 - \sum_{j'=1}^{k} a_{i,j'}}{n-k} \\
&= \sum_{j=1}^{k} a_{i,j} + (1 - \sum_{j'=1}^{k} a_{i,j'}) \\
&= 1,
\end{aligned}
$$

where we used in the second equality that we sum over $(n-k)$ times the same value. We know that the columns of the matrix sum to 1 for all $j \leq k$, since this is the case for P. For $j > k$ we have:

$$
\begin{aligned}
\sum_{i=1}^{n} a'_{i,j} &= \frac{1}{n-k}(\sum_{j=k}^{n}\sum_{i=1}^{n} a'_{i,j}) \\
&= \frac{1}{n-k}(n - \sum_{j=1}^{k}\sum_{i=1}^{n} a'_{i,j}) \\
&= \frac{n-k}{n-k} = 1.
\end{aligned}
$$

Here in the first equality we used that all columns from the k-th column are the same. In the second equality we used that since all rows are summing to 1, the sum of all rows (and therefore also the sum of all columns) equals n. The last equality simply uses the fact that each of the first k columns sums to 1. □

We use this Lemma to prove the generalized Birkhoff-von Neumann theorem. Let $k \leq n$.

Theorem 2.9.2. *Any matrix $P = \{a_{i,j}\}_{i \leq n, j \leq k}$ with $\forall i,j : 0 \leq a_{i,j} \leq 1$, $\forall j : \sum_{i=1}^{n} a_{i,j} = 1$ and $\forall i : \sum_{j=1}^{k} a_{i,j} \leq 1$ can be written as the convex sum $P = \sum_{l=1}^{m} \alpha_l \cdot P_l$ of permutation matrices P_l with coefficients $\alpha_l \in [0,1]$ such that $\sum_{l=1}^{m} \alpha_l = 1$.*

Proof. In Lemma 2.9.1 we show that P can be extended to a doubly stochastic matrix P', i.e. $P = P'|_{i \leq n, j \leq k}$. For this matrix P', the theorem by Birkhoff and von Neumann states that we can find a decomposition into the convex sum of permutation matrices, $P' = \sum_{l=1}^{m} \alpha_l P'_l$, with $\alpha_l \in [0,1]$, $\sum_{l=1}^{m} \alpha_l = 1$ and P'_l permutation matrices. This induces a decomposition of the original matrix P:

$$
P = \sum_{l=1}^{m} \alpha_l P'_l|_{i \leq n, j \leq k}. \quad \square
$$

2.9.2 Complexity of the generalized Birkhoff-von Neumann algorithm

In this section we prove the following claim from Section 2.4.2:

Theorem 2.9.3. *Using the modified top-k algorithm for the generalized Birkhoff-von Neumann theorem, Algorithm 1, a decomposition as described in Theorem 2.4.1 can be obtained with time complexity $\mathcal{O}(k^3n^2)$.*

Proof. The time complexity of Algorithm 1 depends on the complexity of the adjusted Hopcroft-Karp algorithm (line 5) and the number of times it needs to be executed (line 4–9), which is equal to the number of permutation matrices in the decomposition. Hopcroft and Karp [98] show that the time complexity of the Hopcroft-Karp algorithm is $\mathcal{O}((m+l)\sqrt{l})$, where l is the number of vertices and m is the number of edges in the biparate graph. For the baseline approach we have $l = 2n$ and $m = n^2$, therefore the complexity of the Hopcroft-Karp algorithm in this setting would be $\mathcal{O}(n^2\sqrt{(n)})$. Using our approach instead, we have $l = n + (k+1)$ and $m = n \cdot (k+1)$ which reduces the complexity to $\mathcal{O}(kn\sqrt{(n)})$. Furthermore since the maximum length of each augmenting path is bounded by $2 \cdot k$, we can substitute the $\sqrt{n}$ term with k (see Corollary 2 and Theorem 3 of [98]). This gives us a time complexity of $\mathcal{O}(k^2n)$ for the full matching algorithm. For the number of matrices in the decomposition, Johnson et al. [105] define an upper bound of $n^2 - 2n + 2$ permutation matrices, which means that the total complexity of the Birkhoff-von Neumann algorithm equals $\mathcal{O}(n^4\sqrt{n})$. Since for our algorithm, a loose upper bound for the number of permutation matrices is $k \cdot n$, the algorithm proposed in this chapter has a time complexity of only $\mathcal{O}(n^2k^3)$, which makes it much more feasible than the more naive algorithm proposed in Section 2.4.1 for large values of n. □

Conclusion to Chapter 2

Returning to research question RQ A: "*Can we define an exposure-fair ranking policy in situations where the expected exposure distribution is unknown for some rankings?*", we demonstrate that our proposed approach FELIX makes significant progress toward this goal. Specifically, FELIX effectively reduces the frequency with which a probabilistic ranking policy presents ranked lists with unknown exposure distributions to users, thereby improving overall exposure-fairness.

Our findings show that while the complete elimination of such rankings remains an open problem, substantial improvements in exposure-fairness are achievable. It is important to note that our investigation focuses specifically on scenarios where exposure distribution uncertainty arises from inter-document relationships, representing one important class of this broader problem.

Based on these results, we can answer research question RQ A positively, though with important caveats: FELIX does improve exposure-fairness when some rankings have unknown exposure distributions, but it cannot eliminate the problem entirely. This represents a meaningful step forward, even if the broader challenge of handling unknown exposure distributions in all cases remains open for future work.

3

Predictive Uncertainty-based Bias Mitigation in Ranking

In this second chapter on ranking fairness, we investigate how to reduce biases in ranked lists by using the uncertainty inherent in relevance assessments. While the literature frequently assumes that document relevance can be accurately determined, this assumption rarely holds in practice. We exploit this uncertainty to develop fairer ranking approaches.

Our focus in this chapter differs from that of Chapter 2 in two key ways. First, rather than addressing fairness through exposure-based methods, we explicitly target the removal of biased or stereotypical documents from top-ranked positions. Second, we work with text-based rankers, though our approach remains applicable to feature-based ranking systems.

Multi-objective optimization typically involves trade-offs between competing goals and improving fairness or unbiasedness often comes at the cost of user utility. However, since model predictions inherently carry uncertainty, we hypothesize that this trade-off can be managed more strategically. Specifically, we propose making fairness-oriented adjustments primarily where the model exhibits the greatest uncertainty about optimal ranking decisions.

This leads us to address the following research question:

RQ B Can we use the predictive uncertainty of the model prediction to improve ranking fairness?

Note: The notation in this chapter differs slightly from Chapter 2. Most notably, we use L to denote a ranked list (rather than σ), while σ represents the standard deviation of model predictions.

3.1 Introduction

The probability ranking principle (PRP) [175] states that, for optimal retrieval, the documents should be ranked in order of the predicted probability of relevance to the

This chapter was published as M. Heuss, D. Cohen, M. Mansoury, M. de Rijke, and C. Eickhoff. Predictive uncertainty-based bias mitigation in ranking. In *CIKM 2023: 32nd ACM International Conference on Information and Knowledge Management*, pages 762–772. ACM, October 2023.

user. While this principle is ideal with respect to user utility, a ranking approach that solely relies on this principle can lead to an unfair treatment of the documents through unfair exposure and learned historical biases that are implicit in the data [20, 248]. This realization has led to a broad range of work in the field of *fair* ranking, where ways of ranking are explored that do not always strictly follow the PRP, but instead correct for such historical biases and distribute exposure more fairly [31, 63, 157]. Such biases can be reflected in different ways, e.g., models can be biased to over proportionally favor members of one group over another [13]. In this chapter, we follow Rekabsaz et al. [168], and say that a ranking model is *biased*, if documents that contain biases or stereotypes towards a protected group, e.g., people identifying with a certain gender, are being placed in ranked lists for queries that should be inherently neutral.

Using uncertainty to mitigate biases and improve fairness. Recent work has highlighted how learned ranking models violate the PRP – that each score is not well calibrated, and that learned ranking models do not provide an equally reliable estimate of a document's relevance [47, 158]. In this chapter, we take advantage of this violation of assumptions to produce a fair ranking with minimal utility loss. Rather than relying on a deterministic score, we consider the *uncertainty of the model's estimate* to violate the PRP in an informed manner by focusing on the most *uncertain documents*.

Our proposed method, called **P**redictive **U**ncertainty based **F**air **R**anking (PUFR) exploits knowledge about the certainty of the predicted relevance scores for mitigating bias by intervening at the scoring distribution, making it a post-processing method that is easy to use on top of arbitrary ranking models. Furthermore, PUFR does not require any training or fine-tuning of supervised models. Rather, given a ranked list of documents generated by a ranking model (most likely biased), PUFR leverages the uncertainty of the predicted scores assigned to the candidate documents by the ranking model to modify the ranked list among the most uncertain positions to generate a fairer ranking. PUFR aims to reduce the impact of biased documents, while adhering to the PRP as closely as possible, only intervening in places where the ranking model was not very certain to begin with.

Additionally, we introduce an entirely post hoc uncertainty quantification procedure, based on Laplace approximation, that allows PUFR to approximate the uncertainty for any off the shelf model without access to the training data or optimization procedure. This is in contrast to past work that requires a specific training regime to produce the uncertainty scores for each candidate [46, 47, 158, 242].

Motivating example. In Figure 3.1, we visualize our approach to predictive uncertainty-based fairness, PUFR. In this example, the objective is to promote the unbiased documents (marked in green) to appear on top of the ranked result. We start by considering not only the mean ranking score but also the score distribution (uncertainty) as visualized with the cross resp. curve in Figure 3.1a. We chose confidence intervals relative to the standard deviation in which we allow PUFR to adjust the scores for each document, as can be seen in Figure 3.1b. Depending on whether a document is biased or not, we increase the score in this confidence interval if the document is unbiased or decrease it otherwise as visualized with the green/red crosses in Figure 3.1c. As the confidence intervals of the second (D2) and third (D3) documents *intersect*, this changes the order of these scores. After re-ranking with respect to the newly obtained scores, the pro-

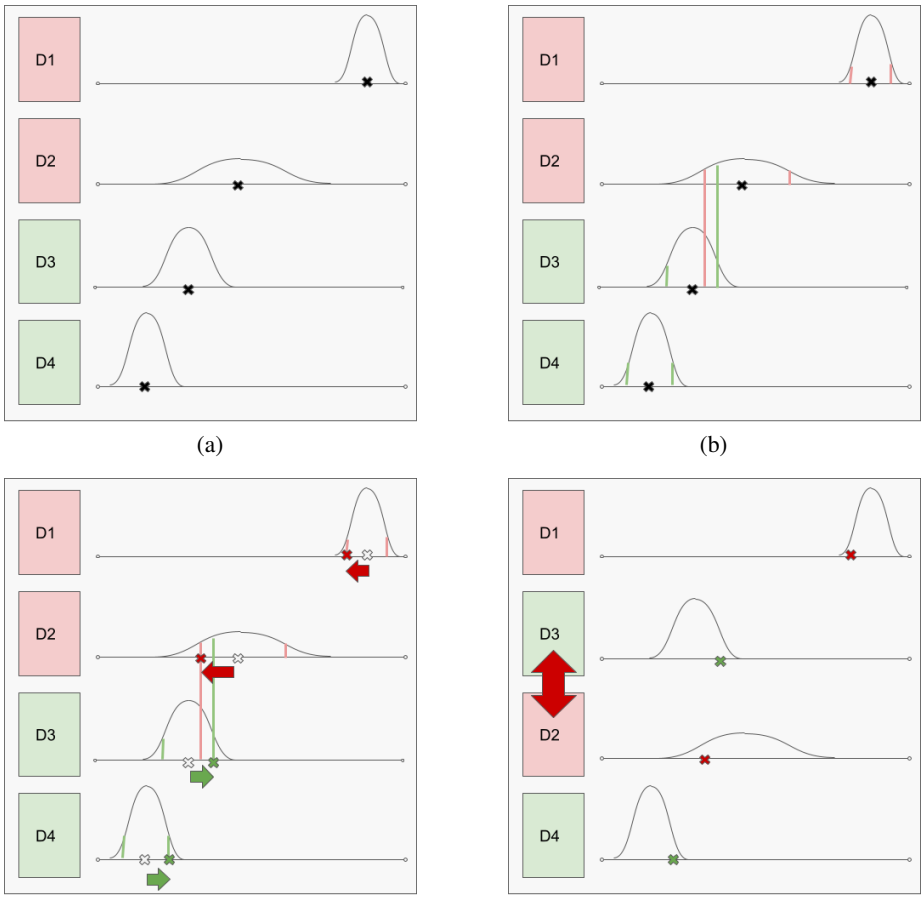


Figure 3.1: Visualization of our method PUFR. Next to the mean ranking scores PUFR also considers the score distribution that we obtained from the ranking model (3.1a). Through intersecting confidence intervals (3.1b) that allow us to adjust the scores (3.1c) such that a not biased document, visualized in green, is swapping place with a higher ranked, biased document (3.1d).

tected document D3 has swapped place with the non-protected document D2 as seen in Figure 3.1d. As there are minimal computational costs for PUFR, developers/users have the freedom to modify the trade-off between utility and fairness with minimal costs for their use-cases.

Our contributions. We summarize our contributions as follows:

- We introduce the notion of uncertainty-based fair ranking and analyze the potential of using the model uncertainty w.r.t. the ranking scores for bias mitigation.
- We define PUFR, an intuitive re-ranking approach that takes as input the ranking score distribution and calculates new ranking scores that can be used to create a less biased ranked list, while still preserving some certainty guarantees.

- We compare PUFR to several in- and post-processing bias mitigation methods and show that it outperforms all baselines, while being computationally much less expensive than some of them. Moreover, we demonstrate that PUFR is easily controllable with respect to the trade-off between fairness and utility, making it practical for use in real-life ranking applications.

3.2 Related Work

3.2.1 Uncertainty in ranking

Zhu et al. [259] introduce the notion of considering a model's confidence when ranking documents. The authors view the confidence of a score based on the probabilistic model's own estimate – the variance. Alternatively, we can assume a Bayesian perspective that considers how well the training data support the current model. As this approach does not rely on a probabilistic ranking model, it complements current ranking regimes. Penha and Hauff [158] first introduce this notion of uncertainty into conversational retrieval by incorporating dropout into a BERT architecture at inference time. The ranking score is then modified by an uncertainty measure to improve the final re-ranking. Cohen et al. [46] suggest a similar approach for ad hoc retrieval where only the last layer's uncertainty is measured to offset both the complexity of a neural model and the size of the document set with similar re-ranking improvements. Yang et al. [241] extend the above work by leveraging the uncertainty estimate to improve the exploration of an online learning to rank model. Rather than performing uncertainty-aware re-ranking, the uncertainty estimate is used to take an optimistic perspective on candidate documents to reduce the exploitation bias commonly found in an online learning to rank setting.

3.2.2 Mitigating bias and fair ranking

Recent years have seen a broad range of research on uncovering and mitigating biases in different information retrieval systems, such as biases in talent pool [82] and resume search [36] and the reinforcement of gender biases through search engines [68]. Rekabsaz and Schedl [167] explore the extent to which documents with gender bias can be found in the retrieved results of different neural retrieval models. Other work focuses more on the mitigation of such biases [e.g., 168, 253], where models are optimized to contain fewer biased documents for queries that are inherently unbiased. Rekabsaz et al. [168] use adversarial learning to remove gender bias from the trained model, Zerveas et al. [253] optimize the query representation from a previously trained architecture instead.

Mitigating biases is often framed as a fairness task. Zehlike et al. [251, 252] introduce a classification framework for fair ranking approaches, which we partly use to position our work in the existing fair ranking literature. As opposed to score-based fairness [33, 112, 205, 239], where the ranking scores are assumed to be known, in this chapter we focus on supervised learning to rank, where the ranking scores need to be determined with a ranking model.

A large body of work focuses on *merit-based* fairness, where the goal is to distribute the user attention in some way proportional to the merit of either individual documents

(individual fairness [e.g., 91, 120, 185]) or groups of documents [e.g., 20, 193, 223]. In contrast, other work [e.g., 248, 250] focuses on *representational* fairness, which is concerned with removing historical biases from the ranking or representing documents from different groups fairly w.r.t. some demographic within the ranking.

Independently of the notion of fairness, we differentiate between pre-processing [119], in-processing [19, 20, 168, 193, 194, 247, 248, 253], and post-processing [56, 113, 250] approaches to fairness interventions. These methods come into play either before the model is being trained, adjust the model or training process itself, or intervene after the model has been trained and the ranking scores are determined.

PUFR is a *post-processing* approach that aims to mitigate bias (*representational* unfairness) as opposed to prior in-processing work on the same task [168, 253]. While other work on post-processing approaches [such as, e.g., 32, 248] intervene at the ranked output, our approach instead adjusts the score distribution. What distinguishes PUFR from prior work on fair ranking is that we aim to exploit the uncertainty that the ranking model has on the predicted relevance scores to increase the fairness of the rankings.

3.2.3 Uncertainty in fair ranking

Prior work at the intersection of uncertainty and fairness can be grouped into two categories. The first category deals with uncertainty introduced when group membership cannot be determined with confidence. Ghosh et al. [84] discover that, when group labels are inferred from data, the usage of fair ranking methods can invalidate fairness guarantees and even increase the disadvantage that protected groups might receive. Mehrotra and Vishnoi [144] follow up on this work and develop a fair ranking framework for cases where socially-salient group attributes cannot be determined with certainty but are assumed to follow a given probability distribution.

The other category, which contains, among others, the work in this chapter, considers the predictive uncertainty stemming from imperfect prediction of merits and ranking scores. Yang et al. [242] are concerned with uncertainty in the relevance estimation. Unlike this chapter, the authors study an online setting where the relevance estimation is constantly updated. We target a static setting, not aiming to reduce the uncertainty for some exploration strategy but to exploit the uncertainty to obtain a better trade-off between fairness and utility.

Lastly, Singh et al. [195] are concerned with uncertainty in merit due to observations of secondary attributes instead of directly observing the merit. The authors suggest a probabilistic fairness framework in the presence of such uncertainty. Their work defines a notion of fairness that takes the uncertainty in the merit prediction into account, while we exploit uncertainty to, for example, correct for historic biases in the data and ranking model.

In summary, where existing methods either ignore the predictive uncertainty of ranking scores, aim to either reduce uncertainty, or take it into account when defining fairness, the work in this chapter is the first to harness uncertainty to improve the fairness-utility trade-off.

3.3 Method

We take an uncertainty-based approach to post hoc bias mitigation in ranking. We exploit the model's uncertainty over the predicted ranking scores to manipulate the ranking in a way that benefits documents that do not contain biases, which results in a fairer ranked list. By staying within a certain confidence range, we minimize the potential cost to utility. Following prior work [144, 168], we frame the task as a fair ranking problem.

Our method operates entirely through principled machinery and allows us to trade-off between user utility and fairness by adjusting a single coefficient. Furthermore, an existing ranker can be used as-is, without the need to retrain it, making it possible to use and adjust it for various levels of fairness, with little additional costs.

Below, in Section 3.3.1, we start by defining our notation and the fair ranking task. In Section 3.3.2, we introduce our method PUFR that, assuming that the predictive uncertainty over the ranking scores is given, uses those uncertainty values to develop a fair ranking approach. Finally, in Section 3.3.3 we follow with a description of how to attain the uncertainty of a given deterministic ranking model over its scores at inference time.

3.3.1 Notation and preliminaries

Given a query q, we consider the task of ranking documents from a candidate set $\mathcal{D}_q = \{d_{q,i}\}_i$ w.r.t. their relevance, to q. Regarding measured user utility only, an ideal ranked list would be ordered by decreasing document relevance. We assume a ranking model has been trained to order the documents w.r.t. the relevance to the query by predicting relevance scores. Most rankers are deterministic, outputting only a single predicted relevance score, $\mu_{q,i}$. In Section 3.3.3 we will describe how to approximate the uncertainty of predicted scores for such a model. We write $\sigma_{q,i}$ for the standard deviation of the predicted score $\mu_{q,i}$ for document $d_{q,i}$. Note that we implicitly assume the score distribution to be Gaussian.

Prior work has shown that models that are trained solely for maximizing the measured utility can be biased and contain unfair representations of the resulting ranked lists [167]. In this chapter, as an additional objective, we aim to decrease the presence of biased documents in the ranked lists. We treat the task as a fair ranking problem, where we want to increase the exposure of the protected group $\mathcal{D}_q^P \subset \mathcal{D}_q$ of documents without biases and decrease the exposure of the non-protected group $\mathcal{D}_q^N \subset \mathcal{D}_q$ of documents that contain biases.

3.3.2 PUFR: Uncertainty-aware fairness

In this section, we introduce our post-processing fairness intervention method **P**redictive **U**ncertainty based **F**air **R**anking, PUFR. The core idea of PUFR is to take advantage of the uncertainty of the model over the predicted ranking scores to adjust these scores proportional to the standard deviation of the predictive distribution for each document, allowing fairness adjustments with minimal cost to the utility. For now, we treat the

score distribution for each document, $\mathcal{N}(\mu_{q,i}, \sigma^2_{q,i})$, as being given, but in Section 3.3.3 we describe how to obtain it for a deterministic ranker.

As the goal of PUFR is to mitigate bias and hence increase the fairness of the ranking system, PUFR accomplishes this by swapping some of the documents of the protected group, $\mathcal{D}^P_q$, with higher ranked documents of the non-protected group, $\mathcal{D}^N_q$. Since the uncertainty of the scores for the documents within the same group can differ greatly, this allows for a tuned adjustment of the ranking scores where swaps only occur in settings where there exists a reasonable chance of the documents being equally relevant, quantified by the model's uncertainty, $\sigma_{q,i}$.

In other words, we allow PUFR to pick ranking scores that maximize fairness in intervals $[\mu_{q,i} - \alpha \cdot \sigma_{q,i}, \mu_{q,i} + \alpha \cdot \sigma_{q,i}]$, without re-ordering the documents within the same group. Here, α is a user defined hyper-parameter that quantifies the chance of a utility violation when performing this procedure. A higher value of α will result in a fairer ranking but at the cost of less accurate predicted scores, and hence potentially a drop in utility.

As shown in Algorithm 3, PUFR initially loops over all documents of the protected group $d_{q,i} \in \mathcal{D}^P_q$, sorted w.r.t. decreasing ranking score, $\mu_{q,i}$, see line 1. PUFR then increases the score as much as possible while staying within the confidence bounds, i.e.,

$$\tilde{\mu}_{q,i} = \mu_{q,i} + \alpha \cdot \sigma_{q,i}. \tag{3.1}$$

See line 2. To avoid intra-group swapping of documents, modified ranking scores are bounded by the lowest score of any higher ranked document within the same group:

$$\tilde{\mu}_{q,i} \leq \min_{\mathcal{D}^P_q, j \leq i}(\tilde{\mu}_{q,j}), \tag{3.2}$$

where j, i are rank positions, see line 3. Equivalently, for all documents of the non-protected group, $d_{q,i} \in \mathcal{D}^N_q$, we decrease the score as follows, this time starting with the document with the lowest ranking score (see line 5):

$$\tilde{\mu}_{q,i} = \mu_{q,i} - \alpha \cdot \sigma_{q,i}, \tag{3.3}$$

see line 6. Again, to avoid the same intra-group swapping for the non-protected group, we lower bound the adjusted scores by the maximum score of all documents in the same group that are ranked lower in the original ranking:

$$\tilde{\mu}_{q,i} \geq \max_{\mathcal{D}^N_q, j \geq i}(\tilde{\mu}_{q,j}). \tag{3.4}$$

See line 7. PUFR then uses these adjusted scores $\tilde{\mu}_{q,i}$ to re-rank the documents (line 9).

Note that even though we define PUFR for a setting with only one protected document group, it can be extended to several protected groups, that need to receive different treatments. Our approach allows us to adjust the strength of the score adjustment individually for each group, e.g., enabling a stronger correction for more disadvantaged groups, by allowing a group-wise choice of hyper-parameter α_g.

Many pre-trained ranking models do not output the uncertainty scores $\sigma_{q,i}$ that PUFR employs to reorder rankings. Thus we need a way to approximate the uncertainty scores $\sigma_{q,i}$ in a post-processing manner. Next, we show how to do this with the help of Laplace approximation.

Algorithm 3 Predictive **U**ncertainty based **F**air **R**anking (PUFR)

Require: mean ranking scores $\{\mu_{q,i}\}_{d_{q,i} \in \mathcal{D}_q}$, standard deviation $\{\sigma_{q,i}\}_{d_{q,i} \in \mathcal{D}_q}$, control parameter α, groups $\mathcal{D}_q^P$, $\mathcal{D}_q^N$

for all $d_{q,i} \in \mathcal{D}_q^P$, sorted by decreasing $\mu_{q,i}$ **do**
    $\tilde{\mu}_{q,i} \leftarrow \mu_{q,i} + \alpha \cdot \sigma_{q,i}$
    $\tilde{\mu}_{q,i} \leftarrow \max_{\mathcal{D}_q^P, j \leq i}(\tilde{\mu}_{q,j})$
end for
for all $d_{q,i} \in \mathcal{D}_q^N$, sorted by increasing $\mu_{q,i}$ **do**
    $\tilde{\mu}_{q,i} \leftarrow \mu_{q,i} - \alpha \cdot \sigma_{q,i}$
    $\tilde{\mu}_{q,i} \leftarrow \min_{\mathcal{D}_q^N, j \geq i}(\tilde{\mu}_{q,j})$
end for
Obtain ranking L by sorting documents $d_{q,i} \in \mathcal{D}_q$ with respect to scores $\tilde{\mu}_{q,i}$
return L

3.3.3 Attaining uncertainty scores from a deterministic ranking model

The goal is to attain effective uncertainty scores, σ, from a ranking model at inference time; conventional uncertainty approaches fail to satisfy this condition [46, 158, 241, 242]. Past approaches have relied on a specific training regime – Monte Carlo (MC) dropout – to achieve an effective Bayesian model. As PUFR is a post hoc method, we leverage an alternative form of uncertainty, *Laplace approximation*, that can be applied to any already trained ranking model.

The standard approach to training a deterministic model f, where there exists a single output for each input, is to learn a set of parameters, θ_{MAP}, that minimizes the loss function

$$\mathcal{L}(\theta) = -\ln P(\theta \mid \mathcal{D}) + r(\theta), \tag{3.5}$$

where r is some regularization on θ and $\mathcal{D}$ is the training dataset. While this is a probabilistic interpretation of the loss function and optimization process, prior work has mapped margin-based ranking losses to this framework [46]. At inference time, the model, f, is evaluated using the single point θ_{MAP}, which minimizes $\mathcal{L}(\theta)$. Alternatively, a Bayesian perspective captures the uncertainty of the model by considering all possible θ values weighed by how likely they are based on the training data using the posterior $P(\theta \mid \mathcal{D})$, with θ_{MAP} as the most likely value. This produces a distribution over outputs, of which the variance σ^2 represents the uncertainty present within the model and $\mathcal{D}$:

$$P(y \mid x, \mathcal{D}) = \int_\theta P(y \mid x, \theta) P(\theta \mid \mathcal{D}) d\theta, \tag{3.6}$$

with x as the input and y as the output of the model. Unfortunately, capturing this distribution is intractable for all but the smallest models due to the nature of computing the posterior $P(\theta \mid \mathcal{D})$. There exists prior work that approximates this distribution using MC Dropout [46, 158, 241, 242]. However, this requires a specific training regime, which would prevent the general application of PUFR to off-the-shelf architectures or previously trained ranking models.

Using Laplace approximation for post-hoc uncertainty approximation. We propose using Laplace approximations (LA), which can turn any conventionally trained deterministic model into a Bayesian model at inference time to produce the necessary σ values for PUFR [140]. LA encompass a family of approaches that fit a local Gaussian around the MAP estimate (3.5) via a second-order Taylor expansion of the log posterior:

$$\begin{aligned} \ln P(\theta \mid \mathcal{D}) \approx & \ln P(\theta_{\text{MAP}} \mid \theta) \\ & \frac{1}{2}(\theta - \theta_{\text{MAP}})^{\intercal} \bar{H} (\theta - \theta_{\text{MAP}}), \end{aligned} \tag{3.7}$$

where $\bar{H}$ is the expected Hessian at θ_{MAP}. The key observation is that the right side only requires the deterministic model, θ_{MAP} to produce the log Bayesian posterior distribution on the left side. Then, to recover the full posterior, exponentiating both sides reveals the Gaussian functional form for θ,

$$\begin{aligned} P(\theta \mid \mathcal{D}) \approx & P(\theta_{\text{MAP}} \mid \mathcal{D}) - \\ & \exp\left(\frac{1}{2}(\theta - \theta_{\text{MAP}})^{\intercal} \bar{H} (\theta - \theta_{\text{MAP}})\right) \\ \approx & \mathcal{N}(\theta_{\text{MAP}}, \bar{H}^{-1}). \end{aligned} \tag{3.8}$$

Thus, this approximation can take any twice differentiable off-the-shelf model and conveniently convert it to a Bayesian model at inference time by inverting the Hessian. While inverting to produce the covariance matrix is intractable for most models, we leverage past work by only inverting the last layers of a neural model to achieve actionable uncertainty estimates with near-zero cost [46, 47] (Algorithm 4, lines 2–3). While there exists a closed form linearization of Eq. 3.8, we are able to achieve sufficient efficiency using Monte Carlo sampling to capture the predictive distribution $P(y \mid x, f)$ by sampling from the Gaussian (line 5), $\mathcal{N}(\theta_{\text{MAP}}, \bar{H}^{-1})$ [54],

$$\begin{aligned} P(y \mid x, \mathcal{D}) = & \int_{\theta} P(y \mid x, \theta) P(\theta \mid \mathcal{D}) d\theta \\ \approx & \frac{1}{N} \sum_{t=1}^{N} p(y \mid x, \theta_t), \theta_t \sim \mathcal{N}(\theta_{\text{MAP}}, \bar{H}^{-1}). \end{aligned} \tag{3.9}$$

Furthermore, as the covariance matrix H^{-1} is viewed as independent to the training process, we do not need to use the original loss function either [115]. Lastly, for further efficiency, we exploit the property that the Hessian, H, is equivalent to the Fisher information matrix, F, at θ_{MAP}. As shown in Algorithm 4, we therefore approximate H by taking the diagonal of F, which is a common approximation regime (line 3) [87, 173].

After estimating $\mathcal{N}(\theta_{\text{MAP}}, \bar{H}^{-1})$ for the last layer of a neural model, we sample this distribution N times to produce N versions of the last layer, in order to produce $\mu_{q,\cdot}$ and $\sigma^2_{q,\cdot}$ as parameters of the predictive distribution $P(y \mid x, \mathcal{D}) = \mathcal{N}(\mu_{q,\cdot}, \sigma^2_{q,\cdot})$ (line 7–8). These parameters are then used by PUFR as described in Section 3.3.2 to debias the ranked list.

Algorithm 4 Post hoc uncertainty estimation for single query

Require: pre-trained l-layer model f_θ, $\theta_{\text{MAP}} = [\theta^1_{\text{MAP}}, \ldots, \theta^l_{\text{MAP}}]$, query q, candidate documents $\mathcal{D}_q = \{d_{q,i}\}_i$, Monte Carlo sample size N.

for all $d_{q,i} \in \mathcal{D}_q$ **do**
  $h_i^{l-1}, y = f_{\theta_{\text{MAP}}}(q, d_{q,i})$
  $H \approx \text{diag}(F) = \text{diag}(\mathbb{E}\big[\nabla_{\theta^l} \ln P(y \mid q, d_{q,i}))^2\big]$
  for all $j \in N$ **do**
    $\{\theta\}_1^j \sim \mathcal{N}(\theta^l{}_{\text{MAP}}, \text{diag}F^{-1})$
  end for
  $\mu_{q,i} = \frac{1}{N} \sum_{t=1}^{N} f_{\theta_t^l}(h_i^{l-1})$
  $\sigma^2_{q,i} = \frac{1}{N} \sum_{t=1}^{N} f_{\theta_t^l}(h_i^{l-1})^2 - \left(\frac{1}{N} \sum_{t=1}^{N} f_{\theta_t^l}(h_i^{l-1})\right)^2$
end for
return $\mu_{q,i}, \sigma_{q,i}\ \forall d_{q,i} \in \mathcal{D}_q$

3.4 Experimental Setup

We aim to answer the following research questions with our experiments: (RQ3.1) Based on empirical findings, are the uncertainty intervals around the ranking scores of a Bayesian ranking model sufficiently intersecting to allow for a re-ranking of documents, while staying within reasonable certainty bounds? (RQ3.2) Can PUFR be used to reduce the number of biased documents that are ranked on top of the list more effectively than prior methods? (RQ3.3) How do the various methods for fairness interventions compare with respect to controllability and computational efficiency?

There are four properties that we consider relevant to answer these questions: (i) We want to improve the fairness within the rankings. (ii) We want to do so with the least loss in utility possible. (iii) The next property is the controllability of the approach at hand. A human user/engineer should be able to easily adjust the trade-off between fairness and utility to fit their purposes. (iv) The last property is computational efficiency since this can also play a role when choosing a fairness method.

Next, we detail our experimental design. Then we discuss the evaluation metrics that we use to measure the four properties mentioned above (Section 3.4.2) and the dataset that we use (Section 3.4.3). Section 3.4.4 summarizes the baselines that we compare against.

3.4.1 Experimental design

We perform our experiments on a web search task, where for each query, the objective is to rank documents that might be relevant to that query. In addition to the requirement of being relevant to the user, the ranked list should not contain any gender biases for queries that are naturally non-gendered [168]. Therefore, we consider only non-gendered queries and expect a fair ranking model to not promote any documents that are biased towards some gender. See Section 3.4.3 for a discussion on the data used for this task.

To get an effective impression of the trade-off between utility and fairness, we perform a range of experiments per baseline, by varying some hyperparameter α. We define this hyperparameter individually for each baseline, based on the respective underlying algorithms (see Section 3.4.4).

To demonstrate the efficacy of PUFR on current search models, we use the BERT ranker introduced by Nogueira and Cho [155] as it represents a common language model architecture in current ranking regimes [58, 96, 132, 183]. Due to hardware constraints, we use Bert-Mini [212], a distilled four-layer version of BERT that performs comparably to the full model in search and other related tasks. We note that in the case of uncertainty modeling, Cohen et al. [46] demonstrate that a distilled model results in less expressive ranking uncertainty compared to larger variants of the same architecture on the same data. Thus, Bert-Mini represents a challenging setting and a conservative estimate of PUFR's performance.

To facilitate reproducibility of the work in this chapter, all code and parameters are made available; see Section 3.7.

3.4.2 Evaluation

User utility and fairness are measured per query. To get a single score to compare across methods, we report the mean over all queries. We measure significance with paired t-tests, where we treat the results of each query as one sample.

User utility. To measure user utility, we use the nDCG metric (normalized discounted cumulative gain). We use different cut-offs to measure the user utility in the top-10 documents, as well as for the first 100 documents.

Fairness. As discussed in Section 3.4.1, our task entails reducing the impact of strongly biased documents in the presented rankings. Therefore, we use the nFaiRR metric as a measure of fairness introduced by Rekabsaz et al. [168]. For a ranked list L, the FaiRR score at cut-off k is defined as:

$$\text{FaiRR@k}(L) = \sum_{\text{rank}_L(d_i) \leq k} n_{d_i} \cdot \frac{1}{\text{rank}_L(d_i)}, \tag{3.10}$$

where $\text{rank}_L(d_i)$ describes the rank of candidate document d_i in L, and the neutrality score $n_{d_i} \in [0, 1]$ is lower, the more biased a document is. Since the possible range of FaiRR scores depends on the distribution of neutrality scores of its candidate documents, to make the results easier to interpret and better comparable among queries, we use the *normalized FaiRR score* (nFaiRR). For this, we normalize the FaiRR score with the highest attainable FaiRR score with the document candidates for this query, similar to how nDCG is calculated from DCG. In our experiments we measure the nFaiRR at a cut-off value of 10 and 50. We select a different cut-off than the utility measure (@100) so as to compare with reported values from the baseline evaluations.

Controllability. We follow prior work [168], and focus on a qualitative analysis of the results by investigating the predictability of the utility-fairness trade-off when adjusting the controllable hyperparameter of each of the methods. An ideal approach should have small change in utility and fairness for a small change in α. To this end, we compare the plots in Figure 3.6 below.

Computational efficiency. For computational efficiency, we measure the run time of our implementation for each approach. We acknowledge that method-specific performance optimization might be able to further improve on the run times observed for the generic implementations used here, but assume that at least a rough execution time comparison can be gleaned. We measure the run time of each query and report the mean run time in Table 3.1.

Significance testing. To test the significance of observed differences in evaluation scores, we perform two-tailed paired t-tests on the metrics, treating the results of an approach of each query as a measurement of the same random variable. In Table 3.1, we mark results with an asterisk if they are significantly different from those of PUFR.

3.4.3 Dataset

The retrieval models that we use are trained on the MS MARCO Passage Retrieval collection [154]. For evaluation, we use MS $\text{MARCO}_{\text{Fair}}$, a subset consisting of 215 queries from the validation set that are non-gendered in nature – i.e., not containing any words or concepts that could be attributed to some gender [168]. However, the top candidate documents for these queries are highly associated with gender [168, 253]. We quantify the degree of gender bias for each document using the neutrality scores provided by Rekabsaz and Schedl [167] in order to measure fairness. We define documents with neutrality score 1 as the protected group for the post-processing baselines and PUFR.

3.4.4 Baselines

The baseline fairness intervention methods that we consider include the two in-processing approaches that have been introduced for the same bias mitigation task and dataset used here [168, 253]. Since PUFR is a post-processing approach, we add two commonly used post-processing fairness approaches that have been slightly adjusted to fit the task. Both post-processing baselines as well as UNFAIR use the mean scores $\mu_{q,i}$, produced by Algorithm 4 in Section 3.3.3 for the BERT-based ranker (see Section 3.4.1) as ranking scores. For each baseline the hyper-parameter α that allows us to control the trade-off between utility and fairness, is defined individually.

UNFAIR. The ranking resulting from ordering the documents with respect to the mean scores $\mu_{q,i}$, without considering fairness.

ADV. The (in-processing) adversarial fairness optimization from [168], which shares the same underlying BERT re-ranking architecture as discussed in Section 3.4.1. However, training is done using an adversarial discriminator head that attempts to predict whether the document is gendered or neutral by optimizing a classification loss function. The gradient from this loss is reversed within the main BERT architecture, therefore moving the parameters away from regions that can effectively capture gender [73]. We implement this model using the source code and suggested hyperparameters provided by the authors. The controlling hyperparameter α (originally λ) is defined by the scale of the reversed gradient.

CODER. This (in-processing) baseline [253] is intended for dense retrieval architectures. The method directly optimizes the query representation from a previously trained

architecture, TAS-B [96], by jointly optimizing thousands of candidate documents in a list-wise manner. While improving overall ranking performance, the large candidate pool within a list-wise loss provides a stable and competitive way to incorporate fairness directly during training. We include this baseline not as a direct comparison with respect to ranking performance, but to provide context on how a direct list-based fairness optimization approach compares to methods that operate entirely within a post hoc framework when viewed from a utility-fairness trade-off perspective. Here, the hyperparameter α (in the original paper λ_r) is defined as the regularization coefficient for the neutrality loss.

CVXOPT. A (post-processing) convex optimization approach similar to [32]. For each query we optimize the ranking L for utility, measured by nDCG, under a constraint on the nFaiRR score, $\mathrm{nFaiRR}(L) \geq \alpha$. To keep computational costs within a reasonable range, we only re-rank the first 50 documents of each query.

FA*IR. A (post-processing) approach suggested in [248]. We use a significance parameter 0.1 as suggested in [248] and vary p, the desired minimal proportions of documents with the protected attribute in the top-k for any value of k. In the remainder of this chapter we use $\alpha := p$, not to be confused with the significance parameter in the original paper, to match the other methods. For a fair comparison w.r.t. to computational efficiency, we use an efficient implementation that pre-computes the required number of protected documents for each rank upfront via an iterative algorithm.

3.5 Experimental Results

We present and discuss answers to our research questions.

3.5.1 Intersections of uncertainty intervals

Recall (RQ3.1): *Based on empirical findings, are the uncertainty intervals around the ranking scores of a Bayesian ranking model sufficiently intersecting to allow for a re-ranking of documents, while staying within reasonable certainty bounds?* To answer (RQ3.1), we analyze the confidence intervals of the ranking scores. If the uncertainty intervals do not intersect much, the ranking model is very certain about the ordering of its ranking scores. In such a case, our approach, or any uncertainty-aware approach in general, would not be able to re-rank the documents within an acceptable utility bracket. Previous work has shown that ranking models tend to be very certain for the ranking scores of highly ranked documents [46], but certainty decays when going down the ranked list. We are interested in how much flexibility a rank-aware fairness approach would offer in swapping documents by allowing the ranking scores to take values in a given certainty $[\mu_{q,i} - \alpha \cdot \sigma_{q,i}, \mu_{q,i} + \alpha \cdot \sigma_{q,i}]$ interval around the mean score value $\mu_{q,i}$. Figure 3.2 shows the median number of documents with intersecting confidence intervals (i.e. the median number of documents that the document at that rank could swap position with) for $\alpha = 1$ resp. $\alpha = 2$ standard deviations.

Even for documents ranked at higher positions, there is flexibility to change the order of the ranking. For a confidence interval of 1 standard deviation, most documents in the top-10 each have at least 6 documents that they could swap rank with. If we

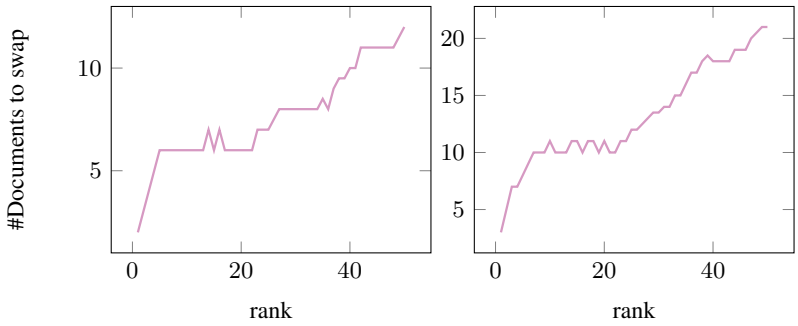


Figure 3.2: MSMARCO$_{\text{FAIR}}$: Median number of documents that have intersecting uncertainty intervals with the document placed at each rank for uncertainty intervals of 1 (left) resp. 2 (right) standard deviations.

look at confidence intervals of two standard deviations, this number increases to $\sim$10 documents that the document at rank 10 can swap place with. We therefore answer (RQ3.1) positively: The uncertainty intervals around the ranking scores of the Bayesian ranking model are sufficiently intersecting to allow for a re-ranking of documents, while staying within acceptable certainty bounds for utility.

Having confirmed that within the uncertainty of the model there is flexibility for an uncertainty-based fairness approach to change the order of documents, we address our second research question that asks whether the proposed approach can improve fairness.

3.5.2 The fairness utility trade-off

Recall (RQ3.2): *Can PUFR be used to reduce the number of biased documents that are ranked on top of the list more effectively than prior methods?* To answer this question we refer to Figure 3.3 and 3.4, where we plot fairness on the x-axis against utility on the y-axis, for PUFR and the baselines discussed in Section 3.4.4, for different values of the respective hyper-parameter α that controls the trade-off. In addition we use Table 3.1, where we compare the experimental outcomes with the best nFaiRR value for a given minimum utility requirement.

Utility-fairness trade-off. In Figure 3.3 and 3.4, we observe that the CODER baseline starts with a better trade-off for the top-10 documents, which can be attributed to better ranking scores that it starts out with (PUFR uses a BERT-based model to obtain ranking scores). CODER's advantage quickly vanishes as the balancing parameter α increases for more weight on fairness. Overall, PUFR offers a better trade-off between fairness and utility than the CODER based and the adversarial fairness optimization baseline (ADV).

If we compare PUFR to the post-processing baselines (CVXOPT and FA*IR), it clearly outperforms those baselines. Once a nFaiRR value of 0.96 is reached the advantage of PUFR over these baselines becomes smaller. For a possible explanation see Section 3.6.

Overall, PUFR outperforms all baselines for a large range of nFaiRR values, which we also highlight by comparing the fairness of the different approaches at two different

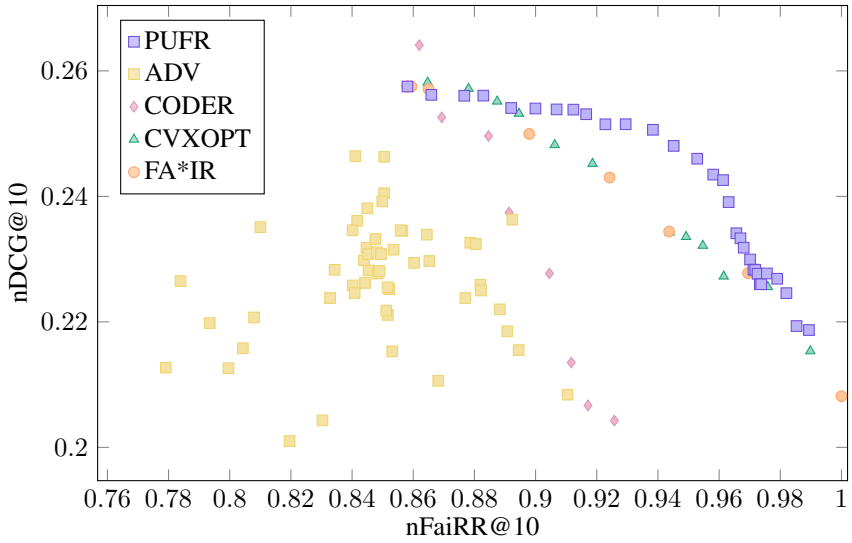


Figure 3.3: Trade-off between fairness and utility evaluated on the first 10 documents.

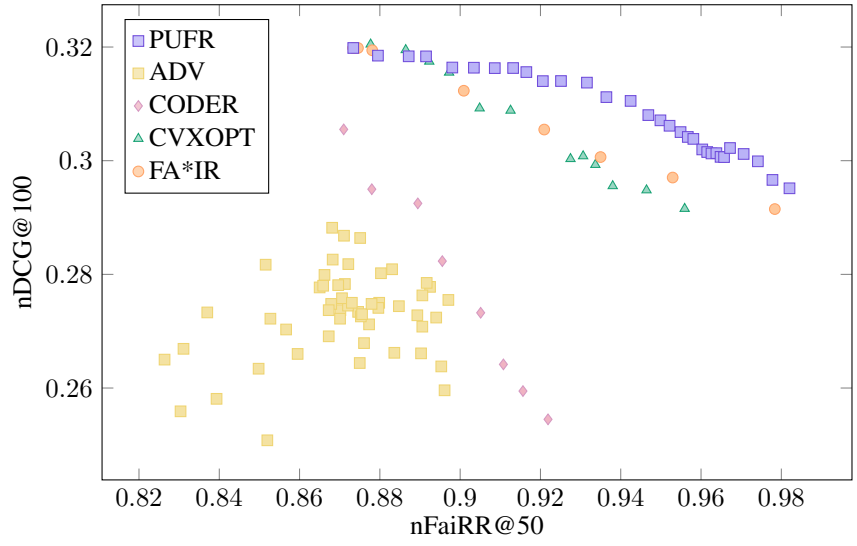


Figure 3.4: Trade-off between fairness and utility evaluated on the first 50 resp. 100 documents.

Table 3.1: Results for experiment with best nFairr value for nDCG decrease not more than 0.01 and 0.02 respectively. ADV baseline does not fulfill the criteria of being at most 0.01 nDCG points worse than UNFAIR. * denotes significance w.r.t. PUFR via two tailed paired students t-test of $p < .05$.

			nDCG↑		nFaiRR↑		re-rank-	req.
	Method	α	@10	@100	@10	@50	time(s)↓	train
	UNFAIR	0.0	0.26	0.32	0.858	0.873	0.00	No
	ADV	2.0	0.21	0.26	0.91	0.896	-	Yes
$nDCG_{100} \geq 0.31$	**PUFR**	2.5	**0.25**	**0.31**	**0.938**	**0.932**	0.014	No
	CODER	3.0	**0.25**	**0.31**	0.920*	0.920*	-	Yes
	CVXOPT	0.8	**0.25**	**0.31**	0.906*	0.905*	0.123	No
	FA*IR	0.7	**0.25**	**0.31**	0.898*	0.901*	0.058	No
$nDCG_{100} \geq 0.30$	**PUFR**	7.0	0.23	**0.30**	**0.970**	**0.960**	0.014	No
	CODER	4.0	**0.24**	**0.30**	0.927*	0.926*	-	Yes
	CVXOPT	0.91	0.23	**0.30**	0.949*	0.931*	0.123	No
	FA*IR	0.85	0.23	**0.30**	0.944*	0.935*	0.058	No

utility levels ($\mathrm{nDCG@100} = 0.31$ and $\mathrm{nDCG@100} = 0.30$) in Table 3.1. We chose these levels of utility, assuming that, when taking a fair ranking approach in production there might be a certain (small) allowance for a drop in utility given, within which the best possible fairness value should be reached. We see that for these levels PUFR reaches significantly higher scores for nFaiRR than all baselines.

Ablation study. To ensure that the uncertainty estimates indeed do contribute to the success of PUFR, we conduct an ablation study. We compare PUFR with a similar approach that, instead of adjusting the scores relative to the standard deviation, in- or decreases all scores by the same, constant value. In our experiments we use the mean uncertainty score over all queries and candidates documents, $\sigma_{\text{mean}} = \text{mean}_{q,i}(\sigma_{q,i})$. The results of this ablation study are presented in Figure 3.5. We see that by using the uncertainty scores instead of a uniform correction factor, we gain a better trade-off. For the top-10, these improvements are less visible (see Figure 3.5 (a)). When considering the top-100 documents instead, the advantages of using uncertainty become much clearer (see Figure 3.5 (b)). This might be due to fact that, as also noted by Cohen et al. [46], for the top-10 documents the uncertainty scores tend to be fairly similar to each other, making our approach, if we only look at a small window, seem similar to the ablation study approach. When we look at a larger window, the uncertainty scores deviate more, emphasizing the advantages of PUFR.

We conclude this section and answer (RQ3.2) in the affirmative. PUFR performs competitively with baselines. In terms of fairness-utility trade-offs it significantly outperforms other post-processing schemes, and clearly beats the two state-of-the-art in-processing baselines. The ablation study confirms that this result is at least partially due to the use of the model's uncertainty in its scores. Hence, PUFR can be used to

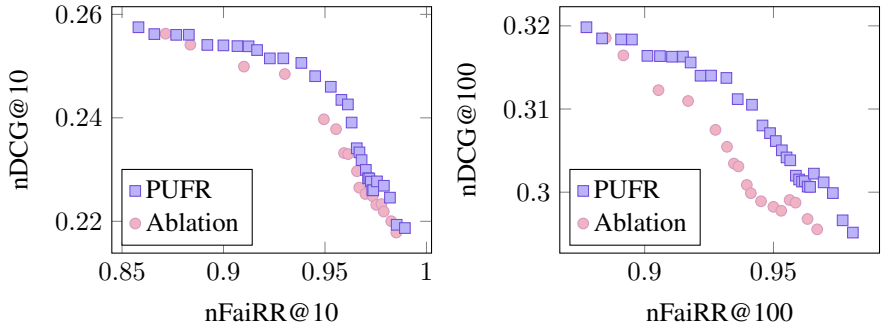


Figure 3.5: Ablation study comparing PUFR (score adjustment proportional to the ranker's uncertainty) with an ablation experiment with uniform score adjustment.

reduce the number of biased documents that are ranked on top of the list more effectively than prior methods.

Since a good utility-fairness trade-off is not the only relevant criterion when choosing a fair ranking method, our next research question (RQ3.3) concerns the degree of controllability and computational costs of the different methods.

3.5.3 Controllability and computational efficiency

Next, we address (RQ3.3): *How do the various methods compare with respect to controllability and computational efficiency?* As discussed in Section 3.4.2, we focus on a qualitative analysis of the α-fairness and α-utility curves, evaluating how predictable and hence controllable the utility-fairness trade-off is. Figure 3.6 shows that for PUFR the nFaiRR score monotonically increases with increasing α. At the same time, utility, measured by nDCG, decreases. Both curves are highly predictable. Furthermore, since re-ranking is computationally very efficient, a broad range of rankings with different trade-offs can be explored to find the right choice of hyper-parameter for the desired trade-off between nFaiRR and nDCG. The CODER-based approach has similarly predictable trade-off curves as PUFR [253]. However, CODER is an in-processing approach, meaning that the model needs to be re-trained for each choice of hyper-parameter α, making it much less controllable in practice. The ADV method on the other hand, seems to be highly unpredictable, on top of the downsides that come with in-processing methods as discussed above. For the FA*IR baseline, although its curve seems to be fairly well controllable, the granularity in which we can produce results is much coarser. Due to space constraints we omit the figure for the convex optimization approach; because of computational efficiency, FA*IR or PUFR should be preferred over it.

With regard to computational efficiency, we recall that both in-processing approaches, ADV and CODER, once trained, do not have the post-processing overhead of the other methods. However, these methods need a large amount of training to gain a reasonable level of performance [168, 253]. Looking at Table 3.1, re-ranking with PUFR is much faster than with the other two post-processing approaches. Obtaining uncertainty labels can be done within microseconds. After adjusting the ranking scores

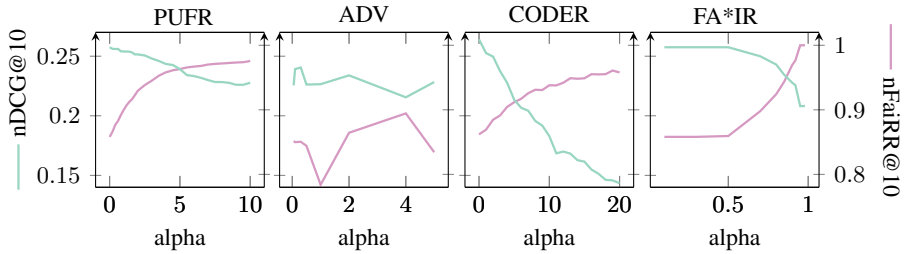


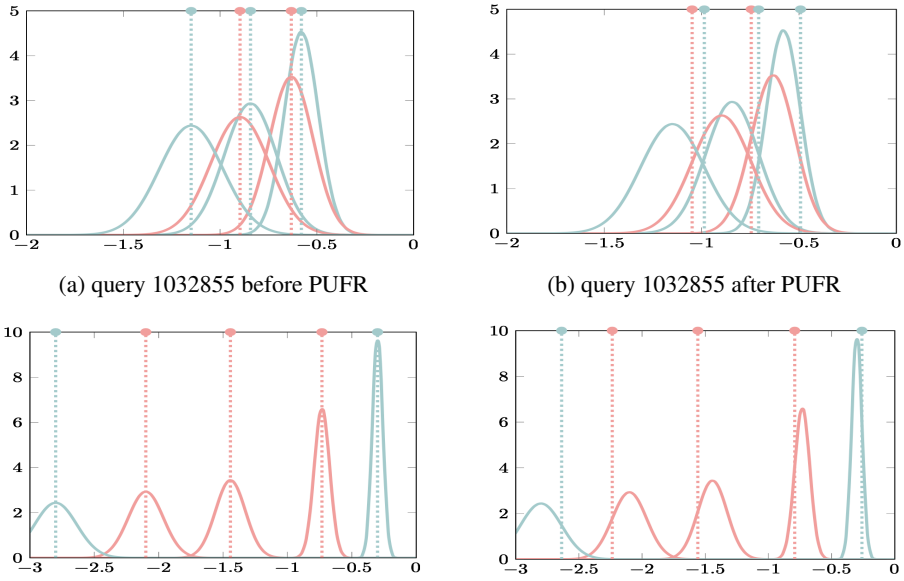
Figure 3.6: Controllability of different approaches visualized by plotting utility and fairness against the controlling hyper-parameter α on the x-axis (see Section 3.4.4 for a description of α for each approach).

there is a single re-sorting of the documents that dominates the execution time. Hence, when using PUFR in production and adjusting the score before the initial ordering of the documents, the execution of PUFR is nearly free.

3.6 Discussion

Exploiting model uncertainty for the fairness-utility trade-off. To increase the fairness of a ranking, we would commonly need to trade-off some predicted utility. Encouraging this trade-off to take place when the ranking model is less certain about the ranking scores will cause roughly equivalently relevant documents that the model cannot confidently rank, to swap place. Assuming that the ranking model is well calibrated, this might be the reason for the overall better trade-off that PUFR achieves, compared to models that do not consider predictive uncertainty. This quality is highlighted in Figure 3.7, where we show the score distribution of the top-5 documents of two queries in the $\text{MSMARCO}_{\text{Fair}}$ dataset. In the case of Figure 3.7a and 3.7b, the larger variance leads to overlapping score distribution, allowing PUFR to swap documents in the re-ranked list. On the other hand, Figure 3.7c and 3.7d show a query where the model is very certain about the order of the documents. PUFR hence does not change the order of the documents, whereas FA*IR and CVXOPT both do adjust the ranking, leading to decreased user utility for those baselines.

Using PUFR outside the models confidence. Our empirical results show that if we allow PUFR to adjust the scores too far outside of its confidence, its performance starts to decay (see Figure 3.3). If α is too high, the natural interpretation of adjusting the scores within plausible error-bounds gets lost and we cannot exploit the models knowledge of its own certainty any further. Without the certainty to back it up, PUFR becomes more arbitrary in its decisions where to trade-off predicted utility with fairness. Hence, PUFR is most effective for small values of α, roughly up to $\alpha = 4$ (see Figure 3.6). This observation means that a purely uncertainty-based fairness method might not be the best choice when the bias we want to correct for is too strong. In such cases, it might be beneficial to use uncertainty in combination with another approach that has proven effective for the task at hand.



(a) query 1032855 before PUFR

(b) query 1032855 after PUFR

(c) query 1089383 before PUFR

(d) query 1089383 after PUFR

Figure 3.7: Examples of score distributions for the top-5 documents for two queries of the MS $\text{MARCO}_{\text{Fair}}$ dataset. Protected documents in green, non-protected in red. Subfigs. 3.7a and 3.7c show the ranking score before PUFR adjusts the scores, 3.7b and 3.7d show them after. Query 1089383 was scaled before plotting.

3.7 Conclusion

We have introduced the notion of predictive uncertainty-based ranking fairness, aiming to exploit a ranking model's uncertainty as an indicator of which documents we should focus on when re-ordering for a fairer ranking which de-emphasizes documents containing biases. Through our empirical analysis we have found that the uncertainty intervals of the ranking scores are sufficiently intersecting to allow us to swap the position of some documents. We have also introduced an intuitive and principled post-processing method, PUFR, that adjusts the predicted ranking scores within some desired confidence bound. We have shown that by considering uncertainty, PUFR can achieve the best utility-fairness trade-off and has superior time complexity and good controllability.

We hope that our contribution makes the adoption of methods to remove bias in ranked results more attractive to practitioners working on real- world search and recommendation systems.

More experimentation is needed to confirm our findings in more settings. We see limitations of our approach as twofold. Firstly, PUFR allows a re-ordering of the documents only within the uncertainty of the model. This might make our method less effective in reducing unfairness when the model is very skewed towards documents containing biases. As a second limitation, we rely on uncertainty scores containing accurate information on which documents are more likely to be in the wrong order.

Furthermore, the uncertainty intervals around the scores need to intersect sufficiently. In our experiments, we are using a neural ranking model on text data, which is a task that inherently carries a fair amount of uncertainty. For other tasks and fairness definitions, more research will be necessary to evaluate whether an uncertainty-based approach can be beneficial for the utility-fairness trade-off.

As to future work, an important next step would be to define ways to evaluate uncertainty scores in a listwise manner for ranking models. Without proper evaluation of the predictive uncertainty, we are unable to put trust on the score distribution and hence on an uncertainty-based fairness approach. Moreover, more work is needed to investigate whether PUFR could be extended to, for example, Bayesian learning-to-rank models or recommender systems. Finally, we see a clear need to create more datasets for large language models with fairness labels, on which methods such as ours can be tested.

Data and code. To facilitate reproducibility of the work in this chapter, all code and parameters are shared at `https://github.com/MariaHeuss/2023-CIKM-uncertainty-based-bias-mitigation`.

Conclusion of Chapter 3

We now return to research question RQ B: "*Can we use the predictive uncertainty of the model prediction to improve ranking fairness?*". Our empirical analysis of the predictive uncertainty-based re-ranking approach PUFR demonstrates that we can reduce the number of biased documents appearing at the top of ranked lists while incurring less utility loss compared to baseline approaches that do not consider model certainty about document ordering. These findings enable us to answer the research question affirmatively: Considering predictive model uncertainty can improve the fairness and unbiasedness of ranking models.

We position this work as a proof of concept, with future research needed to investigate how effectively model uncertainty can enhance the fairness-utility trade-off across different tasks and use cases.

Part II

Explaining Advice-Giving Processes

4

RankingSHAP – Faithful Listwise Feature Attribution Explanations for Ranking Models

This second part of the thesis focuses on the interpretability of advice-giving systems. The motivation behind making model decisions more interpretable includes several potential benefits: helping developers and users identify failure cases, understanding the model's decision process to improve human-model interaction, debugging model behavior, determining when to trust model predictions, and addressing various other interpretability needs.

This chapter investigates how ranking system predictions can be explained through feature attribution explanations. Here, we refer to feature attribution as a proxy for the importance of input features to a specific model prediction. We focus on *local* or *instance-wise* explanations that are specific to individual predictions, rather than explaining the model's behavior as a whole (i.e., *global* explanations).

While extensive research on feature attribution explanations exists in other domains, the field of information retrieval (IR) has seen less progress on this topic. The challenge stems from the fact that ranking predictions take the form of ranked lists (listwise predictions) rather than single prediction values (pointwise predictions), making the application of existing methods less straightforward. In particular, one of the most popular feature attribution explanation approaches, called SHAP, has not been formally defined for listwise ranking models, leaving a gap in the toolbox available to practitioners interested in analyzing their ranking models.

To address this gap, we formalize listwise feature attribution in this chapter and define RankingSHAP, a concrete instantiation that can be flexibly adjusted to explain specific aspects of ranking decisions.

This allows us to answer the following research question:

RQ C How can we generate listwise ranking explanations for listwise ranking models?

This chapter was published as M. Heuss, M. de Rijke, and A. Anand. RankingSHAP–Listwise feature attribution explanations for ranking models. In *SIGIR 2025: 48th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 381–391. ACM, July 2025.

Note: The notation in this chapter differs slightly from previous chapters. Most importantly, we use $\tilde{R}$ for pointwise ranking models, R for listwise ranking models, and π for a specific ranked list.

4.1 Introduction

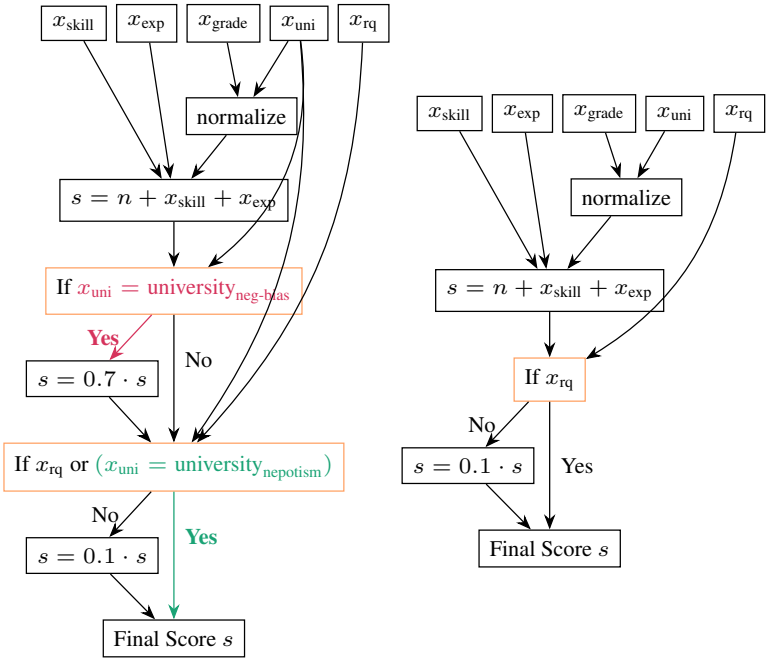
Feature attribution explanations are a posthoc family of explainability approaches that assign scores to features, quantifying their relative contribution to a model's decision. They are used to understand which features most influence the model's predictions, thereby enhancing transparency and trust. Feature attributions are among the most commonly used explanation types for posthoc explanations of trained models in general machine learning (ML) [118, 150, 171, 258].

Typical ML tasks involve pointwise prediction, explaining single classification or regression decisions. However, explaining rankings has different aspects – Why is a document relevant? (pointwise explanations), Why is one document more relevant than another? (pairwise explanations), or Why are the documents ranked in this specific order? (listwise explanations). Listwise explanations encode more context in terms of an entire or partial ranked list and are arguably more accurate/faithful since they are able to find features that affect an entire ranking. This is unlike feature attributions that focus on a single relevant document or a certain preference pair.

Feature attribution often lacks rigorous definition, beyond attributing the highest value to the *most important* feature. Limited work exists on pairwise [159] and listwise explanations [12, 137, 198, 199, 245]. Consequently, listwise feature attribution remains under-explored and in need of further theoretical underpinnings.

4.1.1 A motivating case study – Talent search

To motivate the need for tools that help practitioners arrive at a nuanced understanding of ranking outcomes, we consider talent search. There, systems use learning-to-rank to produce candidate rankings based on features like academic performance, experience, skills, and private attributes such as gender, ethnicity, and university attended. The inclusion of certain attributes in decision-making is debatable, as biases from past decisions can be reflected in the learned model and are best left to human judgment. However, sometimes these attributes are necessary for the model to perform well. Consider the two models in Figure 4.1. Both use the same features, including skills, experience, graduation grade, university, and whether the candidate meets job requirements. The right model (Figure 4.1b) uses the university reasonably by normalizing grades from different institutions, while the left model (Figure 4.1a) discriminates against candidates from certain universities and favors others. Explanations can help differentiate between such models with similar performance to identify which is less biased and more trustworthy. Feature *selection* alone may not provide sufficient insights, as it likely selects the same features (x_{uni} and x_{rq}) for both models. Instead, feature *attribution*, which assigns each feature an importance value, can identify nuanced differences in their relative importance. Furthermore, since candidate ranking scores are only meaningful relative to others, *pointwise explanations* focusing on features for high scores may not reveal



(a) Biased model (b) Unbiased model

Figure 4.1: Flow chart of a biased and an unbiased model for a talent search task. With the help of explanations we would like to be able to differentiate between the two.

the university feature as the key factor in determining the relative order for queries with candidates from universities that the model is biased against. Pairwise and listwise explanations are better suited to explain relative rankings. While pairwise explanations require a specification of the pair of candidates to compare, listwise explanations can provide insight into the model decision as a whole. We will revisit this case study in Section 4.5 to demonstrate listwise feature attribution in practice.

4.1.2 Listwise feature attribution explanations

We are interested in developing a listwise explanation method based on SHAP [136], a method inspired by Shapley values from *game theory*, that quantifies the contribution of each feature to a model's prediction. SHAP has gained significant popularity as a post-hoc explanation approach due to its theoretical properties and versatility [114]. However, SHAP only explains pointwise predictions: Given the contrastive nature of ranking tasks, listwise feature attribution would provide valuable insights into model decisions by explaining the relative order of documents, enabling comparisons across queries and ranking aspects. To address this gap, we introduce RankingSHAP, which extends SHAP to support listwise explanations while maintaining compatibility with existing research on SHAP's limitations and extensions. RankingSHAP provides flexibility in the *listwise explanation objective*, allowing users to determine feature importance for

specific ranking aspects that *faithfully* reflect the model's behavior in the context of ranked lists.

4.1.3 Approach and contributions

Our proposed method, RankingSHAP, preserves the context of ranked lists rather than evaluating documents in isolation. This contextual awareness is crucial because ranking models make decisions about relative document ordering. Therefore, a feature attribution method needs to identify a specific aspect of the model's decision to focus on and define a singular metric that quantifies changes within the ranked list with respect to that aspect. Aspects of interest may include a document's rank, measured by its shift in position, or the overall order of the top-k documents, measured by the number of permutations within the top-k. These diverse aspects underscore the need for a nuanced definition of listwise feature attribution in ranking models, which RankingSHAP provides.

We rigorously assess the faithfulness of RankingSHAP using established learning-to-rank (LtR) benchmark datasets, demonstrating its effectiveness in interpreting ranking models' outputs and providing deeper insights into their decision-making processes.

In summary, (i) we propose and rigorously define listwise feature attribution; (ii) we present a novel instantiation of our feature attribution framework called RankingSHAP; and (iii) we propose multiple evaluation schemes, white box check, preservation and deletion check for ranking feature attributions, and conduct extensive experiments to showcase RankingSHAP's performance.

4.2 Related Work

4.2.1 Shapley values and SHAP

Shapley values, originating from game theory to define a player's marginal contribution [189], are widely used in explainable AI. Efficient approximation techniques have facilitated their application in AI model decisions [206, 207]. SHAP (SHapley Additive exPlanations) [136] is one such technique, approximating the expected marginal contribution of a feature to any feature set excluding it. A comprehensive overview and recent advancements are available in [150]; we build on this work, extending it for ranking models.

Contemporaneously with the work in this chapter, Pliatsika et al. [162] propose a Shapley value-based framework for rankings and preferences, but our research emphasizes listwise explanations, unlike their document-level focus. Concurrently, Chowdhury et al. [43] establish theoretical properties for feature attribution in ranked lists and introduce a method similar to ours that satisfies these properties.

4.2.2 Explainable information retrieval

Explainable IR [12] has focused on models that are explainable by design [125, 257] and on approaches that can posthoc (after model training) explain models [197, 198, 218].

Posthoc approaches operate at the global level (model level) or at the local level (per-query). Global explainability approaches have been used to diagnose ad-hoc neural text rankers with well-understood axioms of text ranking [26, 169, 219] or to probe pre-trained transformer-based ranking models for ranking abilities [220]. We focus on *posthoc*, *local feature attributions*.

Feature selection and attribution for ranking models. Early work on interpreting ranking models was adapted for explaining query-document relevance from popular paradigms of black-box methods [136, 171] or white-box methods [190, 191, 208]. Singh and Anand [197], Verma and Ganguly [218] modify LIME [171], to generate terms as the explanation for a trained black-box ranker. Choi et al. [40], Fernando et al. [70] applied gradient-based feature attribution methods [136, 208] to interpret document relevance scores. Contrary to posthoc feature attribution approaches, local feature selection [97, 124, 125] approaches select a subset of features without distinguishing feature importance. Most work on local feature selection for rankings [97, 125] is not posthoc, and has been performed on text features, not on learning-to-rank data. In this chapter, we work on posthoc approaches for attribution and not selection.

Listwise explanations for ranking models. Typical ML tasks are pointwise prediction tasks, i.e., focusing on a single classification or regression decision. In rankings, even for a single query, we also have to deal with pairwise and listwise explanations, which might be constructed by an aggregation of decisions. There has been limited work on pairwise [159] and listwise explanations [137, 182, 198, 245]. LiEGe [245] tackles the task as text generation. Other work uses simple rankers to approximate the original ranking of a complex black-box model by expanding query terms by solving a combinatorial optimization problem [137, 198]. The work that is closest to ours, on RankLIME [42], approaches the problem with the local surrogate approach LIME, which the authors adapt for ranking models. Again, most of the approaches focus on text features and are not directly applicable to learning-to-rank models.

Explainability in learning-to-rank. Local feature selection approaches can be applied to learning-to-rank [80, 85, 163]. Among the feature-selection approaches, filter methods are model-agnostic [80], while wrapper methods are designed for a particular type of model [85]. In the context of ranking, some work produces local feature selections [163, 196]. Singh et al. [200] propose the notions of validity and completeness based on the information contained in the explanation. While these notions are useful in both conception and evaluation of explanations, they still view the explanation as a *selection* of features. Feature selection methods, however, lack the capability to differentiate between features of varying importance, thereby avoiding a nuanced understanding of which features are substantially more critical in the decision-making process. We focus on feature attributions.

4.2.3 Faithfulness in explainable AI

Faithfulness measures how accurately an explanation represents the reasoning process behind a model's prediction [100]. Evaluating faithfulness is challenging because the model's actual reasoning cannot be directly observed. Hence, various definitions and evaluation frameworks for faithfulness have been proposed [100, 139]. While there

is no clear agreement as to what notion or framework should be used to measure and establish faithfulness [139], there are two dominant frameworks in explainable IR [12]. When locally approximating a ranking model with a proxy model, faithfulness is the degree to which the proxy model approximates the original ranking [137, 198]. An alternative notion of faithfulness is based on an *information-theoretic* notion of feature importance [200, 232]. There, faithfulness refers to the predictive power of the features in the attribution. Specifically, if a feature set is important then masking off or removing the non-relevant features should not result in a big change in model output. While both notions model different aspects of faithfulness, in this chapter we follow the latter framework.

4.3 Feature Attribution for Pointwise Rankers

Early work on local feature explanations has introduced the concept of feature attribution [206]; recent work often lacks a clear definition of what makes a feature *important*, causing ambiguity in evaluating attribution faithfulness. Despite attempts to formalize feature attribution [5], these efforts have not been widely adopted, resulting in inconsistencies and confusion in the field [114]. We build on [136] to define *pointwise feature attribution* for black-box models with one-dimensional model output such as a pointwise ranking model

$$\tilde{R} : \mathcal{D} \to \mathbb{R}, x_{q,l} \mapsto s_{q,l}, \tag{4.1}$$

that predicts the ranking scores $s_{q,l} \in \mathbb{R}$, representing the probability of relevance, for the feature vectors of each document-query pair, $x_{q,l} \in \mathcal{D}$ in the space of all documents $\mathcal{D}$. We consider *instance-wise* feature attribution explanations that assign to each feature i an attribution value $\phi_i(x, \tilde{R})$, directly reflecting the importance of the feature to the model decision for instance x. Hence, *feature attribution explanations* can be understood as dictionaries $\{i \mapsto \phi_i(x, \tilde{R})\}_{i=1,\dots,n}$ containing exactly one attribution value per feature. A well-defined, instance-specific definition of feature attributions should consider the specific combinations of feature values in the input that collectively lead the model to predict a high score. Also, features with greater importance for the prediction should have higher attribution values.

We use marginal contributions to define *pointwise feature attribution*.[1] Our definition is based on SHAP [136]. In Section 4.4, we extend this to *listwise feature attribution* and define RankingSHAP to approximate feature attribution for listwise rankers.

Definition 4.3.1. We define the *attribution* or *importance* of a feature j in terms of marginal contributions. Let $n = dim(\mathcal{D})$ be the input space dimension, and let a *coalition* be a subset $S \subset \{1, \dots, n\} \setminus j$ of the input features excluding j. To measure the marginal contribution of feature j to coalition S, we compare the model output when shown only features in S to the output when shown features in $S \cup \{j\}$. Since we cannot simply erase features, we mask them with samples from a set of feature-vectors $B \subset \mathcal{D}$, called *background data*, which ideally summarizes the data distribution. For masking, we use templates defined by subsets S, indicating the presence $(i \in S)$ or

[1] For a detailed discussion of marginal contributions, see [150].

absence ($i \notin S$) of a feature, and data-points from the background data $b \in \mathcal{D}$. We define $m_{S,b} : \mathcal{D} \to \mathcal{D}$ as:

$$m_{S,b}(x)_i = \begin{cases} x_i, & \text{if } i \in S \\ b_i, & \text{if } i \notin S. \end{cases} \tag{4.2}$$

The marginal contribution of feature j to coalition S for vector b is:

$$\tilde{R}(m_{S\cup\{j\},b}(x)) - \tilde{R}(m_{S,b}(x)). \tag{4.3}$$

We define the **pointwise feature attribution** of feature j to the model decision of $\tilde{R}$ at input x as the expected marginal contribution of feature j to all possible coalitions of features:

$$\phi_j(x, \tilde{R}) = \sum_{S\subset\{1,\dots n\}\setminus j} w_S \cdot \mathbb{E}_{b\sim B}[\tilde{R}(m_{S\cup\{j\},b}(x)) - \tilde{R}(m_{S,b}(x))],$$

with weighting factor $w_S = \frac{1}{n!}|S|!(n - |S| - 1)!$ and uniform sampling from B.

Computational costs. Given the exponential growth of coalitions with the number of features and the need for numerous background examples for a good summary, we approximate pointwise feature attribution using sampling. Following [136], we use SHAP for this approximation. Even though we are approximating the attribution values, SHAP is known to be computationally expensive, especially for high feature dimensions. There have been advances to making the sampling more efficient [102, 255]. Also, since pointwise explanations are usually used as an analysis tool for specific input examples rather than to analyze the whole corpus, it remains a broadly used explanation approach [114, 150] despite its computational costs.

4.4 Feature Attribution for Listwise Rankers

For many machine learning tasks, SHapley Additive exPlanations (SHAP) [136] effectively approximate feature attribution values for individual model decisions, such as regression scores or classification probabilities. However, applying this method to listwise ranking models is challenging because these models output a ranked list rather than a single score. Within this ranked list, different decisions are made regarding the order of individual documents. Pointwise SHAP is only defined for a single one-dimensional model output. While it can explain the model score of an individual document, it does not consider the context of other documents in the list. In this chapter, we extend SHAP to an approach that caters to listwise ranking decisions, called RankingSHAP.

Instead of looking at pointwise ranking models, as we did in Section 4.3, we consider a listwise ranking model

$$R : \{\mathcal{D}_q\}_q \to \text{Sym}, \{x_{q,j}\}_j \mapsto \pi_q \tag{4.4}$$

that maps a set of candidate feature vectors for query q, $\mathcal{D}_q = \{x_{q,j}\}_j$, to some permutation matrix $\pi_q \in \text{Sym}(\mathcal{D}_q)$ representing the ranked list in the Symmetry group of all permutations of the candidate set $\mathcal{D}_q$.

We define two components, *listwise masking* and *listwise explanation objectives* that enable us to establish listwise feature attribution for ranking models, which we will introduce in Section 4.4.1. In Section 4.4.2, we formally define RankingSHAP for approximating listwise attribution values. We define RankingSHAP as a wrapper around SHAP using those two components. We deliberately chose not to modify SHAP's internal algorithm, allowing us to leverage the extensive literature on SHAP directly. Finally, we examine listwise explanation objectives with examples in Section 4.4.3.

4.4.1 Feature attribution for ranking models

Our definition of feature attribution/feature importance for ranking models consists of two parts: (i) Define how masking applies to each document in the ranking $\mathcal{D}_q$ for query q. And (ii) measure the impact of input changes on the model decision, quantified by a single number.

Masking the inputs of a ranking model. We apply a listwise mask $m_{S,b}$ to all documents $\{x_{q,j}\}_j$ in the ranking: $m_{S,b}(\mathcal{D}_q) = \prod_{x_{q,j} \in \mathcal{D}_q} m_{S,b}(x_{q,j})$. By masking the feature vector $x_{q,j}$ of each document with the same mask $m_{S,b}$, we disregard the impact of the masked features on the ranking decision. This helps identify the contributions of non-masked features to the document ordering.

Reducing the model prediction to a single prediction value. Feature attribution is defined by the expected change in the predicted score. We need to reduce the ranking model's decisions to a single value reflecting the change for a perturbed input sample, using a listwise explanation objective that takes a ranked list and maps it to a value, highlighting some property of the ranked list that we want to investigate.

One example for such a function is a rank similarity coefficient like Kendall's tau τ [111], which is commonly used in the interpretability literature to measure rank correlation [137, 198, 200]. By comparing the change in the relative order of documents, we can measure how much the prediction deviates from the optimal order π_q predicted by the model:

$$g_q(\tilde{\pi}) = \tau(\pi_q, \tilde{\pi}). \tag{4.5}$$

For any such *listwise explanation objective* g_q, we define feature importance through the composition with the original ranking model, $g_q \circ R$. Section 4.4.3 provides further examples.

In summary, we have defined how to "remove" a feature from the model input through masking and measure its impact on the model prediction with a single value. This allows us to determine the **listwise feature attribution** using Section 4.3.

4.4.2 Estimating listwise feature attribution with RankingSHAP

With the definition of feature attribution for ranking models, we introduce Ranking-SHAP. This depends on the choice of listwise explanation objective g and aims to explain which features are important for specific aspects of the ranked list. The ability to focus on different aspects of the ranking decision allows RankingSHAP to provide contrastive and flexible instance-wise explanations for rankers.

Algorithm 5 Adjusted model prediction (used in combination with SHAP)

Require: ranking-model R, feature-vectors $\mathcal{D}_q$ for query q, listwise explanation objective g,
Input: masking function $m_{S,b}$
1: **for all** $x_j \in \mathcal{D}_q$ **do**
2: $\quad \tilde{x}_j \leftarrow m_{S,b}(x_j)$
3: **end for**
4: $\pi \leftarrow R(\{\tilde{x}_j\}_j)$
5: $v \leftarrow g(\pi)$
6: **return** v

Following the definition of feature attribution with simultaneous masking of document vectors and a listwise explanation objective, we establish RankingSHAP as a wrapper around SHAP to approximate the marginal contribution of each feature in a ranking model, leveraging prior work.

SHAP samples both coalitions (templates for creating masks) and background data to generate masked perturbations (see Eq. 4.2) of the input, approximating the marginal contribution of a feature to any coalition. Given a sampled mask $m_{S,b}$, we illustrate how RankingSHAP adjusts the model prediction for use with SHAP in Algorithm 5. We loop over all documents $x_j \in \mathcal{D}_q$ (lines 1–3) and perturb the document features with the mask to get $\tilde{x}_j = m_{S,b}(x_j)$. Then, we rank the perturbed feature vectors with the ranking model $\pi = R(\{\tilde{x}_j\}_j)$ (line 4). Finally, we apply the listwise explanation objective $v = g(\pi)$ to measure the change in output according to the specified explanation objective (lines 5 and 6).

Computational costs. Our approach allows for the use of existing SHAP implementations. This also means that it inherits any limitation that SHAP has such as the computational complexity. Nevertheless, it does not introduce any significant new additional computational overhead and allows us to use prior research on SHAP extensions and improvements for ranking without adjustments, such as advances in improving efficiency. Since SHAP is a commonly used explanation approach for pointwise predictions, we do not expect the computational complexity of RankingSHAP to hinder it's adoption in practice.

4.4.3 Listwise explanation objectives

We provide examples of listwise explanation objectives to illustrate the types of contrastive explanations RankingSHAP can generate.

Emphasizing top-ranked documents. Instead of focusing on the entire ranked list, we can emphasize the top-k documents to identify features crucial for their high ranking. For example, we demonstrate RankingSHAP using a weighted rank difference objective with common position weighting:

$$g_q^w(\tilde{\pi}) = \sum_{d \in D_q} \frac{\text{rank}(d|\tilde{\pi}) - \text{rank}(d|\pi_q)}{\log_2(\text{rank}(d|\pi_q))}. \tag{4.6}$$

Explaining feature importance of a singular document. This objective focuses on one particular document d, investigating which features contribute, or would contribute, most to its high ranking compared to others when only a subset of features is considered. This can be implemented using the negative rank[2] of that document:

$$g_q^{\mathrm{rank}(d)}(\tilde{\pi}) = -\operatorname{rank}(d|\tilde{\pi}). \tag{4.7}$$

Alternatively, we can use RankingSHAP to determine the features that are the most beneficial for the document's exposure:

$$g_q^{\exp(d)}(\tilde{\pi}) = \exp(\operatorname{rank}(d|\tilde{\pi})) = 1/\log_2(\operatorname{rank}(d|\tilde{\pi})). \tag{4.8}$$

Explaining the position of a group of documents. RankingSHAP allows us to compare ranking decisions for two groups of documents. We can consider the relative ordering or absolute distance of members of the different groups. Future work could explore explaining model fairness or identifying biases using listwise feature attribution.

4.5 Talent Search: A White Box Example

To demonstrate the application of RankingSHAP and to evaluate the feature attributes generated by different explanation approaches, we create a synthetic example, revisiting the talent search case study from the introduction. We design an interpretable model to estimate the importance of features for various model decisions. This evaluation framework, known as a "White Box Check," is widely used in the explainability community for other ML tasks [152].

In the following sections, we define features and ranking model that we will use as white box in Section 4.5.1. We then describe the experimental setup in Section 4.5.2 and examine various queries modeling different types of model decisions in Section 4.5.3. These queries demonstrate the practical use of listwise feature attribution and qualitatively evaluate three feature explanation approaches. In Section 4.5.4, we show how to use RankingSHAP to zoom in on individual documents and compare it to a pointwise explainer. We conclude with a detailed discussion in Section 4.5.5.

4.5.1 Model design

We design a model using 5 features indicating whether a candidate meets general job requirements, the university the candidate graduated from, skill and experience levels, and average graduation grade. This model ranks candidates for various academic degree-required scenarios, aiming to mimic biases in trained models.

Detailed feature information is in Table 4.1. The model favors candidates from $\mathbf{uni}_{\textbf{nepotism}}$ and disadvantages those from $\mathbf{uni}_{\textbf{neg-bias}}$. A flowchart is in Figure 4.1a in Section 4.1. The ranking score is determined as follows:

[2]We use the negative rank to maintain consistency with higher values being more desirable, explaining why a document ranks high (low rank) rather than low.

Table 4.1: Candidate evaluation criteria for running example

Feature	Description
Requirements	Binary value $x_{\text{rq}} \in \{\text{T}, \text{F}\}$ indicating if the candidate meets the job's minimum requirements.
Experience	Relevant work experience on a scale $x_{\text{exp}} \in [0, 1]$ (1=extensive experience, 0=none)
Skills	Skill fit on a scale $x_{\text{skill}} \in [0, 1]$, (1 = perfect match, 0 = no relevant skills)
University	Institution where the candidate obtained their degree, x_{uni}.
Grades	Mean graduation grade, x_{grade}, with range depending on the university.

- Normalize the grade $\textbf{norm}(x_{\text{grade}}, x_{\text{uni}})$, scaling it so that the minimum possible grade is 0 and the maximum is 1, to make grades from universities with different grading schemes comparable.
- Calculate the sum of x_{skill}, x_{exp}, and $\text{norm}(x_{\text{grade}}, x_{\text{uni}})$.
- For candidates from $\textbf{university}_{\textbf{neg-bias}}$, apply a negative bias by multiplying the score by 0.9.
- If the candidate does not meet the job requirements, multiply the score by 0.25, effectively placing them at the bottom of the list. Candidates from $\textbf{university}_{\textbf{nepotism}}$ are exempt from this penalty.

Candidates are ranked by their scores, with the highest at the top. We then investigate different queries with RankingSHAP to identify biases and compare attribution values to other explanation approaches.

4.5.2 Experimental setup

The main goal of this Section is to showcase the usage of RankingSHAP and demonstrate the need for listwise, as opposed to pointwise, explanations and feature attribution rather than feature selection. Therefore, we compare RankingSHAP to the pointwise SHAP explainer, **PointwiseSHAP** (averaged over all candidates), as well as to the **Greedy** feature selection approach from [199]. The latter iteratively adds features to an initially empty set based on their marginal contribution to the Kendall's tau objective from Eq. 4.5 until the contribution becomes non-positive or the explanation size reaches 2. Section 4.6 contains a more complete empirical comparison with a comprehensive set of baselines, including RankLIME [42] and ShaRP [162]. For background data, we sample 100 candidates from uniform distributions over the possible feature values defined in Section 4.5.1. Detailed feature values and candidate lists for each query are provided in Appendix 4.A.

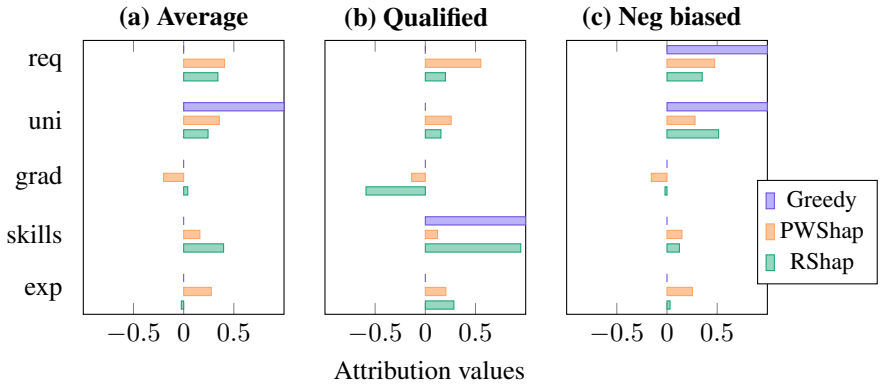


Figure 4.2: Feature attribution values for different query scenarios from Section 4.5.3.

4.5.3 Listwise evaluation across query scenarios

We define scenarios to demonstrate feature attribution for contrastive ranking explanations and evaluate them. We present 5 query scenarios: three in the main body and two in Appendix B.[3] We discuss the setup, candidate constellation, estimated feature importance $imp_{feature}$ for some features on the overall ranking, and evaluate the explanation approaches. In this part of our analysis RankingSHAP uses Kendall's tau explanation objective from Eq. 4.5 to explain the overall order of the candidates.

Average query

Description. This query includes candidates from universities with the same grading scheme, only some meeting the requirements, but none from $\text{university}_{\text{neg-bias}}$ or $\text{university}_{\text{nepotism}}$.

Importance. Since no exceptions for candidates from biased institutes apply and grades are within the same scheme, we expect imp_{rq} to be high, as hiding this feature could change the ranking significantly. We also expect imp_{uni} to have a positive but smaller value since a change of university for all candidates causes ambiguity for the evaluation of the grade.

Evaluation of Feature Attributes. Figure 4.2(a) shows attribution values/selected features (bars with length 1). Both RankingSHAP and PointwiseSHAP identify x_{rq} as an important feature and assign a positive value to x_{uni}. The greedy feature selection approach only selects the university feature.

Qualified query

Description. Similar to the average query, but only candidates meeting the requirements. The model can ignore x_{rq} without bias.

[3] An extended appendix including these additional results is available at `https://github.com/MariaHeuss/RankingShap/blob/main/Paper_RankingSHAP.pdf`.

Importance. While imp_{uni} should still be assinged a positive value, imp_{rq} should be assigned a lower value than before as x_{rq} is irrelevant for these candidates.

Evaluation of the feature attributes. Figure 4.2(b) shows that Greedy and RankingSHAP correctly assign a low value to the x_{rq}. PointwiseSHAP is not able to identify that the feature that is most important for attaining a high ranking score for each individual document, x_{rq}, is not important for this specific query. Furthermore, we notice that RankingSHAP assigns higher values to other features, that are now important to distinguish between the candidates.

Negative bias query

Description. Similar to the average query, but with an additional candidate from university$_{\text{neg-bias}}$ having the best overall profile. The model has a negative bias towards this university.

Importance. We expect imp_{uni} to be higher due to the bias.

Evaluation of the feature attributes. In Figure 4.2(c), both RankingSHAP and Greedy are able to identify the negative bias towards one candidate by correctly assigning a higher attribution value to x_{uni} than for the average query, while PointwiseSHAP is not.

4.5.4 Highlighting feature importance for the rank of individual documents

In this section we zoom in on individual documents and the role of different features on the placement of that documents. For this analysis we use the exposure-based explanation objective from Eq. 4.8, highlighting the impact that the different features for the ranking model have on the exposure of the individual candidates. We compare to the attribution values generated by PointwiseSHAP for the specific document in question. We investigate two of the scenarios in more detail, the results for the other scenarios can be found in Appendix B.[4] Claims made in this subsection on the relative qualities of the candidates can be confirmed with Table 4.A.1 in Appendix 4.A.

Qualified query

Since the university and requirements are the same for all candidates, a recruiter might be interested in which features were particularly important for ranking them. RankingSHAP provides more contrastive insight into the strengths of a document than PointwiseSHAP. For example, RankingSHAP highlights the skill feature as negatively impacting the third candidate's exposure. If a recruiter is more interested in grades, Figure 4.3(a) allows them to make an informed decision to invite the candidate regardless of the model prediction. In contrast, PointwiseSHAP provides similar attribution values for each candidate and does not highlight the grades of the third-ranked candidate as a redeeming quality.

[4] An extended appendix including these additional results is available at `https://github.com/MariaHeuss/RankingShap/blob/main/Paper_RankingSHAP.pdf`.

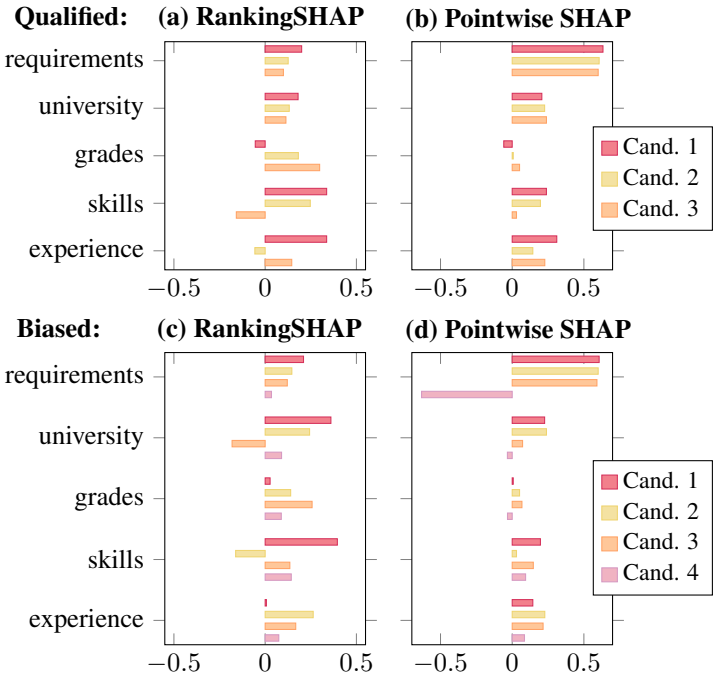


Figure 4.3: Feature attribution values, for RankingSHAP with the $g_q^{exp(d)}$ exposure objective defined in Section 4.4.3 and Pointwise SHAP for individual candidate in the ranked list.

Biased query

The listwise feature attribution analysis of RankingSHAP from Figure 4.2 shows high importance of the university feature for this query, warranting further investigation. Figure 4.3(c) and (d) demonstrate that RankingSHAP can identify the unfair treatment of the third-ranked candidate due to their university, unlike PointwiseSHAP.

4.5.5 Discussion

The contrastive use of feature attribution. We define the estimated feature importance used in this section's evaluation in a contrastive way, comparing them to other queries as well as to other explanation objectives. Prior work [150] suggests that attribution values are hard to interpret in isolation; contextualizing them with other model decisions aids understanding. The use of different explanation objectives makes feature attribution particularly effective for ranking models: since a model decision involves a complex interplay of various decisions about the relative ordering of documents, contrasting different aspects of the decision allows us to uncover nuances that led to a specific model decision.

Using RankingSHAP to identify biases. By comparing attribution values of different queries, we can identify instances where a feature expected to be of moderate importance,

such as x_{uni}, impacts the decision more than anticipated. For example, in the biased query, we can detect hints of bias in the explanations in Section 4.5.3. Zooming in on what features are most important for the model to provide the individual candidates with exposure in Section 4.5.4, we see that RankingSHAP identifies the candidate that got negatively effected by the model bias, as well as qualities that might still speak for them.

Pointwise vs. listwise ranking explanations. From our synthetic example we see that simply using a pointwise explanation approach to explain listwise ranking decisions fails to consider interactions between the feature values of different documents. Features that are important for a high ranking score are assigned a high attribution value, independent of whether they are important for the relative ordering of the list.

Selection is not attribution. While feature selection can be a useful tool for understanding ranking models, more nuanced explanations are sometimes necessary to interpret model decisions. Even if the selection approach correctly identifies the most important features, a feature attribution approach is needed to gain detailed insight into the relative importance of the features impacting for example model bias.

Limitations of white box check evaluation. We acknowledge the limitations of the qualitative evaluation in this section due to the subjective nature of estimated importance, the synthetic experiment setup, and the limited number of queries investigated. Nevertheless, this section is crucial for providing insights into using listwise feature attribution methods like RankingSHAP. To complement this qualitative evaluation, we will quantitatively compare RankingSHAP to a broad range of baselines in Section 4.6.

4.6 Quantitative Feature Attribution Evaluation

The quantitative evaluation of explanations is a difficult task [135]. In contrast to usual machine learning tasks, where labeled data to benchmark different models can be used for the evaluation, for explanations there is nothing like a *ground truth explanation*. Evaluating feature attribution values in particular is challenging, leading to prior work on evaluating feature attribution often defaults to evaluating the feature selection of the top-k features instead [176]. We will follow this strategy, by defining Preservation and Deletion Checks [152] for listwise explanations. We pose the following two research questions on the correctness/completeness of the explanations: (**RQ4.1**) Are explanations generated with RankingSHAP faithful to the model decision in terms of overall order of the documents? And (**RQ4.2**) Can RankingSHAP identify features responsible for the distribution of exposure in the ranked list? We describe our experimental setup in Section 4.6.1, our evaluation framework in Section 4.6.2, and our experimental results in Section 4.6.3.

4.6.1 Experimental setup

Datasets

Following [199] we consider two datasets from LETOR4.0 [165]. MQ2008 consists of 800 queries with pre-computed query-document feature vectors of dimension 46.

The MSLR data set consists of 10k queries with query-document feature vectors of dimension 136. For both, we use the train-val-test split of fold1 and evaluate the explanations on the test data.

Ranking model

We use the LightGBM [110] to train a listwise ranker with LambdaRank, using NDCG as metric.

Listwise explanation objectives

To provide additional evidence for the flexibility of RankingSHAP we use two different explanation objectives: **RShapK** uses Kendall's tau objective from Eq. 4.5 to identify features important for the overall ordering of candidate documents. **RShapW** employs the weighted rank difference objective g^w from Eq. 4.6 to prioritize documents ranked higher by the model.

Baselines

We consider the following baselines:

Random: Random feature attribution, normalized.

PWSHAP Previously used as a baseline in [199], we take the mean over the pointwise SHAP values of the top-5 documents.

PWLime: The mean over the pointwise attribution values generated with LIME of the top-5 documents.

Greedy: A greedy feature selection approach from [199]. The authors iteratively add features with the biggest marginal contribution to the initially empty explanation set until a set size of k is reached.

RLime: Listwise LIME for rankers, inspired by RankLIME [42]. Perturbation is done on each feature of each document independently. Since we are interested in listwise explanations, we report the mean of feature attribution values over all documents.

ShaRP As discussed in Section 4.2, parallel to the work in this chapter, Pliatsika et al. [162] generate feature attribution explanations with SHAP for input features of individual documents, rather than the ranked list as a whole. We use the "Rank Quantity of Interest" for our implementation as it is closest in idea to our Kendall-tau based implementation of RankingSHAP. We use the mean of the individual document explanations to get listwise explanations.

Implementation details

All approaches, except Random, use background data for masking or perturbing input features. For MQ2008, we sample 100 random samples from the training data; for MSLR10k, we sample 20 to compensate for higher feature dimensions. For evaluation,

we sample a different set of 100 background samples for both datasets. We use the KernelSHAP implementation from the SHAP library [136] for RankingSHAP, PWShap and ShaRP and the TabularExplainer from the LIME library [171] for PWLime and RLime, all with default settings.

4.6.2 Experimental evaluation

Due to the lack of ground truth attribution values and evaluation frameworks for rankers, we use the deletion and preservation check strategy [152] from other machine learning tasks, adapted for ranking. A good explanation should replicate the original model output when non-explained features are masked (Preservation check) and significantly alter the output when important features are removed (Deletion check).

Both checks measure the impact of masking features on the model output, evaluated by a function v. We sample masking values b from background data B to substitute for non-explained features, resulting in re-ranked lists $\tilde{\pi}_{e,b}$:

$$\mathrm{Preservation(e)} = \mathbb{E}_{\mathrm{b}\sim\mathrm{B}}[\mathrm{v}(\tilde{\pi}_{\mathrm{e,b}})]\,.$$

Similarly, the deletion check applies the mask to the features included in the explanation.

For ranked list outputs, we use Kendall's similarity τ with the original ranked list π, hence $v^{\tau}(\tilde{\pi}_{e,b}) = \tau(\pi, \tilde{\pi}_{e,b})$. These checks align with the validity and completeness criteria in [199]. Additionally, we evaluate the alignment of the generated explanations with the original model by measuring the exposure difference between each candidate ranked with the original input and the masked input: $v^{\text{exp-diff}}(\tilde{\pi}_{e,b}) = \sum_{d\in\pi} |\exp(\mathrm{rank}(d|\pi)) - \exp(\mathrm{rank}(d|\tilde{\pi}_{e,b}))|$. We conduct evaluations at explanation sizes of 1, 3, 5, 7, and 10 and report the mean values over all evaluated queries.

Note that in this approach, we evaluate feature selection explanations as subsets of features, not attribution values. For feature attribution explanations, we use the top-k features.

4.6.3 Results

The results with the deletion and preservation checks are presented in Figure 4.4.

(RQ4.1) Are explanations generated with RankingSHAP faithful to the model decision in terms of overall order of the documents? To address this research question, we evaluate the *correctness* (how well the explanation aligns with the model's decision) and *completeness* (how much relevant information is captured in the features with the highest attribution values) of the explanations. The preservation check with rank-similarity measures how well the ranked list can be reconstructed using only the most important features identified by each explanation approach. As shown in Figure 4.4 (a), only the Greedy baseline outperforms RankingSHAP, which is expected since Greedy is designed to maximize this metric through feature selection explanations. Conversely, the deletion check (b), which involves removing the features with the highest attribution values, reveals that RankingSHAP outperforms all baselines, including the Greedy and all pointwise baselines. These findings are consistent for the MSLR dataset, as illustrated in Figure 4.4 (c) and (d). Overall using an explanation size of 10 features, we

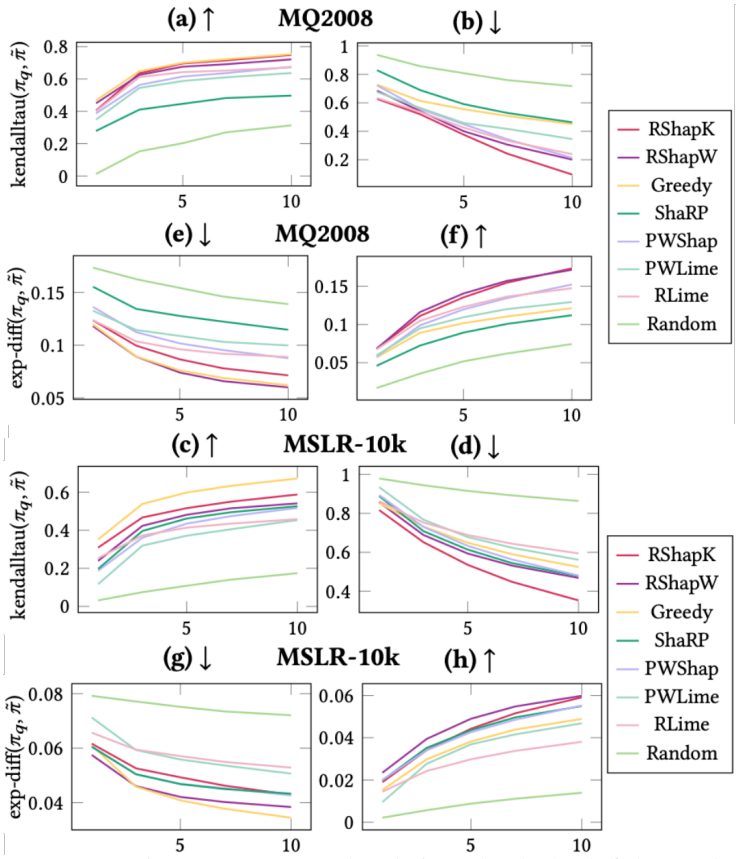


Figure 4.4: Preservation (a, c, e, g) and Deletion Check (b, d, f, h). Only features top-k of the explanations are kept/ masked. For the Kendall τ measure, higher numbers represent higher similarity with the original rank, so for the Preservation check higher is better while for the Deletion check lower is better. For the exposure-base measure, it is exactly the other way around since lower numbers represent exposure closer to the original one.

achieve approximately 0.7 rank similarity for the MQ2008 data and 0.6 rank similarity for the MSLR-10k data. In contrast, the rank similarity drops to less than 0.2 and 0.4, respectively, when removing these 10 features with the highest attribution values from the model input. Thus, we answer our first research question in the affirmative: RankingSHAP is capable of faithfully explaining the model decision.

(RQ4.2) Can RankingSHAP identify features responsible for the distribution of exposure in the ranked list? We compare explanation approaches using the Preservation Check (Figure 4.4 (e) and (g)) and the Deletion Check (Figure 4.4 (f) and (h)), alongside the exposure difference metric $v^{\text{exp-diff}}$ from Section 4.6.2. The Preservation Check indicates that the exposure difference decreases for all explanation approaches as the explanation size increases. RankingSHAP and the Greedy approach perform best

in the Preservation Check, reducing the exposure difference by 1/2 to 1/3 compared to the random baseline. In the Deletion Check, RankingSHAP clearly outperforms all other approaches, producing an exposure difference 3 to 5 times greater, depending on the dataset, when the most important features identified by RankingSHAP are omitted, as opposed to random features. These findings provide evidence that RankingSHAP effectively identifies features responsible for the distribution of exposure in a ranked list, thus positively answering the research question.

4.6.4 Reflections

Using different explanation objectives for focusing on different aspects of the ranking decision. The performance difference between the two versions of RankingSHAP, each with distinct explanation objectives, highlights RankingSHAP's ability to emphasize different aspects of the ranked list for specialized explanations. A listwise similarity objective, like Kendall's tau in RShapK, identifies features critical for the overall ranking. Conversely, an objective like the weighted rank difference in WShapK focuses on the top of the ranked list, improving faithfulness for top documents, as evidenced by exposure-based evaluation. Hence, when using RankingSHAP for generating ranking explanations, it is crucial to carefully consider which aspects of the ranking decision should be elucidated.

Using SHAP advances in RankingSHAP for enhanced interpretability. Since we define RankingSHAP as a wrapper around SHAP, it is possible to apply improvements developed for SHAP to RankingSHAP. This allows for the use of numerous advances in the field, such as handling correlated features [1], increasing the efficiency of SHAP [102, 107], and making adjustments to the sampling of background data [83], or the weighting of different coalitions when calculating SHAP values [118]. Some of these advances can be applied directly to RankingSHAP, although future research will need to investigate how easily transferable these improvements are to the ranking task.

4.7 Conclusion

In this chapter, we have defined the concept of listwise feature attribution for ranking tasks, allowing flexible and contrastive examination of ranking decisions through a listwise explanation objective. We show that our proposed approach RankingSHAP results in delivering faithful feature attributions and RankingSHAP can aid in meaningfully understanding model decisions and detecting biases.

However, we note that RankingSHAP has limitations, including high computational costs for high-dimensional input spaces and the challenge of interpreting SHAP values, which may not always align with human expectations [116], potentially lacking contrastiveness [148], and it can be susceptible to adversarial attacks [201]. Additionally, SHAP assumes uncorrelated features, leading to unrealistic out-of-distribution data if ignored [1]. Some of these limitations have been addressed in prior literature, and due to RankingSHAP's structure as a SHAP wrapper, these improvements could potentially be applied to RankingSHAP (see Section 4.6.4).

For future work, we see the need for a more thorough evaluation framework that goes beyond faithfulness. Furthermore, future research should examine whether using listwise SHAP attribution values in a contrastive manner can bridge the gap between mathematically well-defined explanations and practical applications in real-life scenarios.

Data and code. To facilitate reproducibility, code and parameters are available at `https://github.com/MariaHeuss/RankingShap`.

Conclusion to Chapter 4

This chapter addresses research question RQ C: "*How can we generate listwise ranking explanations for listwise ranking models?*". We have formally defined listwise feature attribution by focusing on one specific aspect of ranking decisions at a time. By implementing RankingSHAP as a wrapper around SHAP, a well-established feature attribution approach in other domains, we develop a Shapley value-based method for generating listwise ranking explanations. Our approach uses the listwise explanation objective to explain specific aspects of the ranking decision, thereby providing an answer to research question RQ C: We can use RankingSHAP to generate listwise ranking explanations for listwise ranking models.

4.A Appendix A

Here we include the explicit set-up of the simulated example from Section 4.5. In Table 4.A.1 we give an overview over all candidates that were used for the different query scenarios. The different universities have different grading schemes, which the

Table 4.A.1: Feature values for the individual candidates.

candidate	experience	skills	grades	university	req
qual-1	0.8	0.55	3.5	uni_{us}	True
qual-2	0.7	0.75	3.3	uni_{us}	True
qual-3	0.9	0.8	3	uni_{us}	True
non-qual	0.7	0.7	3	uni_{us}	False
privileged	0.8	0.6	3.6	uni_{nep}	False
qual-net	0.7	0.9	8	uni_{net}	True
qual-ger	0.8	0.8	1	uni_{ger}	True
qual-biased	0.8	0.7	3.6	uni_{bias}	True

models from Figure 4.1 depends on. Table 4.A.2 shows an overview over the different universities that are used in the query scenarios. We show the best possible and the worst passing grade as well as whether the biased model is biased towards the university in question. Those candidates were then used for different queries. Which candidates

Table 4.A.2: Comparison of grading schemes and model bias across universities.

university	highest grade	lowest grade	model bias
uni_{us}	4	1	None
uni_{nep}	4	1	Positive
uni_{bias}	4	1	Negative
uni_{ger}	1	4	None
uni_{net}	10	6	None

were used for what queries can be found in Table 4.A.3. The table entries indicate the rank of the candidate for the biased ranker, with 0 indicating that they were not included.

Table 4.A.3: Query-candidate matrix - numbers indicate the rank for the biased ranker, 0 that they were not considered.

candidate	average	nepotism	qualified	internat.	biased
qual-1	2	3	3	0	3
qual-2	1	2	2	0	2
qual-3	0	0	1	2	0
non-qual	3	4	0	4	4
privileged	0	1	0	0	0
qual-net	0	0	0	3	0
qual-ger	0	0	0	1	0
qual-biased	0	0	0	0	1

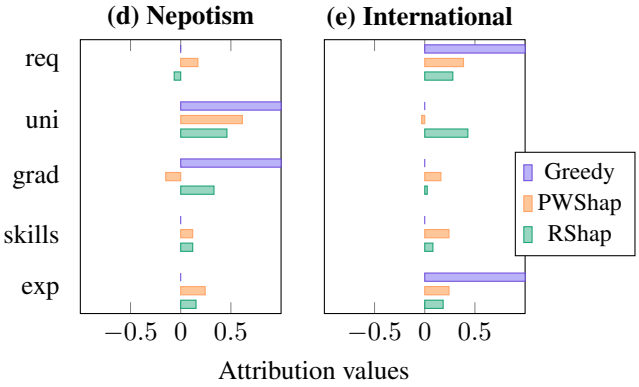


Figure 4.B.1: Feature attribution values for different query scenarios from Section 4.5.3.

4.B Appendix B – Simulated Experiment – Additional Results

Here we present additional results for the simulated experiment for some more query scenarios, as well as for the unbiased model from the flowchart in Figure 4.1b.

4.B.1 Additional query scenarios

Nepotism query

Description. For this query, one additional candidate from university$_{\text{nepotism}}$ with good records for x_{skill}, x_{exp} and x_{grade} is considered, but lacking some of the job requirements.

Importance. As we know, the model has picked up on a bias in the data, favoring candidates coming from university$_{\text{nepotism}}$, which coincidentally or not is the same university that some people that made past hiring decisions graduated from. Hence, for this query we estimate imp_{rq} to take a smaller value, and imp_{uni} to take a higher importance value.

Evaluation of the feature attributes. In Figure 4.B.1(d) we see that all approaches

correctly pick up on the bias towards university$_{\text{nepotism}}$ by assigning a high value to x_{uni}, while assigning a low value to/ not selecting the usually important x_{req}.

International query

Description. This query considers candidates from universities with different grading schemes. Most candidates meet the job requirements, and none are from university$_{\text{nepotism}}$ or university$_{\text{neg-bias}}$

Importance. For this query we estimate imp_{uni} to take a higher value than for the average query. Since candidates from universities with different grading schemes are compared, knowing which university the candidate went to is important for the interpretation of the grades.

Evaluation of the feature attributes. By comparing Figure 4.B.1(e), with the plot for the average query from Figure 4.2(a) we see that RankingSHAP is the only approach assigning x_{uni} a higher value than for the average query.

4.B.2 Unbiased model explanations

The bar chart in Figure 4.B.2 shows the feature attribution values from the three considered approaches from Section 4.5.2 for the same query scenarios as defined in Section 4.5.3. Comparing the attribution values of different models for different query scenarios like in Figures 4.B.1 and 4.B.2 can help us with selecting the least biased model when we have a choice of models of similar performance.

4.B.3 Additional per candidate analysis

Here we provide additional results of the per candidate analysis from Section 4.5.4, which can be found in Figures 4.B.3 and 4.B.4.

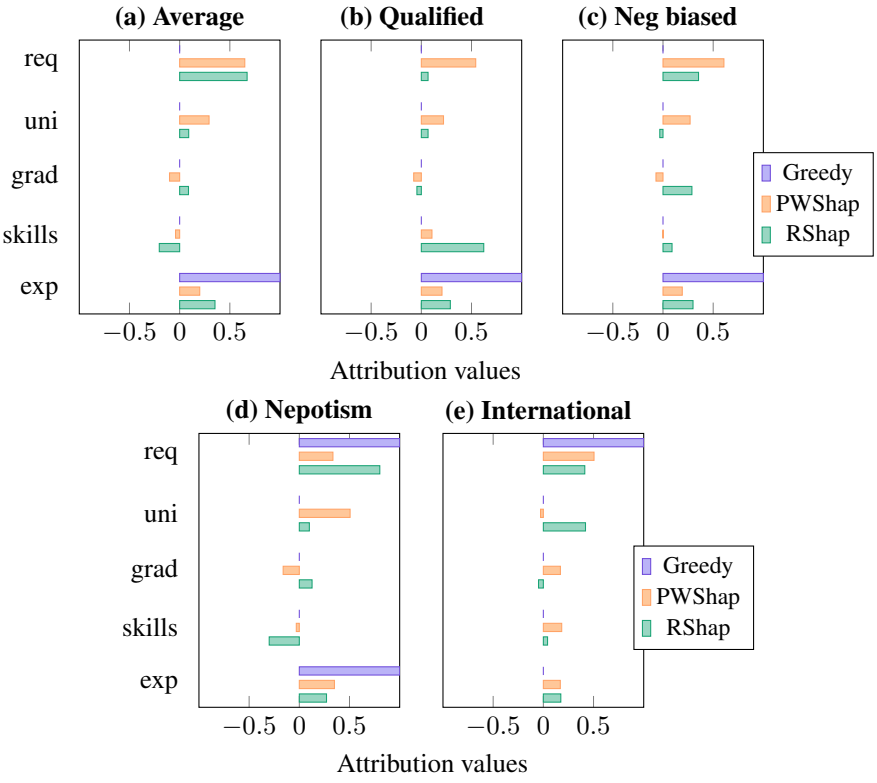


Figure 4.B.2: Feature attribution values of the unbiased model for different query scenarios

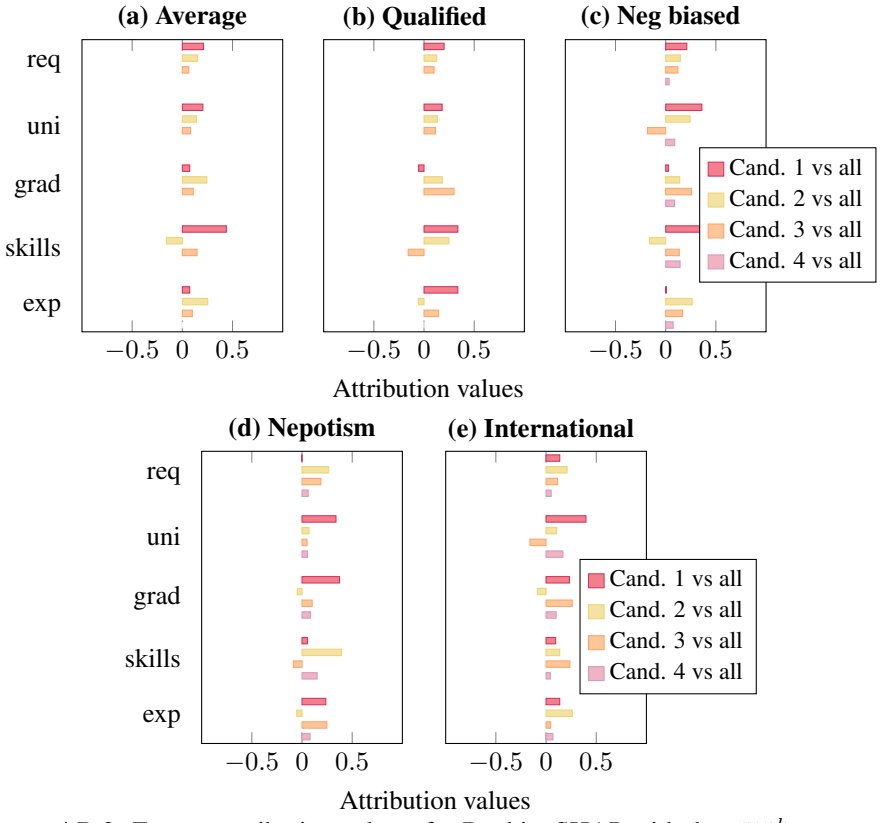


Figure 4.B.3: Feature attribution values, for RankingSHAP with the g_q^{rank} exposure objective defined in Section 4.4.3

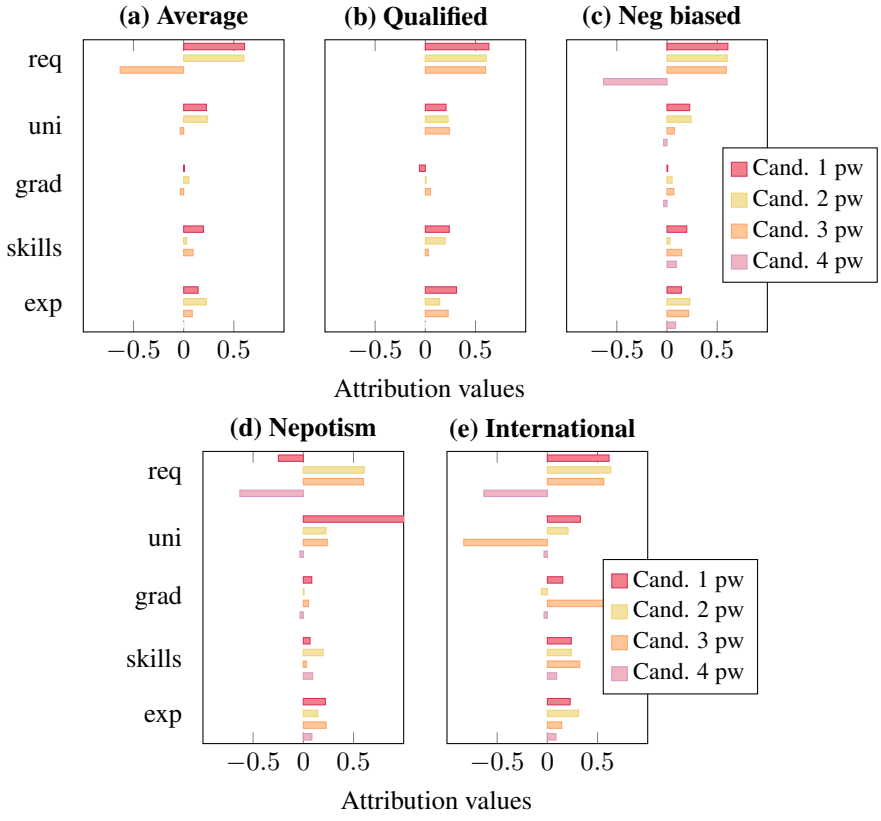


Figure 4.B.4: Feature attribution values, Pointwise SHAP for each individual candidate in the ranked list.

5

Correctness is not Faithfulness in Retrieval Augmented Generation Attributions

In the second chapter of Part II on interpretability in advice-giving systems, and thereby the last technical chapter of this thesis, we move away from explaining ranking systems toward another component of responsible advice-giving. We examine retrieval augmented generation (RAG) systems that, in addition to retrieving and sorting documents with respect to relevance scores, interface with human users through a chat-based interface operated by a large language model.

Recent progress in the field of LLMs has made RAG systems more attractive for enabling easy and customized interaction with information. However, while the ever-growing complexity of LLMs enables them to perform impressive tasks, their decision-making processes are notoriously difficult to interpret.

Self-explanations such as chain-of-thought, in which the model generates a reasoning chain that explains the final answer, or, in the case of RAG, citations that explain the origin of certain pieces of information, might help shed light on parts of the answer generation process. However, such self-explanations should only be trusted when they are faithful to the answer generation process; in other words, when they accurately reflect the internal model processes that led to the generation of the answer.

This leads us to ask the following research question:

RQ D Do RAG citations faithfully reflect the source of the information used in the answer generation process?

Note: In this chapter we use A as notation for the set of candidate documents, as opposed to $\mathcal{D}$ from prior chapters.

This chapter was published as J. Wallat, M. Heuss, M. de Rijke, and A. Anand. Correctness is not faithfulness in RAG attributions. In *ICTIR 2025: The 15th International Conference on the Theory of Information Retrieval*. ACM, July 2025.

5.1 Introduction

Recent years have shown great improvements in LLMs and a steep increase in the adoption of chat systems for different tasks, such as information access. They can improve information accessibility through their interactive nature, the possibility to interact with information in a foreign language or the use of simple language. The adoption of these systems spans lots of different societal applications, ranging from healthcare [240] and legal systems [186] to education [108]. Trustworthiness of AI systems is key to their responsible deployment and usage in high-stakes scenarios, particularly in such high stakes domains [95, 101].

A critical challenge in these systems are hallucinations, where large language models (LLMs) generate plausible but incorrect or fabricated information, potentially undermining their reliability and disproportionately affecting vulnerable populations who may rely on these systems for critical information access [9].

One promising approach to address hallucinations is enabling text generation that is explicitly grounded in retrieved source documents and accompanied by citations [24, 166] which is often operationalized through **retrieval augmented generation (RAG)**. RAG employs a two-step process: first, retrieve relevant documents and then use them to generate answers. While citations cannot eliminate hallucinations, they enhance verifiability by explaining the origin of information [133]. This grounded text generation approach [74] has been successfully applied to various NLP tasks, including summarization and question answering. Recent implementations of RAG mechanisms [126] ensure that content remains coherent, contextually relevant, and anchored in verifiable sources [24].

In this chapter, we investigate the faithfulness of citations in RAG, examining whether cited documents genuinely contribute to the answer generation process or are merely superficially referenced. We conceptualize citations as a form of LLM (self-) explanation that should give insight into the source of generated information, analogous to how chain-of-thought explanations reveal a model's reasoning process. This analogy raises important concerns, as recent research [39] has demonstrated that even reasoning models, that should benefit from coherent chains of thought, frequently exhibit unfaithful behavior by omitting crucial information from their reasoning chains that was evidently used in generating answers.

Current evaluation practices for attributed text focus primarily on two aspects: the **correctness of the answer** and the **correctness of citations**, which is based on the agreement between attributed statements and the information found in referenced source documents. Citation correctness, sometimes called answer faithfulness [77], measures the extent to which cited documents support a generated statement.

We argue that ensuring mere correctness is insufficient for reliable information retrieval systems. This is particularly evident in domains such as legal information retrieval [142] and medical question answering [123], where documents are complex and responses are vulnerable to model biases. In these contexts, simple fact-checking or correctness evaluation may prove inadequate, requiring instead a nuanced understanding of document content. Both unwarranted trust and excessive skepticism toward model outputs can have significant consequences.

Moreover, research has shown that the presence of explanations can paradoxically

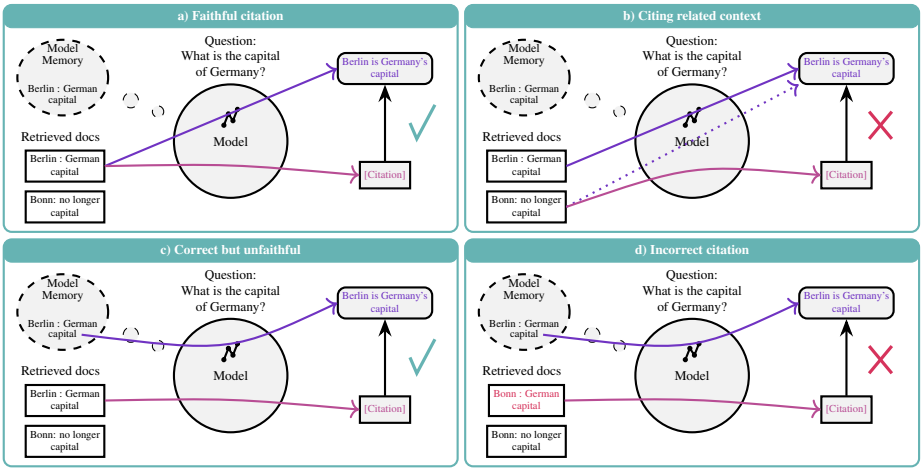


Figure 5.1: Different answer scenarios for the query "What is the capital of Germany?" (a) The ideal case, i.e., a correct citation that is faithful to the answer's generation process. (c) A correct but unfaithful citation, where the model post-rationalizes a citation to fit its prior. (b) A citation referring to the context that was used during the answer generation but does not contain the statement itself. (d) An incorrect citation.

increase user trust, even when these explanations are misleading [181], particularly in hard to assess tasks where output verification is challenging. To address this concern, we need to understand the model's reasoning process to verify that it correctly used the cited documents rather than answering from its parametric memory through **post-rationalization**, where models may cite sources to fit preconceived notions rather than genuine retrieval. We introduce the term **citation faithfulness** to describe whether the citation accurately reflects the model's reasoning process. Figure 5.1 illustrates the differences between faithful and unfaithful behavior as well as correct and incorrect citations.

When building trustworthy IR systems that offer self-explanations – in this case, citations – we should strive to convey the system's decisions accurately. Only if the produced citations are faithful to the underlying processes can we enable justified trust (as opposed to misplaced trust if faithfulness breaks down).

Our contributions are threefold: First, we offer coherent notions of attribution and citation in the context of grounded generation and introduce the concept of citation faithfulness. Second, we propose desiderata for citations that go beyond correctness and accuracy and are needed for trustworthy and usable systems. Third, we emphasize the need to evaluate the faithfulness of citations by studying post-rationalization. Our experiments reveal the existence of unfaithful behavior, with up to 57% of citations being post-rationalized.

Our work on disentangling citation correctness and faithfulness in grounded text generation using LLMs aims to create more reliable IR systems by ensuring accurate and contextually faithful citations. By focusing on post-rationalization, we enhance

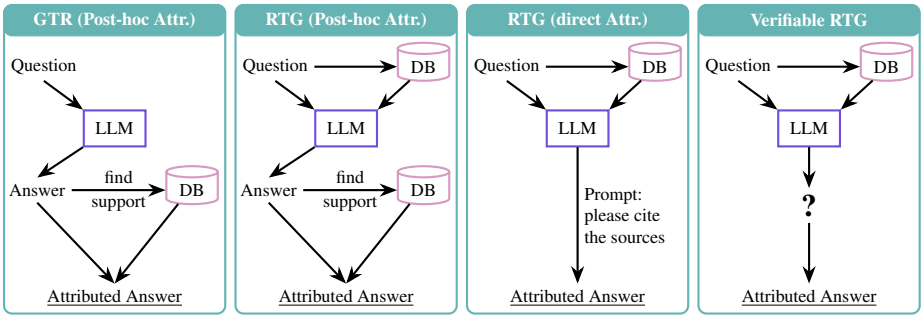


Figure 5.2: Different methods of attribution generation, using information from the database (DB) at different stages of the generation pipieline. The likelihood for unfaithful behavior and post-rationalization decreases from left to right.

accountability, helping IR systems avoid propagating biases or misinformation, thus promoting ethical standards in information dissemination and ensuring these systems effectively serve all users, regardless of their technical expertise or background.

5.2 Related Work

We summarize relevant background and position our work w.r.t. risks of LLMs, the evaluation of attributed generation, faithfulness in interpretability, and faithfulness of self-explanations. The area of knowledge conflicts [236] examines information flow and whether answers originate from parametric memory or the contextual [153, 227]. Its goal of understanding models is similar, but has a different focus (full answers vs. citations) and is therefore out of scope.

5.2.1 Risks of LLMs

Recent work in the field of responsible AI has identified numerous risks associated with the deployment of LLMs in real-world applications. These risks span multiple domains, from security vulnerabilities and susceptibility to adversarial attacks [233], environmental concerns [172], and challenges related to the trustworthiness of these systems and their alignment with social norms, values, and regulations [134]. This chapter focuses on the risk of unreliable or incorrect information being presented as authoritative and trustworthy. LLMs are known to produce hallucinated information that may be inconsistent with real-world facts or entirely unverifiable [129]. These fabricated "facts" can become sources of misinformation, since the presence or absence of citations can influence users' trust in the presented content [177]. The risk of misinformation becomes particularly concerning when considering demographic variations in susceptibility. Research has shown that certain population groups, including younger individuals, those with lower levels of education, and racial minorities, are especially vulnerable to health misinformation [151]. This vulnerability is particularly troubling given the increasing use of LLMs in high-stakes domains such as healthcare, where

misinformation has long been a significant concern among public health practitioners and researchers [41, 204]. The expansion of LLM applications into sensitive domains such as emotional support, financial advice, medical advice, and legal assistance [90] raises additional concerns. For instance, the use of LLMs for self-diagnosis purposes has been identified as a potential new vector for health misinformation [16]. These applications highlight the critical need for robust safeguards and regulatory frameworks. Current regulatory efforts, such as the EU AI Act, attempt to address these risks, though some argue that existing frameworks are inadequate for the challenges posed by generative language models [89]. Although legislative frameworks may need refinement, the documented risks associated with LLM-powered information systems underscore the technical community's responsibility to anticipate potential failures and develop responsible solutions. Our work in this chapter aims to contribute to that effort by examining post-rationalization and unfaithful citations in LLM-powered advice systems which might become sources of misinformation.

5.2.2 LLMs and attributions

Supplying LLM-generated answers with attributions aims to improve the quality of the generated answers [76], reduce hallucination [211], and improve users' trust [147] in the generated outputs. Methods for generating attributed answers range from prompting [76], adding post-hoc attributions [76, 210], and training paradigms [14, 34, 147, 210, 244] to generation-planning for more fine-grained citations [202]. Figure 5.2 provides an overview of common methods. The simplest method is generate-then-retrieve (GTR), a paradigm in which a model produces an answer (without attributions), and supporting evidence is added in a subsequent step [24, 76]. Retrieve-then-generate (RTG) operates similarly, but the model produces the (unattributed) answer after seeing both the question and the retrieved documents. As with GTR, RTG produces attributions in a second retrieval step, independent of the initially retrieved documents [244]. Thus, both GTR and RTG have post-hoc attributions, which are unfaithful to the model by design, i.e., the citation does not reflect the model's decision-making during the answer generation process. It is, however, possible to directly generate attributed answers by prompting the RTG model to do so [24, 76]. The resulting attributed answer *may* be faithful to the model's decision process, but we lack guarantees. As we show below, there is a significant chance of unfaithful behavior. The ultimate goal of attributed answer methodologies is to verify that certain information in the answer *originates* from the source document.

5.2.3 Evaluation of attributed generation

Attributed generation is a complex process that requires evaluation across multiple dimensions. One dimension is the *usability* of the generated response, which includes factors like fluency and perceived utility [133]. Traditionally, these factors have been assessed through user studies and automatic evaluation methods [76]. Other important dimensions include *answer relevance*, which measures how well the response addresses the question, and context relevance, which looks at the compactness of the retrieved context [66]. Datasets like HAGRID [106] are useful for evaluation, with human

evaluations of the informativeness and attributability of the responses, which can be used to measure overlap with gold citations [57]. Weller et al. [229] use the QUIP-Score, a method based on n-gram comparisons, to measure grounding and quoting from model pre-training data.

Next to the generated answer, the citation to the referenced document needs to be evaluated, too. To this end, prior work often uses natural language inference (NLI) classifiers [24, 75]. These help evaluate citation precision, which measures the average correctness of citations, and comprehensiveness/citation recall, which quantifies the proportion of accurately cited statements in all statements [57, 127]. The correctness of citations is a major focus in prior work [4, 57, 77, 127, 143, 166, 178, 180, 256]. We differentiate between citation correctness and the related but distinct aspect of **citation faithfulness**. Citation faithfulness requires a causal relationship between the cited document and the generated statement, an area that has so far received little attention.

5.2.4 Faithfulness in interpretability

In retrieval-augmented generation (RAG) attributions, (citation) faithfulness has not been studied much. In contrast, the evaluation of faithfulness of explanations has been studied extensively. Here, faithfulness refers to how accurately an explanation reflects the model's decision-making process, clearly differentiating it from explanation plausibility [100]. It lacks a universally accepted formal definition and is often defined in an ad-hoc manner [139]. Faithfulness establishes a causal relationship. Various methods have been proposed for evaluating faithfulness: (i) axiomatic evaluation, (ii) predictive power evaluation, (iii) robustness evaluation, (iv) perturbation-based evaluation, (v) white-box evaluation, and (vi) human perception evaluation [139]. Twelve desirable properties of explanations have been identified by Nauta et al. [152], including correctness (of explanations), which is equated with faithfulness. Overall, the concepts of faithfulness and correctness appear entangled in the explainability literature. We take a step towards disentangling those two aspects for attributed text. Inspired by Lyu et al. [139], we consider the causal relationship between the attributed text and generated answer to be a fundamental condition of faithful attribution.

5.2.5 Faithfulness of LLM self-explanations

Self-explanations are explanations that an LLM is prompted to generate along with the answer to a posed question. Self-explanations have been divided into (i) chain-of-thought (CoT) reasoning, which involves generating a sequence of intermediate steps that lead to the response [228], (ii) token importance, which highlights tokens that significantly influence the response generation [128, 235], and (iii) counterfactual explanations, which provide insights into how different inputs might lead to a different response [7]. Faithfulness of self-explanations has recently received attention [7, 121, 138, 213**?**], with work on evaluating faithfulness [121, 213] and its importance in contrast to plausibility [7]. There is high variation in how much LLMs use CoT on different tasks, some relying upon it heavily, others merely generating it in a post-hoc manner [121]. Recent work on evaluating the faithfulness of reasoning models reveals that even models explicitly trained for reasoning tasks exhibit an astonishing level of unfaithful CoT

reasoning [39, 44]. We view attributed generation that generates citations along with the text, rather than post-hoc, as a special class of self-explanation. We use a similar evaluation strategy as was previously used for the evaluation of faithfulness for CoT explanations [213] to show that similar faithfulness concerns arise for attributed generation as for CoT reasoning. We identify the problem of post-rationalization, which is closely related to post-hoc reasoning [121].

5.3 Attributions

RAG systems provide a way of grounding LLM-generated answers in documents that are retrieved from a corpus. By ensuring high quality of information in the corpus, this can improve the quality of the generated answers. RAG operates in two stages, where the first stage retrieves documents that match the information need/query of the user, and the second stage uses the retrieved documents to generate an answer. In the context of attributed text generation in RAG, an answer may be accompanied by references to documents, emphasizing that certain information originates from the referenced document. Merriam-Webster defines the verb *to attribute* as explaining by indicating a *cause*, emphasizing the causal nature.[1]

5.3.1 Notation

Let $A = \{a_i\}_i$ be a set of retrieved documents and let s be a text snipped a factual statement that needs to be grounded in the retrieved documents A. A citation $cit : s \mapsto a_j \in A$, or simply (s, a_j), connects a statement to a document that supports the stated statement. We use the term *attribution* to refer to the referenced document a_j or the process of referencing source documents.

Example 1: Attributed Answer

Question: Whats the biggest penguin in the world?
Answer: The Emperor Penguin [0] is the tallest [0] or biggest penguin in the world.

In Example 1, "tallest" would be a factual statement s attributed to document $a = 0$ through the citation ("tallest", 0). We note that many attributed statements are underspecified. Therefore, we distinguish between the statement ("tallest") and the underlying *claim* ("Emperor penguin: tallest: in the world"). Ideally, a citation should map claims to documents, but it is currently operationalized as statement to document, which can cause problems of misalignment between those two.

When attribution generation is integrated with answer generation, citations can be considered a form of self-explanation, others being chain-of-though explanations [228], explain-then-predict and predict-then-explain frameworks [27], and counterfactuals [38].

[1] https://www.merriam-webster.com/dictionary/attribute#h2

5.3.2 Desiderata for good attributions

Here we define several dimensions that can make attribution good or bad (Overview in Table 5.1).

Table 5.1: Desiderata for good attributions.

Desideratum	**Description**
Correctness	The attribution accurately represents the content of the cited document.
Faithfulness	The attribution accurately represents how the model derived its answer.
Appropriateness	The attribution is relevant and meaningful, not noisy or irrelevant.
Comprehensiveness	The attributions cover all the key points in the answer.

Correctness. Most importantly, good citations should be correct, meaning that the cited documents should support the generated statement. Ensuring correctness in attribution is crucial for maintaining the integrity and reliability of the information being presented. However, there are several ways in which the outputs of an LLM can be right or wrong.

Wrong answers. A direct way in which an LLM-generated answer can be wrong is if the statement itself is wrong, not matching the ground truth answer. This is the property that is evaluated most frequently in the open-domain QA and attribution literature [e.g., 24, 57, 126]. Wrong answers can result from hallucinations or correct attributions from a document containing false information. Therefore, an answer can be wrong despite having proper citations.

Hallucinated attributions. Attributions that do not exist, i.e., when a model hallucinates a reference to a non-existing document, are relatively easy to spot. LLMs without a retrieval component, such as the early versions of ChatGPT, especially, commonly generate broken links or hallucinate titles and authors of the source document from which certain information should come.

Wrong citations. Attributions can be incorrect, for example, by misrepresenting the content of the attributed documents or by attributing claims from document a to document b. In these cases, the citation (s, b) is incorrect. Compared to answer correctness, less work focuses on the correctness of attributions. Attributions are usually evaluated by testing if the attributed document implies the statement. To do so, recent work employs NLI models [24, 57, 75].

Appropriateness & comprehensiveness – *What* do we cite? Besides unfaithful behavior and incorrect attributions, bad citations may (appear to) be inappropriate or non-comprehensive and, therefore, dilute our understanding or evaluation of the answer. Appropriateness of attributions means that the attribution should be relevant, understandable, and meaningful; comprehensiveness refers to covering all the key points in the answer. The question of how much we need to cite and whether attributions cover

the important claims are less prominent in current evaluations frameworks, but these aspects may heavily skew the results of other evaluation metrics like correctness.

Example 2: Inappropriate Citations

Question: how long was gabby in a coma in the choice
Answer: In the novel [0, 4] the choice [0, 3, 4], Gabby is in a coma for three months.

Inappropriate citations. In Example 2, neither citation offers much value given the question. Attributing the title "the choice" provided in the question to documents 0, 3, and 4 offers no additional insights. On the contrary, when evaluating the quality of the provided citations, common approaches average over all existing citations. A large number of such *low-value* citations, which re-state information from the question, may heavily skew the evaluation metrics.

Short statements – What is the actual claim? Capturing a comprehensive, standalone statement in an LLM-generated response that maintains its specificity even when separated from the rest of the text can be a complex task. The statement is often reduced to a single word or concept, subtly referring to other parts of the generated response. In our example, it remains ambiguous as to what the highlighted word "novel" pertains to (i.e., the actual claim). This lack of clarity makes interpreting and evaluating such references more challenging.

For which statements do we need a citation? An answer may contain several citations, but one may be missing for the factual answer to the question. In the above example, the focus of the question is the time that Gabby spent in a coma ("three months"). This is the most critical statement in the answer and should be attributed to a source document. The above answer is not comprehensive since a central requested fact is not attributed to any source.

Faithfulness – Right for the *wrong* reason? Can an attribution be correct and still be bad? Like model explanations, attributions can be right for the wrong reason. To judge whether an attribution is right for the wrong reason, it is key to understand the internal model processes and understand whether a document a was considered during answer generation. If a is cited for another reason, then the attribution is not faithful to the underlying model behavior. Importantly, unfaithful attributions might still be factually correct and, therefore, difficult to spot – yet unfaithful attributions foster misguided trust.

Post-rationalization. We hypothesize that post-rationalized attributions are a special case of unfaithful behavior. In this setting, an LLM's parametric memory produces an answer to the question, and the model looks for support in the documents in some shallow way (e.g., by token-matching). The resulting citation is not faithful since the attribution superficially maps to a document, while using the model's internal knowledge. Let us consider Example 3.

Example 3: Faithfulness, Post-rationalization, Correctness

Question: What is the capital of Germany?
Answer: The capital of Germany is Berlin [1, 2]

Document 1: The capital of Germany is Berlin [...]
Document 2: Berlin has the best night-life [...]

Faithful (right for the right reason): Citing document 1 because the LLM used document 1's information to generate the answer.
Post-rationalized but **correct** (right for the wrong reason): Citing document 1 because the model knows the answer and finds a document that agrees with its priors.
Post-rationalized and **wrong**: Citing document 2 because the model knows the answer, and the answer token is mentioned in document 2.

Since the outputs in the faithful and unfaithful cases are identical (citing document 1), unfaithful behavior is hard to identify. We propose that a comprehensive evaluation of faithfulness must consider both the *attributions* themselves and the *process* through which they are derived. Given that citation faithfulness and correctness have often been conflated in previous research, we provide a detailed discussion and definition of citation faithfulness in Section 5.4.

5.4 Citation Faithfulness

The Cambridge Dictionary defines *faithful* as "true or not changing any of the details, facts, style, etc. of the original."[2] In explainability literature, a "faithful explanation should accurately reflect the reasoning process behind the model's prediction" [100]. Lyu et al. [139] further clarify that faithfulness establishes causality, distinguishing between "what is known by the model" and "what is actually used in making predictions."

Prior work on attributed answer generation defines *answer faithfulness* as the extent to which the cited document supports the generated statement [256]. Answer faithfulness considers the answer itself rather than the citation. In the context of the citation, this property is often called the correctness of the citation. In this chapter, we define **citation faithfulness** and disentangle the concepts of *answer faithfulness/correctness* and *citation faithfulness*. Prior work on attributed answers often has defined faithfulness loosely, for example, as "whether the selected documents influence the LLM during the generation" [164]. We take inspiration from the rich literature on the faithfulness of explanations and define the faithfulness of citations through a casual dependency of the generated answer and referenced document.

[2] https://dictionary.cambridge.org/us/dictionary/english/faithful

Definition 1: Citation Faithfulness

Let s be a generated statement underlying claim c. Let $A = \{a_i\}_i$ be a set of documents that the model has retrieved as context. We call (s, a_j) a faithful citation if:

- $a_j \in A$,
- The underlying claim c is supported by a_j (correctness), and
- c is causally impacted by a_j.

The second condition, often referred to as the "correctness" of the citation, has been a focal point in previous studies evaluating RAG attribution. Correctness usually tests whether a statement or claim is supported by the attributed document (measured by NLI models). However, while correctness is a necessary condition for faithfulness, it is insufficient. For a citation to be deemed faithful, the model must also rely causally on the cited document to generate the answer in a way that the flow of information goes from the document to the generated claim. The evaluation of this causal dependence of the model output on the cited statement has been largely overlooked, which is why we advocate for increased attention to the topic in future research.

We recognize that our definition of faithfulness is somewhat abstract. As Lyu et al. [139] observe, formulating a concrete definition with a single, comprehensive test to evaluate explanation faithfulness remains an open challenge – one that extends beyond our field to explanation methods in general. Thus, a set of more tangible necessary conditions with corresponding tests should be established in practice. These can assist in approximating the level of faithfulness of specific explanations. Consider the following examples of more concrete necessary conditions for faithful attribution. For a citation (s, a) to be considered faithful, the following should hold:

(1) If the relevant information in the cited document a is altered, the model should either provide a different generated statement s or modify the decision-making process. This could involve using different evidence a' or the model's memory to generate the answer.

(2) Adding irrelevant documents to the context should not affect the attribution, provided that the answer remains unchanged.

In Section 5.5, we design and implement an experiment to test the second necessary condition, providing empirical evidence for post-rationalization. While the first condition might offer broader insights into model faithfulness, testing it directly would require a deeper understanding of the model's internal decision-making process. Current analytical techniques are insufficient for this level of investigation. Therefore, we leave this analysis to future work.

5.5 Post-Rationalization – A Study of Unfaithful Behavior

We study attributions of a prominent RAG model and produce evidence of unfaithful behavior. As Jacovi and Goldberg [100] argue, faithfulness, as opposed to plausibility, should not be measured through human evaluation. Therefore rather than doing a human study to evaluate the quality of the citation self-explanations, we deploy a test based on input-output relationships. We investigate a particular case of unfaithful behavior, post-rationalization, i.e., the process in which a model generates a prior answer from model memory without regard to the documents and then searches retrieved documents to find supporting evidence.

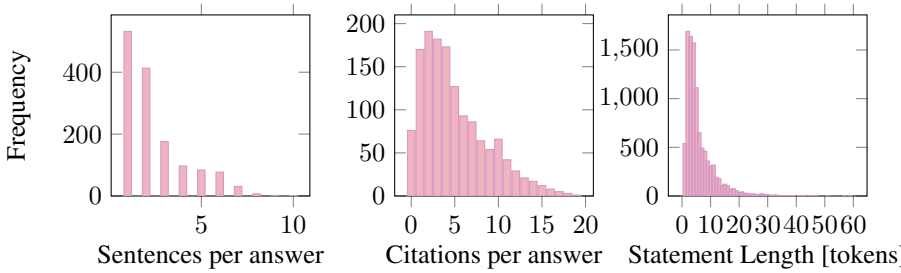


Figure 5.3: Statistical analysis of citations performed by Command-R+ on NaturalQuestions.

5.5.1 Setup

Cohere's COMMAND-R+ model is a "RAG-optimized" LLM specifically trained to produce grounded answers.[3] It has 104B parameters and a context length of 128k tokens, which we use in 4-bit quantization to run on a single NVIDIA A100 GPU. We evaluate COMMAND-R+'s attributions on the NaturalQuestions QA dataset, containing 1,444 real user questions answered by Wikipedia pages [117]. We use the temporally-aligned KILT [160] Wikipedia dump[4] as a retrieval base. Following [49], we split passages into chunks of 100 tokens and prepend the title of the page to the chunks. We index the resulting chunks and, for each query, retrieve the top 30 documents using BM25. We rerank the 30 retrieved documents using ColBERT v2 [183] and feed the top 5 documents together with the question into COMMAND-R+.

We use the grounded generation prompt template provided by Cohere.[5] The grounded generation pipeline with COMMAND-R+ follows four steps: (i) predict the relevance of the retrieved documents; (ii) predict which documents should be cited; (iii) produce an answer without citations, and (iv) one with citations. This setup makes COMMAND-R+ a retrieve-then-generate (RTG) model with direct attributions

[3] https://cohere.com/blog/command-r-plus-microsoft-azure
[4] Available here: https://huggingface.co/datasets/facebook/kilt_tasks.
[5] https://huggingface.co/CohereForAI/c4ai-command-r-plus

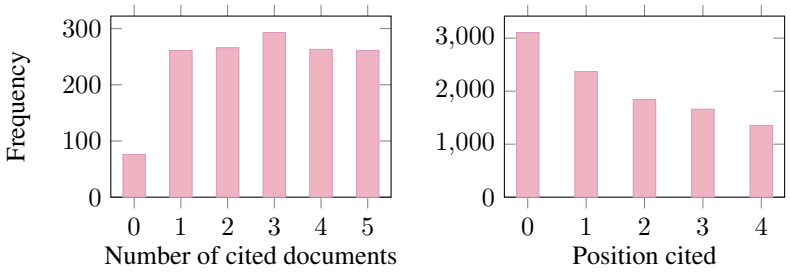


Figure 5.4: Number of cited documents and position of the cited documents in the input.

via prompting (see Figure 5.2). We selected an instance from this class of models since its chances of faithful behavior are higher than in the case of post-hoc attributions.

5.5.2 Citing behavior

As an initial step, we study the answers and attributions performed by COMMAND-R+. Figure 5.3 provides an overview. The model produces relatively short answers with on average 2.4 sentences and roughly five citations per answer. The cited spans (statements) have a median length 4 tokens and are, therefore, relatively short. Further looking at individual documents (Figure 5.4), we see that COMMAND-R+ cites on average 3 documents for a given statement, with almost equal frequency of 1–5 documents being cited. We also find 76 instances where the model did refrain from answering and, therefore, cited nothing. With regard to the position of the cited documents, we observe a tendency to citing the first documents. The first document is cited more than twice as much as the fifth document. However, since we order the input documents by reranking scores, it is to be expected that the earlier documents are more relevant.

To better understand COMMAND-R+'s grounded generation process, we also investigate whether the model cites the documents it predicted to be relevant and to be cited (step 1 and 2, c.f. Section 5.5.1). We present the results in Table 5.2. While it is expected that the model does not cite all documents it predicted to be relevant, it is somewhat surprising that it only cited 46% of the documents it predicted to be cited. In the remaining 54%, the model cited either nothing, fewer documents, or some documents it did not predict to be cited (1%). We hypothesize that the model was specifically trained to cite only the documents selected in the earlier processes. Furthermore, we did not find the model hallucinating attributions (e.g., citing document IDs other than the five retrieved documents). Nevertheless, the large mismatch between the documents predicted to be cited and the actual citations lets us question the faithfulness of the model's attribution behavior.

5.5.3 Unfaithful attributions

We devise the following experiments to better understand the extent to which COMMAND-R+ post-rationalizes citations. One possible way of post-rationalization could be finding documents to cite by token matching, so we (i) generate attributed answers for QA pairs, and (ii) select statements from these answers and append them to other documents.

Table 5.2: Investigation into the citing behavior of Command-R+. We explore whether all documents that the model predicted as relevant (step 1 in Command-R+'s grounded generation) and predicted to be cited (step 2) were cited in the grounded answer (step 4).

Split	Pred. Rel.	Pred. Cited
Cited all selected documents	636	820
Cited less than all selected documents	708	522
Cited not selected documents	8	12
Cited nothing	76	76

Since statements are, usually, around 4 tokens, they mostly contain short concepts such as "Emperor penguin" or "The Choice," which should not be cited when appearing without factual context. We append these adversarial statements into three kinds: random documents retrieved using BM25 for arbitrary questions, documents predicted to be relevant in step 1 of Section 5.5.1 but never cited, and documents cited for other statements in the attributed answer step 4 of Section 5.5.1. The created dataset with adversarial documents consists of 1,344 QA pairs (random), 702 (relevant but not cited), and 829 (cited for other reasons). In step (iii) we again generate attributed answers, but this time with our adversarial documents. In the case that the adversarial document was created from a random document, we append it to the list of documents in the context. If the original document was part of the context, we substitute it with the adversarial one. Lastly, (iv) we observe whether the model now cites our adversarial documents for the statements selected in step (ii). We operate under the assumption that citing documents that just randomly contain the statement ("Emperor penguin") indicates *post-rationalization*. This process is also depicted in Figure 5.5.

The results are presented in Figure 5.6. First and foremost, we note that recovering the old statement in the newly generated answer worked in 63–70% of the cases, while the generated answer changed at least to some extent in the remainder of the cases. Since a change in answer statement might reflect a change in the used attention mechanisms and makes it impossible to compare citations for previously generated statements with newly generated ones, we discard those cases. This is necessary to understand if the adversarial document has been cited for the same statement. By injecting the statement into random documents and passing them to the model, the model cited these documents in 12% (116/936) cases. Interestingly, the number of adversarial documents cited is much higher when forging relevant but uncited documents (57%) and documents cited for different reasons (55%). Based on our results, we conjecture post-rationalization to be a *common phenomenon*. We additionally present an examples of COMMAND-R+'s post-rationalization behavior, citing a random adversarial document below:

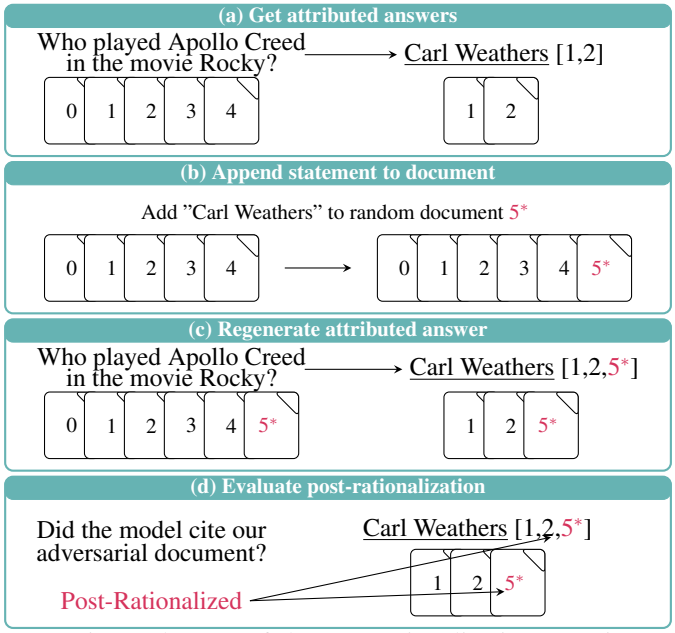


Figure 5.5: Experimental setup of the post-rationalization experiment. We inject attributed statements into random documents and regenerate the answer to see if the model cites unrelated documents when injected with statements.

Example 4: QA pairs with adversarial random documents added

Question: who played apollo creed in the movie rocky
Answer: Carl Weathers [1,2,5*] played Apollo Creed in the Rocky films.
Adversarial Document 5*: 1974 State of the Union Address
The 1974 State of the Union Address was given to the 93rd United States Congress, on Wednesday, January 30, 1974, by Richard Nixon, the 37th President of the United States. He said, "We meet here tonight at a time of great challenge and great opportunities for America. We meet at a time when we face great problems at home and abroad that will test the strength of our fiber as a nation. But we also meet at a time when that Carl Weathers

Are the adversarial documents actually adversarial? Our first experiment is based on the assumption that adversarial documents, generated by appending statements to unrelated documents, do not contain the actual claims. If this assumption does not hold, the model might be able to use the information within the document to generate the answer, hence the citation referencing this document might in fact be faithful. Therefore, to verify our estimation of post-rationalization, we examine whether adversarial documents alone might be sufficient to generate the investigated claims. Recall from Section 5.3.1 that we differentiate between the text snippet that the citation

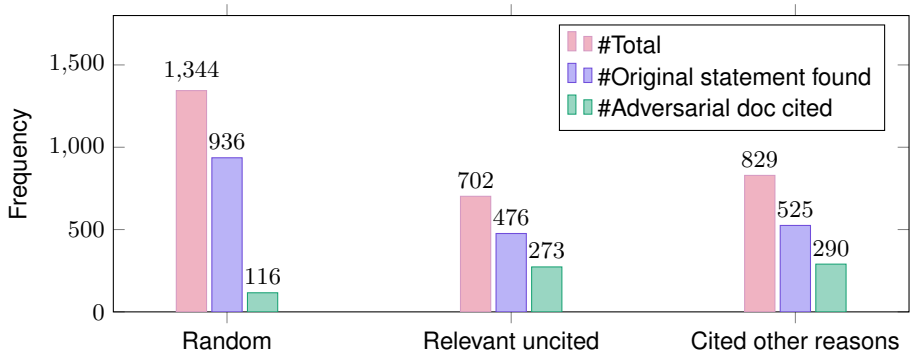


Figure 5.6: Results of the post-rationalization tests. We measure the cases in which the model cited our adversarial document (which had the previously cited statement appended). Since we also change the input, the model is not guaranteed to produce the same statements again. Therefore, we also include the number of cases where we could match the old statement.

is referencing called statement s ("Carl Weathers" in Figure 5.5) and the underlying claim c ("Carl Weathers played Apollo Creed in Rocky"). Since statements are typically quite short (median of 4 tokens, cf. Figure 5.4, right), we do not expect that adding statements alone provides sufficient context to generate an answer (or claim) using only information from the adversarial documents. To validate this assumption, we conducted an additional experiment focusing on instances where the model cites adversarial documents.

We ran inference using three different context configurations (instead of the full list of retrieved documents (e.g., [0,1,2,3,4]):

(a) The complete set of originally cited documents for the corresponding statement (e.g., [1,2] in Figure 5.5).

(b) One randomly sampled document that was originally cited for the corresponding statement (e.g., [2] in in Figure 5.5).

(c) The adversarial document alone (e.g., [5] in Figure 5.5).

We hypothesize that the model should recover original statements more frequently from previously cited documents than from adversarial documents, assuming that at least some of the original citations were faithful to their sources.

The results are presented in Figure 5.7. We found that adversarial documents alone are only sufficient to recover the statement in 14–21% of cases, varying by the document type used for adversarial generation. In contrast, contexts containing cited documents yielded much higher recovery rates of 30–43%. These findings were somewhat unexpected, as we anticipated higher recovery rates from originally cited documents and near-zero recovery from adversarial documents.

The relatively low recovery rates from cited documents might stem from the inherent instability of language generation, as can also be seen in Figure 5.6, where simply adding

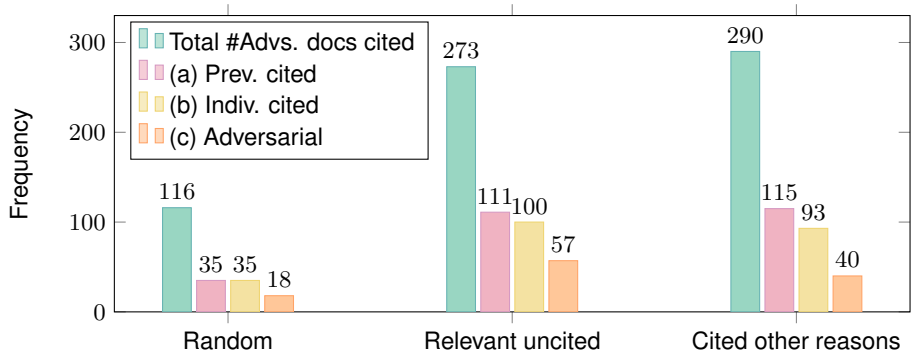


Figure 5.7: Frequency of statement recovered when providing different types of context documents to the model. The bars show the total number of adversarial instances investigated (green), and the number of recovered statements when all previously cited documents (pink), only a single cited document (yellow), or only the adversarial document (orange) is used as context.

irrelevant adversarial documents to the context reduces answer consistency, as well as the potential existence of unfaithful citations in the original answers.

On the other hand, several factors may explain the non-zero recovery from adversarial contexts: (1) the model might generate answers from its parametric memory and then match tokens to create citations, though this process is not directly observable; (2) as shown by Chen et al. [39], reasoning models change their answers based on subtle hints in prompts without faithfully reflecting these changes in their reasoning — similarly, the appended statements in our adversarial documents might serve as subtle hints pointing to plausible answers; and (3) a few adversarial documents may in fact contain the target claims, although preliminary qualitative analysis suggests otherwise.

Nevertheless, considering the big differences in recovery rate between the adversarial only setup (c) and the two baselines (a), (b) we conclude that at least most of the adversarial documents do **not** contain the claim that is necessary to generate the correct answer and can hence be considered adversarial.

5.6 Discussion

Citing parametric memory. Our results are the first step toward understanding unfaithful behavior in RAG systems due to post-rationalization. We focus on attributions, where a *faithful* attribution should signify the origin of the corresponding information. In contrast to past work, which values high citation recall, we argue that statements that were not generated from context but rather from model memory should not be accompanied by a citation. If the parametric model memory is used to generate an answer, a faithful model should either omit the citation or acknowledge their use of parametric memory rather than attempting to provide potentially misleading citations. This could for example be done by adding "model memory" as an explicit source, increasing transparency about the true origin of information.

The importance of faithfulness evaluation. Our work underscores the importance of establishing control settings that yield conclusive evidence regarding faithfulness in model-generated content. Several issues necessitate a principled approach to measuring faithfulness in future research. The challenges we encountered are reminiscent of those seen in explainability research within IR and other fields, where ensuring validity in attribution metrics remains difficult [17, 28, 138]. A lack of ground truth, as well as the inherently interpretative nature of attribution for RAG systems, presents a challenge for constructing evaluation criteria that can accurately identify unfaithful outputs. We suggest using evaluation strategies from explainability in IR, such as deliberate data contamination techniques [99, 198], model probing to gain first insights into specific model capabilities [72, 187, 220, 221], or reverse engineering parts of decision process [35]. However, validating LLM-based attributions introduces new challenges that call for the development of novel evaluation paradigms. We have proposed a preliminary test designed to assess faithfulness. This test, however, implicitly assumes that the model internals, or in other words, where the model looks and based on what it generates the answer, do not change through the insertion of additional irrelevant documents. To verify this assumption, an investigation of the model's internal states during answer generation would be necessary. Subsequent work could apply recent findings in understanding internal model processes [25, 64, 81] to the problem of faithful attributions.

5.7 Conclusion

In this chapter, we have demonstrated that citation faithfulness is a crucial yet often overlooked aspect of reliable information retrieval systems. We have defined desiderata of faithful attribution and defined and disentangled the notions of citation correctness and citation faithfulness. We provide empirical evidence of unfaithful citation behavior through post-rationalization in Command-R+, a state-of-the-art LLM trained for the RAG task, by measuring the impact of short text insertions into irrelevant documents on the generated citations. Our investigation reveals that up to 57% of such insertions result in post-rationalization, highlighting a significant gap between correctness measured through mere token matching and true faithfulness in citation behavior. This highlights the importance of evaluating faithfulness, along with correctness, especially in high-stakes decision-making and decision-support.

Our study has several limitations that warrant careful consideration. First, the relatively small scale of our empirical analysis may limit the generalizability of our findings across different contexts. Second, our research builds on the assumption that citations in AI-generated responses enhance user trust. While there is some initial evidence that misleading explanations can increase user trust [181], the impact of misleading citations on user trust still requires further empirical validation. Third, we acknowledge that a user study would be necessary to empirically measure the actual impact of misleading citations on information consumption and trust in reliable sources, particularly among vulnerable populations. Lastly, the conducted experiments raise several questions that cannot be addressed within the scope of this work, such as the instances where the adversarial document alone suffices to generate the model's answer, as observed in our experiments (Section 5.5).

These limitations point to several directions for future research. First, larger-scale studies are needed to validate our findings on post-rationalization and unfaithful attribution across a more diverse range of language models and datasets. Second, systematic human studies should investigate how different user groups, particularly vulnerable populations, interpret and interact with AI-generated citations. Third, researchers should develop robust evaluation frameworks for algorithmic accountability that specifically address attribution faithfulness in RAG systems. Finally, there is a need to explore alternative citation mechanisms that clearly distinguish between information drawn from model memory versus document-sourced statements.

Conclusion to Chapter 5

Returning to research question RQ D: "*Do RAG citations faithfully reflect the source of the information used in the answer generation process?*", in this chapter we investigated a state-of-the-art RAG model that was explicitly fine-tuned for the attributed generation task. We employed adversarial documents that had been specifically designed to trigger citations while not containing the corresponding supporting statements. This approach demonstrated that the model relies on superficial textual cues rather than genuine semantic grounding to generate citations, essentially post-rationalizing citations instead of anchoring them in the answer generation process.

Based on this counterexample using a state-of-the-art model, we must answer research question RQ D negatively: RAG citations do not always faithfully reflect the sources of information actually used during the answer generation process.

While research in this field is advancing rapidly with new LLMs appearing frequently, future investigations of citation faithfulness may require more sophisticated adversarial methods to demonstrate unfaithfulness or more comprehensive evaluation frameworks to build evidence supporting faithful citation behavior.

We consider this chapter a foundational step toward developing principled approaches for evaluating citation faithfulness in RAG systems – a critical requirement for creating trustworthy systems capable of supporting interactive and customizable automated advice-giving.

6 Conclusions

6.1 Summary of Findings

In this section we take a step back and discuss our main findings across the two parts of this thesis.

6.1.1 Fairness in ranking systems

The first part of this thesis focused on the fairness of IR systems. In Chapter 2, we took steps towards answering the first research question:

RQ A Can we define an exposure-fair ranking policy in situations where the expected exposure distribution is unknown for some rankings?

To answer this research question, in a context where the exposure distribution is unknown due to inter-document relationships, we formulated the task of fair ranking under incomplete exposure estimation. We generalized the convex optimization approach for fair ranking to be applicable to top-k rankings and provided an efficient algorithm that solves the convex optimization problem. Furthermore, we defined an approach that re-shuffles documents between ranked lists within the same ranking policy to provide a fairer ranking policy that mitigates rankings with unknown exposure distribution. Through careful experimentation on the example of outliers, which have been shown in past work to impact user exposure to documents, we demonstrated that our approach can substantially reduce the number of rankings with unknown exposure distribution that the policy produces. While this does not completely solve the problem, it represents a significant step toward addressing exposure-fairness in situations where the expected exposure distribution is unknown for some types of ranked lists due to inter-document relationships. This provided a positive but partial answer to the first research question: we can define an exposure-fair ranking policy when exposure distributions are unknown for some rankings, as FELIX successfully reduces the occurrence of such unknown rankings, yet it cannot completely eliminate them from the policy, leaving some room for improvement for future fair policies.

Then, we turned our attention to the task of bias mitigation for a language-based ranking model by answering the following research question:

RQ B Can we use the predictive uncertainty of the model prediction to improve ranking fairness?

We started our investigation by approximating the predictive model uncertainty about the order of documents in the ranked list with the help of Laplace approximation. We used these uncertainty estimates to define a method that swaps the ranking positions of documents with scores that have intersecting confidence intervals if it benefits the fairness or unbiasedness of the model. We showed empirically that with this intuitive approach, purely by exploiting the model's internal uncertainty about the ordering of documents, we could achieve a better utility-fairness trade-off than any of the baselines we test against, while remaining very computationally efficient. This lead us to answering the second research question positively: We can effectively use the predictive uncertainty of the model prediction to improve ranking fairness.

6.1.2 Explaining advice-giving processes

In the second part of this thesis, we turned to another aspect of responsible advice-giving: the ability to explain model predictions, thereby enabling monitoring of internal model processes (including biases), providing a tool for debugging, and giving users more information to help them decide whether to trust the model output.

Our first research question on this topic concerned explaining ranking models through feature attribution values:

RQ C How can we generate listwise ranking explanations for listwise ranking models?

Since listwise feature attribution explanations in past literature have not been properly defined, we started Chapter 4 by rigorously defining listwise feature attribution. For this, we defined a masking strategy that masks features within all documents simultaneously to determine the listwise importance of said feature, and the listwise explanation objective, which can be used to zoom in on certain properties of the ranked list, giving us a flexible tool for thoroughly investigating different aspects of the ranking decision. We defined a method that uses these two building blocks to extend SHAP, a Shapley-value based approach to explain pointwise (regression/classification) predictions, for the ranking use case. We introduced two evaluation paradigms to evaluate the explanations produced by our method against existing feature explanation approaches and show that our method performs competitively. Hence, we answered RQ C positively: We can generate listwise ranking explanations through the use of Shapley values.

To conclude this thesis, we investigated citation as a means of explaining the generated output of a language model that generates grounded answers in a RAG setup by asking the following research question:

RQ D Do RAG citations faithfully reflect the source of the information used in the answer generation process?

Since faithfulness and correctness of citations have previously been entangled concepts, we started Chapter 5 by defining desiderata of good and responsible attribution and highlighted the difference between citation faithfulness and citation correctness (also

called answer faithfulness). We introduced the phenomenon of post-rationalization of citations, where the model, rather than citing documents based on their use during answer generation, cites them in a post-hoc manner just for the sake of citing related information. We showed empirically that post-rationalization is a common phenomenon in state-of-the-art grounded generation, highlighting the need for further research in this field. We concluded by answering the research question negatively: RAG citations do not always faithfully reflect the source of information used in the answer generation process.

6.2 Impact of this Thesis

6.2.1 Questioning standard assumptions in fair information retrieval

Any kind of theoretical framework for fair advice-giving systems is built upon fundamental assumptions about task definitions, data availability, and measurable characteristics. For instance, group fair ranking approaches (as discussed in Chapter 2) typically assume complete knowledge of document group membership, merit/relevance scores, and exposure distributions across ranked results. However, these assumptions often fail to hold in practice, creating uncertainty about the validity and applicability of existing approaches.

This thesis addresses this challenge by examining two scenarios where standard assumptions break down. In Chapter 2, we investigated the case of incomplete knowledge about exposure distributions, demonstrating how fair ranking approaches must be adapted when this foundational assumption is violated. On the other hand, in Chapter 3, we explored uncertainty in merit assessment, showing how this uncertainty can actually be leveraged to design effective fairness mechanisms.

These investigations identify a key area for further research in fair ranking research: moving beyond the assumption that complete information is available or achievable. The approaches developed here offer both warnings about current methods' fragility and practical alternatives that maintain fairness when foundational assumptions fail. This shift in perspective is necessary for deploying fair information retrieval systems in practice, where incomplete information is the norm rather than the exception.

6.2.2 Explaining complex model outputs

The explainability literature has predominantly focused on classification and regression tasks [152], leaving advice-giving processes such as information access systems relatively underexplored. Explaining advice-giving processes presents distinct challenges due to the complex nature of their outputs. While classification models can be explained by identifying factors that increase class probabilities, and regression models through factors that influence prediction scores, advice-giving systems produce inherently more complex outputs that resist such straightforward analysis.

Consider ranking models, which generate ordered lists rather than single numerical values. Early efforts to explain these systems have largely adapted existing classification

and regression techniques, focusing on pointwise explanations that identify factors contributing to high ranking scores for individual documents. However, this approach fundamentally misses the contrastive nature of ranking decisions. A document's ranking score carries little meaning when considered in isolation. It gains meaning only when viewed relative to scores assigned to other documents in the ranking process.

In Chapter 4, we addressed this limitation by advocating for listwise explanations that directly capture how documents influence each other's positions in ranking decisions. Rather than explaining why individual documents receive high scores, listwise explanations reveal the factors responsible for the relative ordering of documents within the ranked list. Since in past literature a formal definition for explanations of ranking models is missing, we formally defined listwise explanations and showed how an explanation approach commonly used for other tasks, called SHAP, can be extended to explain specific aspects of a ranking decision.

This work represents an initial step toward principled explanations for information access/advice-giving systems with complex outputs, such as ranked lists. We anticipate that future research will build upon this foundation by establishing clear definitions and theoretical frameworks that enable systematic approaches to explaining the intricate outputs of those systems.

6.2.3 Highlighting challenges of self-explanations for advice-giving systems

While recent literature has devoted considerable attention to the trustworthiness and faithfulness of self-explanations such as chain-of-thought reasoning [7, 39, 44, 121, 138, 213], the trustworthiness of citations, that are frequently employed in RAG systems to explain the origin of information, has received comparatively little scrutiny. Although citation frameworks have traditionally focused on evaluating correctness, this metric alone fails to guarantee that citations accurately represent their information sources, as we argued in Chapter 5.

Our work in Chapter 5 provided a first investigation into citation faithfulness, moving beyond simple content matching to examine whether citations genuinely reflect the sources of information used during answer generation. We revealed that state-of-the-art models might fail to provide faithful citations, instead generating them post-hoc through content or token matching rather than accurately tracing the information flow from source documents or parametric memory during the generation process.

Our findings highlight critical gaps in current evaluation methodologies that fail to establish causal links between citations and generated answers, underscoring a fundamental challenge in developing trustworthy self-explanations. While our approach, similar to recent work by Chen et al. [39], examines how input modifications affect both outputs and explanations, such input-output relationship testing has inherent limitations, it can identify clearly unfaithful behavior but cannot guarantee that passing these tests means the model's actual decision process aligns with the provided explanation. More promising approaches may emerge from recent research examining the internal mechanisms models use during answer generation. So although our evaluation methodology successfully exposes flaws in current models, it may prove insufficient for detecting unfaithful citation behavior in more sophisticated future systems. We therefore view this

work as an important first, but definitely not final, step toward developing LLM-based advice-giving systems that faithfully provide information sources through citations rather than mere post-hoc justifications, setting the foundation for future research on trustworthy AI systems that users can rely on to verify the information they provide.

6.3 Limitations

6.3.1 Notions and definitions of fairness in information retrieval

Fairness is a multidimensional and complex topic that requires careful consideration of several key aspects [141]: the exact notion of fairness that should be applied, biases that might exist in the training data, technical biases that might be introduced through the model's functionality or presentation of results, and the definition of potentially disadvantaged groups.

A significant limitation of current fairness research in IR is its heavy reliance on theoretical assumptions about what constitutes fairness in real-world applications, without a general consensus on what notion of fairness is appropriate for what use case [149, 161].

Take, for instance, the concept of fairness of exposure, which assumes that a ranking policy is fair when documents or document groups receive adequate exposure. However, the definition of "adequate" varies considerably. Some approaches advocate for disparate treatment, where exposure should be proportional to the estimated utility [20], while other notions of fairness consider more impact oriented metrics, accounting for example for user interaction metrics like click-through rate [193].

The practical implications of these different approaches become clear when comparing two ranking policies: one that places only irrelevant documents from one group at the top while favoring relevant documents from another group, versus a policy that distributes similarly relevant documents from both groups evenly. While the first policy might satisfy certain representational fairness criteria such as disparate treatment, the second policy offers more meaningful opportunities for both groups to receive user attention.

Group fairness presents additional challenges, such as choosing appropriate aggregation functions to measure group utility and exposure. Consider a scenario where a single high-quality document from one group consistently ranks well and receives substantial exposure, while other relevant documents from the same group remain hidden from users. Especially in scenarios where exposure benefits have diminishing returns, as in hiring scenarios (where one candidate can only fill one job at a time), focusing visibility on one document while hiding other qualified group members fails to achieve meaningful group fairness.

While this thesis examined the implications of relaxing two common assumptions in fair ranking systems, that exposure distributions across ranked lists are known (Chapter 2) and that document merit can be accurately determined (Chapter 3), our work, like much of the existing literature, still relies on fundamental assumptions on the definition of fairness that should be applied, which are developed without substantial input from social science experts. To enhance the practical applicability of fair IR

research, closer collaboration with social scientists and stakeholders is essential to better understand what constitutes fair ranking in real-world contexts and how to implement fairness in advice-giving systems in practice (see, for example, the work by Green [88] on uniting social and technical aspects of fairness).

Moving forward, user studies examining both the behavioral impact of fairness interventions and users' perceptions of bias in ranked lists could provide valuable insights into the effectiveness of these approaches. Such research would help bridge the gap between theoretical fairness measures and their practical implications.

6.3.2 Explanation evaluation in information retrieval

The evaluation of explanations remains a fundamental challenge in the field. Unlike model performance evaluation, which relies on established benchmarks and metrics, creating ground truth datasets for "correct explanations" of model processes is particularly challenging. The complexity is further increased by the multiple aspects of explanations that require evaluation. While Nauta et al. [152] have identified 12 distinct categories of evaluation approaches across various machine learning tasks, the IR community still lacks consensus on which aspects are most crucial and how explanations should be evaluated.

In this thesis, we primarily focused on evaluating the faithfulness of generated explanations. In Chapter 4, we used a white box ranker to evaluate explanations by comparing them against our interpretation of the model, taking advantage of our understanding of its straightforward decision process. We also adapted two widely-used tests from other domains, the deletion and preservation checks [152], for ranking tasks. In Chapter 5, we developed an evaluation approach using adversarial documents to detect unfaithful behavior in model citations.

However, our evaluation approaches only address a fraction of the relevant aspects, even within the scope of faithfulness itself. Lyu et al. [139] discuss various concrete aspects of explanation faithfulness and their corresponding evaluation approaches. To develop a more comprehensive evaluation framework, future research should explore how existing evaluation methods from related fields, such as NLP and the broader ML literature, can be adapted for IR and advice-giving tasks. Additionally, researchers must identify which explanation aspects are particularly important in this context and may require developing novel evaluation methods.

As models continue to grow in size and complexity, and new explanation techniques emerge (such as mechanistic interpretability [18]), we will need to regularly update and enhance our evaluation frameworks. This ongoing refinement is crucial to ensuring that explanations serve their intended purpose of supporting decision-making, rather than generating misleading, unreliable, or unnecessary information.

6.3.3 The impact of explanations on user trust

Another common assumption in explainable AI research is that explanations directly influence user trust. However, this relationship is complex: a plausible but unfaithful explanation might increase trust inappropriately [181], while a faithful but counterintuitive explanation could decrease trust even when the model's prediction is reliable [7].

In Chapter 5, we assumed a similar dynamic with citations, where unfaithful citations may boost user trust in generated answers – a potentially harmful outcome when the cited information is hallucinated. To better support users in evaluating system-generated advice, we need a deeper understanding of the relationship between explanations and trust in advice-giving systems. This includes investigating how to optimally assist users through explanations in their information-seeking and decision-making process. Future research should examine this relationship through well-designed user studies that assess not only how misleading citations immediately affect user trust, but also their lasting implications over time.

6.4 Vision and Future Directions

6.4.1 Gaining insight into advice-giving systems through modern interpretability tools

The landscape of advice-giving systems has undergone significant expansion over the past decades. While traditional search-based tools continue to serve important roles, the field has broadened to include systems that leverage large language models through RAG architectures. As we look toward the future, emerging paradigms such as generative information retrieval [131] may represent yet another addition to this diverse ecosystem. Each approach within this expanding landscape brings with it distinct requirements for explanation and interpretation.

Modern advice-giving systems draw upon methodologies from multiple disciplines such as natural language processing, information retrieval and question answering. This interdisciplinary nature presents both challenges and opportunities for the field of explainable systems. The complexity introduced by combining different model components necessitates explanations that can address the unique characteristics of each component while providing coherent insights into the system as a whole. Depending on how these systems continue to develop, we must be prepared to adapt our explanation frameworks to match the specific approaches and architectures employed.

The introduction of LLM-based advice-giving systems has brought particular challenges that demand novel explanation methodologies. While traditional approaches in explainable information retrieval have largely focused on feature importance and input-output relationships, the increasing complexity of these newer models calls for more nuanced explanation techniques. The sensitivity of large language models to subtle variations in input, combined with the inherent instability in their generation processes, suggests that understanding internal model mechanisms may necessary to gain insight into the generation process. In this context, recent advances in mechanistic interpretability offer promising avenues for developing deeper insights into model behavior and decision-making processes.

Looking ahead, the emergence of generative information retrieval [131] as a competitive paradigm could add yet another dimension to our current ecosystem of advice-giving systems. If such systems prove capable of leveraging parametric memory more directly, potentially bypassing the explicit retrieval steps that characterize current RAG approaches, then established frameworks for citation and source attribution, such as

those discussed in Chapter 5, may require substantial re-conceptualization. This would necessitate developing entirely new explanation paradigms focused on assessing the trustworthiness and provenance of information drawn from parametric knowledge. As these technological advances continue to reshape the landscape of automated advice-giving, our approaches to making these systems interpretable and trustworthy must evolve in tandem, ensuring we maintain the ability to understand and validate their outputs regardless of the underlying approach.

6.4.2 Toward trustworthy model self-explanations

Self-explanations represent a promising avenue for enhancing the transparency and trustworthiness of model outputs. Different types of self-explanations provide distinct windows into how models operate: chain-of-thought explanations reveal the step-by-step reasoning processes, while citations expose the underlying sources that inform the generated content. However, both recent literature on chain-of-thought explanations and our investigation of citations in Chapter 5 demonstrate that these explanations are not consistently faithful, raising fundamental questions about their reliability.

Ensuring faithfulness in self-explanations presents a significant challenge that calls for research in two key areas: we need better ways to evaluate how faithful explanations actually are, and we must explore training approaches that encourage models to generate more reliable self-explanations. Recent work on reasoning models offers promising insights. When researchers explicitly train models to develop reasoning capabilities, the resulting self-explanations tend to be more faithful to the underlying processes. Building on this finding, we hypothesize that training or fine-tuning models specifically on attributed generation tasks could lead to similar improvements in citation faithfulness. However, this remains a hypothesis that will require careful empirical testing to confirm.

While self-explanations and citations certainly have their current limitations, they also bring several notable benefits to the table. For one, they give us valuable insight into a model's information sources without requiring significant computational overhead. Furthermore, they might enhance model performance analogous to how reasoning capabilities improve responses to complex queries. Perhaps most importantly, self-explanations have a key advantage over traditional post-hoc explanation methods. Post hoc approaches must infer model processes from outside observations. In contrast, self-explanations are generated using the same model and computational processes as the original outputs, theoretically enabling more authentic representations of the model's information processing.

Moving forward, research should prioritize the development of training paradigms that ensure faithful model self-explanations. Importantly, training should not target human interpretations of faithfulness directly, as this approach risks teaching models to simulate rather than genuinely exhibit faithful behavior. Instead, we advocate for using rigorous evaluation frameworks to identify training methodologies that naturally promote faithful self-explanations, similar to how reasoning-focused training simultaneously improves both model capabilities and the faithfulness of reasoning chains.

6.4.3 Advancing fair information retrieval through model explainability

A frequently cited motivation for research on explainable or interpretable models is their potential to help investigate model biases and fairness issues [15, 53]. In Chapter 4, we demonstrated this concept using a deliberately biased white box model to showcase how listwise feature attribution can help examine such biases. While some research exists on using explainability methods to identify dataset and model biases in various domains [11, 146, 156], there remains a notable gap in research specifically addressing real-world applications of IR and recommendation systems. On the other hand, current fairness research in IR frequently assumes prior knowledge of model biases, including awareness of marginalized groups and group membership characteristics, which in the real world cannot necessarily be guaranteed, or might even be impossible due to legislation forbidding the collection of those characteristics [67]. Moving forward, research should bridge the gap between interpretability and fairness studies, exploring how explainability techniques can effectively uncover biased or unfair model behavior and help ensuring fairness in cases where the group membership is unknown.

6.4.4 Responsible advice-giving as a whole

While existing research, including this thesis, has made significant progress in understanding individual aspects of responsible advice-giving systems, less attention has been paid to how these components work together to create a truly responsible system. This thesis examined specific elements, such as fair ranking mechanisms and explainable LLM interfaces for information presentation. However, an advice-giving pipeline consists of numerous other crucial components: from data curation and indexing to document retrieval, post-processing steps, and various design choices like document selection thresholds and interface design. Although developing responsible and interpretable individual components is valuable, we still lack a comprehensive understanding of how these elements interact and what risks might emerge when we fail to consider the system holistically. Moving forward, researchers should adopt a more integrated approach, examining responsible advice-giving systems as complete entities and studying their overall impact on end users.

Bibliography

[1] K. Aas, M. Jullum, and A. Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298:103502, 2021. (Cited on page 75.)

[2] T. Abdullahi, L. Mercurio, R. Singh, and C. Eickhoff. Retrieval-based diagnostic decision support: mixed methods study. *JMIR Medical Informatics*, 12:e50209, 2024. (Cited on page 1.)

[3] T. Abdullahi, R. Singh, C. Eickhoff, et al. Learning to make rare and complex diagnoses with generative AI assistance: qualitative study of popular large language models. *JMIR Medical Education*, 10(1):e51391, 2024. (Cited on page 1.)

[4] V. Adlakha, P. BehnamGhader, X. H. Lu, N. Meade, and S. Reddy. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:775–793, 2024. (Cited on page 88.)

[5] D. Afchar, V. Guigue, and R. Hennequin. Towards rigorous interpretations: A formalisation of feature attribution. In *International Conference on Machine Learning*, pages 76–86. PMLR, 2021. (Cited on page 62.)

[6] A. Agarwal, I. Zaitsev, X. Wang, C. Li, M. Najork, and T. Joachims. Estimating position bias without intrusive interventions. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 474–482, 2019. (Cited on page 16.)

[7] C. Agarwal, S. H. Tanneru, and H. Lakkaraju. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*, 2024. (Cited on pages 88, 106, and 108.)

[8] S. Agarwal and S. Mishra. *Responsible AI*. Springer, 2021. (Cited on page 1.)

[9] M. A. Ahmad, I. Yaramis, and T. D. Roy. Creating trustworthy LLMs: Dealing with hallucinations in healthcare AI. *arXiv preprint arXiv:2311.01463*, 2023. (Cited on page 84.)

[10] Q. Ai, K. Bi, C. Luo, J. Guo, and W. B. Croft. Unbiased learning to rank with unbiased propensity estimation. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 385–394, 2018. (Cited on page 16.)

[11] K. Alikhademi, B. Richardson, E. Drobina, and J. E. Gilbert. Can explainable AI explain unfairness? A framework for evaluating explainable AI. *arXiv preprint arXiv:2106.07483*, 2021. (Cited on page 111.)

[12] A. Anand, P. Sen, S. Saha, M. Verma, and M. Mitra. Explainable information retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3448–3451, 2023. (Cited on pages 58, 60, and 62.)

[13] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications, 2016. (Cited on page 36.)

[14] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024. (Cited on page 87.)

[15] E. Balkir, S. Kiritchenko, I. Nejadgholi, and K. C. Fraser. Challenges in applying explainability methods to improve the fairness of nlp models. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 80–92, 2022. (Cited on page 111.)

[16] F. Barnard, M. Van Sittert, and S. Rambhatla. Self-diagnosis and large language models: A new front for medical misinformation. *arXiv preprint arXiv:2307.04910*, 2023. (Cited on page 87.)

[17] J. Bastings, S. Ebert, P. Zablotskaia, A. Sandholm, and K. Filippova. "Will you find these shortcuts?" A protocol for evaluating the faithfulness of input salience methods for text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 976–991. Association for Computational Linguistics, 2022. (Cited on page 100.)

[18] L. F. Bereska. Mechanistic interpretability for AI safety—a review. In *Proceedings of The 1st Conference on Lifelong Learning Agents*, 2022. (Cited on pages 2 and 108.)

[19] A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi, and C. Goodrow. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 2212–2220. ACM, 2019. (Cited on page 39.)

[20] A. J. Biega, K. P. Gummadi, and G. Weikum. Equity of attention: Amortizing individual fairness in rankings. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 405–414, 2018. (Cited on pages 14, 16, 17, 36, 39, and 107.)

[21] G. Birkhoff. *Lattice Theory*. AMS, 1940. (Cited on page 22.)
[22] G. Birkhoff. Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucuman, Ser. A*, 5:147–154, 1946. (Cited on page 18.)
[23] J. S. Black and P. van Esch. AI-enabled recruiting: What is it and how should a manager use it? *Business horizons*, 63(2):215–226, 2020. (Cited on page 1.)
[24] B. Bohnet, V. Q. Tran, P. Verga, R. Aharoni, D. Andor, L. B. Soares, M. Ciaramita, J. Eisenstein, K. Ganchev, J. Herzig, et al. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*, 2022. (Cited on pages 84, 87, 88, and 90.)
[25] C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. (Cited on page 100.)
[26] A. Câmara and C. Hauff. Diagnosing BERT with retrieval heuristics. *Advances in Information Retrieval*, 12035:605, 2020. (Cited on page 61.)
[27] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom. e-SNLI: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018. (Cited on page 89.)
[28] O.-M. Camburu, E. Giunchiglia, J. Foerster, T. Lukasiewicz, and P. Blunsom. Can I trust the explainer? Verifying post-hoc explanatory methods. *arXiv preprint arXiv:1910.02065*, 2019. (Cited on page 100.)
[29] D. Campregher, Y. Chen, S. Hoffman, and M. Heuss. Tracing facts or just copies? A critical investigation of the competitions of mechanisms in large language models. *Transactions on Machine Learning Research*, 2025.
[30] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136, 2007. (Cited on page 26.)
[31] C. Castillo. Fairness and transparency in ranking. In *ACM SIGIR Forum*, pages 64–71, 2019. (Cited on pages 13 and 36.)
[32] L. E. Celis, D. Straszak, and N. K. Vishnoi. Ranking with fairness constraints. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*, page 28:1–28:15, 2018. (Cited on pages 16, 39, and 47.)
[33] L. E. Celis, A. Mehrotra, and N. K. Vishnoi. Interventions for ranking in the presence of implicit bias. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 369–380, 2020. (Cited on page 38.)
[34] A. Chen, P. Pasupat, S. Singh, H. Lee, and K. Guu. PURR: efficiently editing language model hallucinations by denoising language model corruptions. *arXiv preprint arXiv:2305.14908*, 2023. (Cited on page 87.)
[35] C. Chen, J. Merullo, and C. Eickhoff. Axiomatic causal interventions for reverse engineering relevance computation in neural retrieval models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 1401–1410. ACM, 2024. (Cited on page 100.)
[36] L. Chen, R. Ma, A. Hannák, and C. Wilson. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 chi conference on human factors in computing systems*, page 1–14. ACM, 2018. (Cited on page 38.)
[37] Y. Chen and C. Suh. Top-k ranking: An information-theoretic perspective. In *2015 IEEE Information Theory Workshop - Fall (ITW)*, pages 212–213, 2015. (Cited on page 15.)
[38] Y. Chen, R. Zhong, N. Ri, C. Zhao, H. He, J. Steinhardt, Z. Yu, and K. McKeown. Do models explain themselves? Counterfactual simulatability of natural language explanations. In *Forty-first International Conference on Machine Learning*, 2024. (Cited on page 89.)
[39] Y. Chen, J. Benton, A. Radhakrishnan, J. Uesato, C. Denison, J. Schulman, A. Somani, P. Hase, M. Wagner, F. Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025. (Cited on pages 84, 89, 99, and 106.)
[40] J. Choi, J. Choi, and W. Rhee. Interpreting neural ranking models using grad-CAM. *arXiv preprint arXiv:2005.05768*, 2020. (Cited on page 61.)
[41] W.-Y. S. Chou, A. Oh, and W. M. Klein. Addressing health-related misinformation on social media. *Jama*, 320(23):2417–2418, 2018. (Cited on page 87.)
[42] T. Chowdhury, R. Rahimi, and J. Allan. Rank-LIME: Local model-agnostic feature attribution for learning to rank. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 33–37, 2023. (Cited on pages 61, 67, and 72.)
[43] T. Chowdhury, Y. Zick, and J. Allan. RankSHAP: a gold standard feature attribution method for the

ranking task. *arXiv preprint arXiv:2405.01848*, 2024. (Cited on page 60.)

[44] J. Chua and O. Evans. Are DeepSeek R1 and other reasoning models more faithful? In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025. (Cited on pages 89 and 106.)

[45] A. Chuklin, I. Markov, and M. de Rijke. *Click Models for Web Search*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, August 2015. (Cited on page 16.)

[46] D. Cohen, B. Mitra, O. Lesota, N. Rekabsaz, and C. Eickhoff. Not all relevance scores are equal: Efficient uncertainty and calibration modeling for deep retrieval models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 654–664, 2021. (Cited on pages 36, 38, 42, 43, 45, 47, and 50.)

[47] D. Cohen, K. Du, B. Mitra, L. Mercurio, N. Rekabsaz, and C. Eickhoff. Inconsistent ranking assumptions in medical search and their downstream consequences. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2572–2577. ACM, 2022. ISBN 9781450387323. (Cited on pages 36 and 43.)

[48] N. C. Coombs, W. E. Meriwether, J. Caringi, and S. R. Newcomer. Barriers to healthcare access among US adults with mental health challenges: A population-based study. *SSM-population health*, 15:100847, 2021. (Cited on page 2.)

[49] F. Cuconasu, G. Trappolini, F. Siciliano, S. Filice, C. Campagnano, Y. Maarek, N. Tonellotto, and F. Silvestri. The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729, 2024. (Cited on page 94.)

[50] J. Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, 2022. (Cited on page 2.)

[51] P. R. Daugherty and H. J. Wilson. *Human+ Machine, Updated and Expanded: Reimagining Work in the Age of AI*. Harvard Business Press, 2024. (Cited on page 1.)

[52] M. de Rijke, B. van den Hurk, F. Salim, A. Al Khourdajie, N. Bai, R. Calzone, D. Curran, G. Demil, L. Frew, N. Gießing, M. K. Gupta, M. Heuss, S. Hobeichi, D. Huard, J. Kang, A. Lucic, T. Mallick, S. Nath, A. Okem, B. Pernici, T. Rajapakse, H. Saleem, H. Scells, N. Schneider, D. Spina, Y. Tian, E. Totin, A. Trotman, R. Valavandan, D. Workneh, and Y. Xie. Information retrieval for climate impact: Report on the MANILA24 workshop. *SIGIR Forum*, 59(2), June 2025.

[53] L. Deck, J. Schoeffer, M. De-Arteaga, and N. Kühl. A critical survey on fairness benefits of explainable AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1579–1595, 2024. (Cited on page 111.)

[54] Z. Deng, F. Zhou, and J. Zhu. Accelerated linearized Laplace approximation for Bayesian deep learning. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 2695–2708, 2022. (Cited on page 43.)

[55] P. M. Deshpande, D. P, and K. Kummamuru. Efficient online top-k retrieval with arbitrary similarity measures. In *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology*, pages 356–367, 2008. (Cited on page 15.)

[56] F. Diaz, B. Mitra, M. D. Ekstrand, A. J. Biega, and B. Carterette. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 275–284, 2020. (Cited on pages 14, 16, 17, 26, and 39.)

[57] H. Djeddal, P. Erbacher, R. Toukal, L. Soulier, K. Pinel-Sauvagnat, S. Katrenko, and L. Tamine. An evaluation framework for attributed information retrieval using large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5354–5359, 2024. (Cited on pages 88 and 90.)

[58] Q. Dong, Y. Liu, S. Cheng, S. Wang, Z. Cheng, S. Niu, and D. Yin. Incorporating explicit knowledge in pre-trained language models for passage re-ranking. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1490–1501. ACM, 2022. ISBN 9781450387323. (Cited on page 45.)

[59] A. Dotsinski, U. Thakur, M. Ivanov, M. H. Khan, and M. Heuss. On the generalizability of "Competition of mechanisms: Tracing how language models handle facts and counterfactuals". *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

[60] F. Dufossé, K. Kaya, I. Panagiotas, and B. Uçar. Further notes on Birkhoff–von Neumann decomposition of doubly stochastic matrices. *Linear Algebra and its Applications*, 554:68–78, 2018. (Cited on page 23.)

[61] W. H. Dutton. *Social transformation in an information society: Rethinking access to you and the world*, volume 13. Citeseer, 2004. (Cited on page 1.)

[62] M. D. Ekstrand, A. Das, R. Burke, and F. Diaz. Fairness and discrimination in information access systems. *arXiv preprint arXiv:2105.05779*, 2021. (Cited on pages 13 and 15.)
[63] M. D. Ekstrand, G. McDonald, A. Raj, and I. Johnson. Overview of the TREC 2021 fair ranking track. In *The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings*, 2022. (Cited on pages 13 and 36.)
[64] N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1 (1):12, 2021. (Cited on page 100.)
[65] A. Elliott. *The culture of AI: Everyday life and the digital revolution*. Routledge, 2019. (Cited on page 1.)
[66] S. Es, J. James, L. E. Anke, and S. Schockaert. 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*, pages 150–158, 2024. (Cited on page 87.)
[67] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. URL `https://data.europa.eu/eli/reg/2016/679/oj`. (Cited on page 111.)
[68] A. Fabris, A. Purpura, G. Silvello, and G. A. Susto. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Information Processing & Management*, 57:102377, 2020. (Cited on page 38.)
[69] Z. Fang, A. Agarwal, and T. Joachims. Intervention harvesting for context-dependent examination-bias estimation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 825–834, 2019. (Cited on page 16.)
[70] Z. T. Fernando, J. Singh, and A. Anand. A study on the interpretability of neural retrieval models using DeepSHAP. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 1005–1008, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6172-9. (Cited on page 61.)
[71] L. Floridi. *The fourth revolution: How the infosphere is reshaping human reality*. OUP Oxford, 2014. (Cited on page 1.)
[72] T. Formal, B. Piwowarski, and S. Clinchant. Match your words! A study of lexical matching in neural information retrieval. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II*, volume 13186 of *Lecture Notes in Computer Science*, pages 120–127. Springer, 2022. (Cited on page 100.)
[73] Y. Ganin and V. S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1180–1189, 2015. (Cited on page 46.)
[74] J. Gao, B. Peng, C. Li, J. Li, S. Shayandeh, L. Liden, and H.-Y. Shum. Robust conversational AI with grounded text generation. *arXiv preprint arXiv:2009.03457*, 2020. (Cited on page 84.)
[75] L. Gao, Z. Dai, P. Pasupat, A. Chen, A. T. Chaganty, Y. Fan, V. Y. Zhao, N. Lao, H. Lee, D.-C. Juan, et al. RARR: Researching and revising what language models say, using language models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023. (Cited on pages 88 and 90.)
[76] T. Gao, H. Yen, J. Yu, and D. Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, 2023. (Cited on page 87.)
[77] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023. (Cited on pages 84 and 88.)
[78] W. Garney, K. Wilson, K. V. Ajayi, S. Panjwani, S. M. Love, S. Flores, K. Garcia, and C. Esquivel. Social-ecological barriers to access to healthcare for adolescents: a scoping review. *International journal of environmental research and public health*, 18(8):4138, 2021. (Cited on page 2.)
[79] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56 (Suppl 1):1513–1589, 2023. (Cited on page 4.)
[80] X. Geng, T. Liu, T. Qin, and H. Li. Feature selection for ranking. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, editors, *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 407–414. ACM, 2007. (Cited on page 61.)
[81] M. Geva, R. Schuster, J. Berant, and O. Levy. Transformer feed-forward layers are key-value memories.

In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, 2021. (Cited on page 100.)

[82] S. C. Geyik, S. Ambler, and 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*, pages 2221–2231, 2019. (Cited on pages 14 and 38.)

[83] S. Ghalebikesabi, L. Ter-Minassian, K. DiazOrdaz, and C. C. Holmes. On locality of local explanation models. *Advances in neural information processing systems*, 34:18395–18407, 2021. (Cited on page 75.)

[84] A. Ghosh, R. Dutt, and C. Wilson. When fair ranking meets uncertain inference. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1033–1043, 2021. (Cited on page 39.)

[85] A. Gigli, C. Lucchese, F. M. Nardini, and R. Perego. Fast feature selection for learning to rank. In B. Carterette, H. Fang, M. Lalmas, and J. Nie, editors, *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval, ICTIR 2016, Newark, DE, USA, September 12- 6, 2016*, pages 167–170. ACM, 2016. (Cited on page 61.)

[86] T. Gillespie. The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society*, 167(2014):167, 2014. (Cited on page 1.)

[87] S. Gorantla, A. Deshpande, and A. Louis. On the problem of underranking in group-fair ranking. In *International Conference on Machine Learning*, pages 3777–3787. PMLR, 2021. (Cited on page 43.)

[88] B. Green. Escaping the impossibility of fairness: From formal to substantive algorithmic fairness. *Philosophy & Technology*, 35(4):90, 2022. (Cited on page 108.)

[89] P. Hacker, A. Engel, and M. Mauer. Regulating ChatGPT and other large generative AI models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1112–1123, 2023. (Cited on page 87.)

[90] C. E. Haupt and M. Marks. AI-generated medical advice—GPT and beyond. *Jama*, 329(16):1349–1350, 2023. (Cited on page 87.)

[91] M. Heuss, F. Sarvi, and M. de Rijke. Fairness of exposure in light of incomplete exposure estimation. In *SIGIR 2022: 45th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 759–769. ACM, July 2022. (Cited on page 39.)

[92] M. Heuss, D. Cohen, M. Mansoury, M. de Rijke, and C. Eickhoff. Predictive uncertainty-based bias mitigation in ranking. In *CIKM 2023: 32nd ACM International Conference on Information and Knowledge Management*, pages 762–772. ACM, October 2023.

[93] M. Heuss, C. Chen, A. Anand, C. Eickhoff, and S. Verberne. Workshop on explainability in information retrieval. In *SIGIR 2025: 48th international ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, July 2025.

[94] M. Heuss, M. de Rijke, and A. Anand. RankingSHAP–Listwise feature attribution explanations for ranking models. In *SIGIR 2025: 48th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 381–391. ACM, July 2025.

[95] High-Level Expert Group on AI. Ethics guidelines for trustworthy AI. Technical report, EU, 2019. `https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai`. (Cited on page 84.)

[96] S. Hofstätter, S.-C. Lin, J.-H. Yang, J. Lin, and A. Hanbury. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 113–122. ACM, 2021. ISBN 9781450380379. (Cited on pages 45 and 47.)

[97] S. Hofstätter, B. Mitra, H. Zamani, N. Craswell, and A. Hanbury. Intra-document cascading: Learning to select passages for neural document ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1349–1358, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. (Cited on page 61.)

[98] J. E. Hopcroft and R. M. Karp. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM Journal on Computing*, 2(4):225–231, 1973. (Cited on pages 22 and 34.)

[99] M. Idahl, L. Lyu, U. Gadiraju, and A. Anand. Towards benchmarking the utility of explanations for model debugging. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, pages 68–73, 2021. (Cited on page 100.)

[100] A. Jacovi and Y. Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, pages 4198–4205, 2020. (Cited on pages 61, 88, 92, and 94.)

[101] A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 624–635, New York, NY, USA, 2021. Association for Computing Machinery. (Cited on page 84.)

[102] N. Jethani, M. Sudarshan, I. C. Covert, S.-I. Lee, and R. Ranganath. FastSHAP: Real-time Shapley value estimation. In *International conference on learning representations*, 2021. (Cited on pages 63 and 75.)

[103] T. Joachims. Fairness and control of exposure in two-sided markets. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, 2021. (Cited on page 13.)

[104] T. Joachims, A. Swaminathan, and T. Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 781–789, 2017. (Cited on page 16.)

[105] D. M. Johnson, A. Dulmage, and N. Mendelsohn. On an algorithm of G. Birkhoff concerning doubly stochastic matrices. *Canadian Mathematical Bulletin*, 3(3):237–242, 1960. (Cited on pages 22 and 34.)

[106] E. Kamalloo, A. Jafari, X. Zhang, N. Thakur, and J. Lin. HAGRID: A human-LLM collaborative dataset for generative information-seeking with attribution. *arXiv preprint arXiv:2307.16883*, 2023. (Cited on page 87.)

[107] S. Kariyappa, L. Tsepenekas, F. Lécué, and D. Magazzeni. SHAP@k: Efficient and probably approximately correct (PAC) identification of top-k features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13068–13075, 2024. (Cited on page 75.)

[108] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023. (Cited on page 84.)

[109] M. Kay, C. Matuszek, and S. A. Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, pages 3819–3828, 2015. (Cited on page 2.)

[110] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, volume 30, 2017. (Cited on page 72.)

[111] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. (Cited on page 64.)

[112] J. Kleinberg and M. Raghavan. Selection problems in the presence of implicit bias. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, pages 33:1–33:17. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018. (Cited on page 38.)

[113] T. Kletti, J.-M. Renders, and P. Loiseau. Pareto-optimal fairness-utility amortizations in rankings with a DBN exposure model. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 748–758, 2022. (Cited on page 39.)

[114] S. Krishna, T. Han, A. Gu, S. Wu, S. Jabbari, and H. Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *Transactions on machine learning research (TMLR)*, 2024. (Cited on pages 59, 62, and 63.)

[115] A. Kristiadi, M. Hein, and P. Hennig. Being Bayesian, even just a bit, fixes overconfidence in ReLU networks. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5392–5402, 2020. (Cited on page 43.)

[116] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler. Problems with Shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pages 5491–5500. PMLR, 2020. (Cited on page 75.)

[117] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. (Cited on page 94.)

[118] Y. Kwon and J. Y. Zou. WeightedSHAP: Analyzing and improving Shapley based feature attributions. *Advances in Neural Information Processing Systems*, 35:34363–34376, 2022. (Cited on pages 58 and 75.)

[119] P. Lahoti, K. P. Gummadi, and G. Weikum. iFair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1334–1345. IEEE, 2019. (Cited on page 39.)

[120] P. Lahoti, K. P. Gummadi, and G. Weikum. Operationalizing individual fairness with pairwise fair

representations. *Proceedings of the VLDB Endowment*, pages 506–518, 2019. (Cited on page 39.)

[121] T. Lanham, A. Chen, A. Radhakrishnan, B. Steiner, C. Denison, D. Hernandez, D. Li, E. Durmus, E. Hubinger, J. Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. (Cited on pages 88, 89, and 106.)

[122] V. Laurim, S. Arpaci, B. Prommegger, and H. Krcmar. Computer, whom should i hire?–acceptance criteria for artificial intelligence in the recruitment process. 2021. (Cited on page 1.)

[123] M. Lee, J. Cimino, H. R. Zhu, C. Sable, V. Shanker, J. Ely, and H. Yu. Beyond information retrieval—medical question answering. In *AMIA annual symposium proceedings*, volume 2006, page 469. American Medical Informatics Association, 2006. (Cited on page 84.)

[124] T. Lei, R. Barzilay, and T. Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, 2016. (Cited on page 61.)

[125] J. Leonhardt, K. Rudra, and A. Anand. Learnt sparsity for effective and interpretable document ranking. *arXiv preprint arXiv:2106.12460*, 2021. (Cited on pages 60 and 61.)

[126] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020. (Cited on pages 5, 84, and 90.)

[127] D. Li, Z. Sun, X. Hu, Z. Liu, Z. Chen, B. Hu, A. Wu, and M. Zhang. A survey of large language models attribution. *arXiv preprint arXiv:2311.03731*, 2023. (Cited on page 88.)

[128] J. Li, X. Chen, E. Hovy, and D. Jurafsky. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691. Association for Computational Linguistics, 2016. (Cited on page 88.)

[129] J. Li, J. Chen, R. Ren, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, 2024. (Cited on page 86.)

[130] L. Li, T. Lassiter, J. Oh, and M. K. Lee. Algorithmic hiring in practice: Recruiter and HR professional's perspectives on AI use in hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 166–176, 2021. (Cited on page 1.)

[131] X. Li, J. Jin, Y. Zhou, Y. Zhang, P. Zhang, Y. Zhu, and Z. Dou. From matching to generation: A survey on generative information retrieval. *ACM Transactions on Information Systems*, 43(3):1–62, 2025. (Cited on page 109.)

[132] J. Lin, R. Nogueira, and A. Yates. Pretrained transformers for text ranking: BERT and beyond. *Synthesis Lectures on Human Language Technologies*, 14(4):1–325, 2021. (Cited on page 45.)

[133] N. F. Liu, T. Zhang, and P. Liang. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, 2023. (Cited on pages 84 and 87.)

[134] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. Guo, H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li. Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment. In *Socially Responsible Language Modelling Research*, 2023. (Cited on page 86.)

[135] A. Lucic, M. Srikumar, U. Bhatt, A. Xiang, A. Taly, Q. V. Liao, and M. de Rijke. A multistakeholder approach towards evaluating AI transparency mechanisms. In *ACM CHI Workshop on Operationalizing Human-Centered Perspectives in Explainable AI*. ACM, May 2021. (Cited on page 71.)

[136] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017. (Cited on pages 5, 59, 60, 61, 62, 63, and 73.)

[137] L. Lyu and A. Anand. Listwise explanations for ranking models using multiple explainers. In *European Conference on Information Retrieval*, pages 653–668. Springer Nature Switzerland Cham, 2023. (Cited on pages 58, 61, 62, and 64.)

[138] Q. Lyu, S. Havaldar, A. Stein, L. Zhang, D. Rao, E. Wong, M. Apidianaki, and C. Callison-Burch. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, 2023. (Cited on pages 88, 100, and 106.)

[139] Q. Lyu, M. Apidianaki, and C. Callison-Burch. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, pages 1–67, 2024. (Cited on pages 61, 62, 88, 92, 93, and 108.)

[140] D. J. C. Mackay. A practical Bayesian framework for backpropagation networks. *Neural Computation*,

4:448–472, 1992. (Cited on page 43.)
[141] K. Makhlouf, S. Zhioua, and C. Palamidessi. Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, 58(5):102642, 2021. (Cited on page 107.)
[142] K. T. Maxwell and B. Schafer. Concept and context in legal information retrieval. In *Legal Knowledge and Information Systems*, pages 63–72. IOS Press, 2008. (Cited on page 84.)
[143] J. Mayfield, E. Yang, D. J. Lawrie, S. MacAvaney, P. McNamee, D. W. Oard, L. Soldaini, I. Soboroff, O. Weller, E. S. Kayi, K. Sanders, M. Mason, and N. Hibbler. On the evaluation of machine-generated reports. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 1904–1915. ACM, 2024. (Cited on page 88.)
[144] A. Mehrotra and N. Vishnoi. Fair ranking with noisy protected attributes. In *Advances in Neural Information Processing Systems*, volume 35, pages 31711–31725, 2022. (Cited on pages 39 and 40.)
[145] R. Mehrotra, J. McInerney, H. Bouchard, M. Lalmas, and F. Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 2243–2251, 2018. (Cited on page 14.)
[146] G. I. Melsión, I. Torre, E. Vidal, and I. Leite. Using explainability to help children understandgender bias in AI. In *Proceedings of the 20th Annual ACM Interaction Design and Children Conference*, pages 87–99, 2021. (Cited on page 111.)
[147] J. Menick, M. Trebacz, V. Mikulik, J. Aslanides, F. Song, M. Chadwick, M. Glaese, S. Young, L. Campbell-Gillingham, G. Irving, et al. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022. (Cited on page 87.)
[148] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019. (Cited on page 75.)
[149] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application*, 8(1):141–163, 2021. (Cited on pages 3 and 107.)
[150] C. Molnar. *Interpreting Machine Learning Models with SHAP*. Independently published, 2023. (Cited on pages 58, 60, 62, 63, and 70.)
[151] X. Nan, Y. Wang, and K. Thier. Why do people believe health misinformation and who is at risk? A systematic review of individual differences in susceptibility to health misinformation. *Social Science & Medicine*, 314:115398, 2022. (Cited on page 86.)
[152] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, 55(13s):1–42, 2023. (Cited on pages 66, 71, 73, 88, 105, and 108.)
[153] E. Neeman, R. Aharoni, O. Honovich, L. Choshen, I. Szpektor, and O. Abend. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10056–10070. Association for Computational Linguistics, 2023. (Cited on page 86.)
[154] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches 2016*, 2016. (Cited on page 46.)
[155] R. F. Nogueira and K. Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019. (Cited on page 45.)
[156] I. Palatnik de Sousa, M. M. Vellasco, and E. Costa da Silva. Explainable artificial intelligence for bias detection in covid ct-scan classifiers. *Sensors*, 21(16):5657, 2021. (Cited on page 111.)
[157] G. K. Patro, L. Porcaro, L. Mitchell, Q. Zhang, M. Zehlike, and N. Garg. Fair ranking: A critical review, challenges, and future directions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, page 1929–1942. ACM, 2022. ISBN 9781450393522. (Cited on page 36.)
[158] G. Penha and C. Hauff. On the calibration and uncertainty of neural learning to rank models for conversational search. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 160–170. Association for Computational Linguistics, 2021. (Cited on pages 36, 38, and 42.)
[159] G. Penha, E. Krikon, and V. Murdock. Pairwise review-based explanations for voice product search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, pages 300–304,

2022. (Cited on pages 58 and 61.)

[160] F. Petroni, A. Piktus, A. Fan, P. S. H. Lewis, M. Yazdani, N. D. Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, V. Plachouras, T. Rocktäschel, and S. Riedel. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2523–2544. Association for Computational Linguistics, 2021. (Cited on page 94.)

[161] E. Pitoura, K. Stefanidis, and G. Koutrika. Fairness in rankings and recommendations: an overview. *The VLDB Journal*, pages 1–28, 2022. (Cited on page 107.)

[162] V. Pliatsika, J. Fonseca, T. Wang, and J. Stoyanovich. ShaRP: Explaining rankings with shapley values. *arXiv preprint arXiv:2401.16744*, 2024. (Cited on pages 60, 67, and 72.)

[163] A. Purpura, K. Buchner, G. Silvello, and G. A. Susto. Neural feature selection for learning to rank. In *European Conference on Information Retrieval*, pages 342–349. Springer, 2021. (Cited on page 61.)

[164] J. Qi, G. Sarti, R. Fernández, and A. Bisazza. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6037–6053, 2024. (Cited on page 92.)

[165] T. Qin and T.-Y. Liu. Introducing LETOR 4.0 datasets. *arXiv preprint arXiv:1306.2597*, 2013. (Cited on page 71.)

[166] H. Rashkin, V. Nikolaev, M. Lamm, L. Aroyo, M. Collins, D. Das, S. Petrov, G. S. Tomar, I. Turc, and D. Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023. (Cited on pages 84 and 88.)

[167] N. Rekabsaz and M. Schedl. Do neural ranking models intensify gender bias? In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2065–2068, 2020. (Cited on pages 38, 40, and 46.)

[168] N. Rekabsaz, S. Kopeinik, and M. Schedl. Societal biases in retrieved contents: Measurement framework and adversarial mitigation of bert rankers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 306–316, 2021. (Cited on pages 36, 38, 39, 40, 44, 45, 46, and 51.)

[169] D. Rennings, F. Moraes, and C. Hauff. An axiomatic approach to diagnosing neural IR models. In *European Conference on Information Retrieval*, pages 489–503. Springer, 2019. (Cited on page 61.)

[170] D. L. Rhode. *Access to justice*. Oxford University Press, 2004. (Cited on page 2.)

[171] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016. (Cited on pages 2, 5, 58, 61, and 73.)

[172] M. C. Rillig, M. Ågerstrand, M. Bi, K. A. Gould, and U. Sauerland. Risks and benefits of large language models for the environment. *Environmental Science & Technology*, 57(9):3464–3466, 2023. (Cited on page 86.)

[173] H. Ritter, A. Botev, and D. Barber. A scalable Laplace approximation for neural networks. In *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*. International Conference on Representation Learning, 2018. (Cited on page 43.)

[174] S. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, pages 333–389, 2009. (Cited on page 26.)

[175] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33:294–304, 1977. (Cited on page 35.)

[176] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *Proceedings of the 39th International Conference on Machine Learning*, pages 18770–18795, 2022. (Cited on page 71.)

[177] J. Rowley and F. Johnson. Understanding trust formation in digital information sources: The case of Wikipedia. *Journal of Information Science*, 39(4):494–508, 2013. (Cited on page 86.)

[178] S. Roychowdhury, S. Soman, H. Ranjani, N. Gunda, V. Chhabra, and S. K. Bala. Evaluation of RAG metrics for question answering in the telecom domain. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024. (Cited on page 88.)

[179] C. Rus, J. Kareem, C. Xu, Y. Liu, Z. Deng, and M. Heuss. AMS42 at the NTCIR-18 FairWeb-2 task. *Proceedings of NTCIR-18*, May 2025.

[180] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia. 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)*, pages 338–354, 2024. (Cited on page 88.)

[181] M. Sadeghi, D. Pöttgen, P. Ebel, and A. Vogelsang. Explaining the unexplainable: The impact of misleading explanations on trust in unreliable predictions for hardly assessable tasks. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pages 36–46, 2024. (Cited on pages 85, 100, and 108.)

[182] S. Saha, H. Agarwal, S. Mohanty, M. Mitra, and D. Majumdar. ir_explain: a Python library of explainable IR methods. *arXiv preprint arXiv:2404.18546*, 2024. (Cited on page 61.)

[183] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, and M. Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734. ACM, July 2022. (Cited on pages 45 and 94.)

[184] P. Sapiezynski, W. Zeng, R. E. Robertson, A. Mislove, and C. Wilson. Quantifying the impact of user attention on fair group representation in ranked lists. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 553–562, 2019. (Cited on page 16.)

[185] F. Sarvi, M. Heuss, M. Aliannejadi, S. Schelter, and M. de Rijke. Understanding and mitigating the effect of outliers in fair ranking. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 861–869, February 2022. (Cited on pages 14, 15, 16, 19, 20, 25, 26, 27, 28, 29, and 39.)

[186] T. Seabrooke, E. Schneiders, L. Dowthwaite, J. Krook, N. Leesakul, J. Clos, H. Maior, and J. Fischer. A survey of lay people's willingness to generate legal advice using large language models (LLMs). In *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*, pages 1–5, 2024. (Cited on page 84.)

[187] P. Sen, D. Ganguly, M. Verma, and G. J. F. Jones. The curious case of IR explainability: Explaining document scores within and across ranking models. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 2069–2072. ACM, 2020. (Cited on page 100.)

[188] C. Shah and E. M. Bender. Envisioning information access systems: What makes for good tools and a healthy web? *ACM Transactions on the Web*, 18(3):1–24, 2024. (Cited on page 1.)

[189] L. S. Shapley. A value for n-person games. In *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton, 1953. (Cited on pages 5 and 60.)

[190] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017. (Cited on page 61.)

[191] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Workshop Track Proceedings*, 2014. (Cited on page 61.)

[192] A. Singh and T. Joachims. Equality of opportunity in rankings. In *Workshop on Prioritizing Online Content (WPOC) at NIPS*, page 31, 2017. (Cited on page 16.)

[193] A. Singh and T. Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2219–2228, 2018. (Cited on pages 2, 14, 15, 16, 17, 18, 20, 23, 26, 32, 39, and 107.)

[194] A. Singh and T. Joachims. Policy learning for fairness in ranking. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 5426–5436, 2019. (Cited on pages 16 and 39.)

[195] A. Singh, D. Kempe, and T. Joachims. Fairness in ranking under uncertainty. In *Advances in Neural Information Processing Systems*, volume 34, pages 11896–11908. Curran Associates, Inc., 2021. (Cited on pages 16 and 39.)

[196] J. Singh and A. Anand. Posthoc interpretability of learning to rank models using secondary training data. *arXiv preprint arXiv:1806.11330*, 2018. (Cited on page 61.)

[197] J. Singh and A. Anand. EXS: Explainable search using local model agnostic interpretability. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pages 770–773, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-5940-5. (Cited on pages 60 and 61.)

[198] J. Singh and A. Anand. Model agnostic interpretability of rankers via intent modelling. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 618–628, 2020. (Cited on pages 58, 60, 61, 62, 64, and 100.)

[199] J. Singh, M. Khosla, W. Zhenye, and A. Anand. Extracting per query valid explanations for blackbox learning-to-rank models. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 203–210, 2021. (Cited on pages 58, 67, 71, 72, and 73.)

[200] J. Singh, Z. Wang, M. Khosla, and A. Anand. Valid explanations for learning to rank models. *International Conference on the Theory of Information Retrieval*, 2021. (Cited on pages 61, 62, and 64.)

[201] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020. (Cited on page 75.)

[202] A. Slobodkin, E. Hirsch, A. Cattan, T. Schuster, and I. Dagan. Attribute first, then generate: Locally-attributable grounded text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3344, 2024. (Cited on page 87.)

[203] L. Soldaini, A. Cohan, A. Yates, N. Goharian, and O. Frieder. Retrieving medical literature for clinical decision support. In *Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29-April 2, 2015. Proceedings 37*, pages 538–549. Springer, 2015. (Cited on page 1.)

[204] B. G. Southwell, J. Niederdeppe, J. N. Cappella, A. Gaysynsky, D. E. Kelley, A. Oh, E. B. Peterson, and W.-Y. S. Chou. Misinformation as a misunderstood challenge to public health. *American Journal of Preventive Medicine*, 57(2):282–285, 2019. (Cited on page 87.)

[205] J. Stoyanovich, K. Yang, and H. Jagadish. Online set selection with fairness and diversity constraints. In *21st International Conference on Extending Database Technology, EDBT 2018*, pages 241–252, 2018. (Cited on page 38.)

[206] E. Štrumbelj and I. Kononenko. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11:1–18, 2010. (Cited on pages 60 and 62.)

[207] E. Štrumbelj and I. Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41:647–665, 2014. (Cited on page 60.)

[208] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017. (Cited on page 61.)

[209] T. Tao, X. Wang, Q. Mei, and C. Zhai. Language model information retrieval with document expansion. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 407–414, 2006. (Cited on page 26.)

[210] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al. LaMDA: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022. (Cited on page 87.)

[211] S. T. I. Tonmoy, S. M. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024. (Cited on page 87.)

[212] I. Turc, M. Chang, K. Lee, and K. Toutanova. Well-read students learn better: The impact of student initialization on knowledge distillation. *arXiv preprint arXiv:1908.08962*, 2019. (Cited on page 45.)

[213] M. Turpin, J. Michael, E. Perez, and S. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36, 2024. (Cited on pages 88, 89, and 106.)

[214] A. Vardasbi, H. Oosterhuis, and M. de Rijke. When inverse propensity scoring does not work: Affine corrections for unbiased learning to rank. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1475–1484, October 2020. (Cited on page 16.)

[215] A. Vardasbi, G. Bénédict, S. Gupta, M. Heuss, P. Khandel, M. Li, and F. Sarvi. The University of Amsterdam at the TREC 2021 fair ranking track. In *TREC*, 2021.

[216] A. Vardasbi, M. de Rijke, and I. Markov. Mixture-based correction for position and trust bias in counterfactual learning to rank. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, page 1869–1878, November 2021. (Cited on page 16.)

[217] A. Vardasbi, F. Sarvi, and M. de Rijke. Probabilistic permutation graph search: Black-box optimization for fairness in ranking. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, July 2022. (Cited on page 16.)

[218] M. Verma and D. Ganguly. LIRME: Locally interpretable ranking model explanation. In *Proceedings of the 42nd International ACM SIGIR*, 2019. (Cited on pages 60 and 61.)

[219] M. Völske, A. Bondarenko, M. Fröbe, M. Hagen, B. Stein, J. Singh, and A. Anand. Towards axiomatic explanations for neural ranking models. *International Conference on the Theory of Information Retrieval*, 2021. (Cited on page 61.)

[220] J. Wallat, F. Beringer, A. Anand, and A. Anand. Probing BERT for ranking abilities. In *European Conference on Information Retrieval*, pages 255–273. Springer Nature Switzerland Cham, 2023. (Cited on pages 61 and 100.)

[221] J. Wallat, H. Hinrichs, and A. Anand. Causal probing for dual encoders. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2292–2303, 2024. (Cited on page 100.)

[222] J. Wallat, M. Heuss, M. de Rijke, and A. Anand. Correctness is not faithfulness in RAG attributions. In *ICTIR 2025: The 15th International Conference on the Theory of Information Retrieval*. ACM, July 2025.

[223] L. Wang and T. Joachims. User fairness, item fairness, and diversity for rankings in two-sided markets. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 23–41, 2021. (Cited on pages 14, 16, 17, 23, and 39.)

[224] W. Wang, F. Feng, X. He, H. Zhang, and T.-S. Chua. Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1288–1297, 2021. (Cited on page 16.)

[225] X. Wang, M. Bendersky, D. Metzler, and M. Najork. Learning to rank with selection bias in personal search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115–124, 2016. (Cited on page 16.)

[226] X. Wang, N. Golbandi, M. Bendersky, D. Metzler, and M. Najork. Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 610–618, 2018. (Cited on page 16.)

[227] Y. Wang, S. Feng, H. Wang, W. Shi, V. Balachandran, T. He, and Y. Tsvetkov. Resolving knowledge conflicts in large language models. In *First Conference on Language Modeling*, 2024. (Cited on page 86.)

[228] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing systems*, volume 35, pages 24824–24837, 2022. (Cited on pages 88 and 89.)

[229] O. Weller, M. Marone, N. Weir, D. Lawrie, D. Khashabi, and B. Van Durme. "According to...": Prompting language models improves quoting from pre-training data. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2288–2301, 2024. (Cited on page 88.)

[230] J. White. PubMed 2.0. *Medical reference services quarterly*, 39(4):382–387, 2020. (Cited on page 1.)

[231] T. Wiegman, L. Perotti, V. Pravdová, O. Brand, and M. Heuss. Reproducibility study of: "Competition of mechanisms: Tracing how language models handle facts and counterfactuals". *Transactions on Machine Learning Research*, 2025.

[232] S. Wiegreffe and Y. Pinter. Attention is not not explanation. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. (Cited on page 62.)

[233] F. Wu, N. Zhang, S. Jha, P. McDaniel, and C. Xiao. A new era in LLM security: Exploring security concerns in real-world LLM-based systems. *arXiv preprint arXiv:2402.18649*, 2024. (Cited on page 86.)

[234] X. Wu, J. Nian, Z. Tao, and Y. Fang. Evaluating social biases in LLM reasoning. *arXiv preprint arXiv:2502.15361*, 2025. (Cited on page 2.)

[235] Z. Wu, Y. Chen, B. Kao, and Q. Liu. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics., 2020. (Cited on page 88.)

[236] R. Xu, Z. Qi, Z. Guo, C. Wang, H. Wang, Y. Zhang, and W. Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, 2024. (Cited on page 86.)

[237] H. Yadav, Z. Du, and T. Joachims. Policy-gradient training of fair and unbiased ranking functions. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1044–1053, 2021. (Cited on pages 14, 16, and 17.)

[238] K. Yang and J. Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, 2017. (Cited on page 15.)

[239] K. Yang, V. Gkatzelis, and J. Stoyanovich. Balanced ranking with diversity constraints. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI*, 2019. (Cited on page 38.)

[240] R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting, and N. Liu. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263, 2023.

(Cited on pages 1 and 84.)

[241] T. Yang, C. Luo, H. Lu, P. Gupta, B. Yin, and Q. Ai. Can clicks be both labels and features? Unbiased behavior feature collection and uncertainty-aware learning to rank. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 6–17. ACM, 2022. (Cited on pages 38 and 42.)

[242] T. Yang, Z. Xu, Z. Wang, A. Tran, and Q. Ai. Marginal-certainty-aware fair ranking algorithm. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 24–32, 2023. (Cited on pages 36, 39, and 42.)

[243] X. Yang, H. Steck, Y. Guo, and Y. Liu. On top-k recommendation using social networks. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, pages 67–74, 2012. (Cited on page 15.)

[244] X. Ye, R. Sun, S. Ö. Arik, and T. Pfister. Effective large language model adaptation for improved grounding and citation generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6237–6251. Association for Computational Linguistics, 2024. (Cited on page 87.)

[245] P. Yu, R. Rahimi, and J. Allan. Towards explainable search results: A listwise explanation generator. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 669–680, 2022. (Cited on pages 58 and 61.)

[246] Y. Yue, R. Patel, and H. Roehrig. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1011–1018, 2010. (Cited on page 16.)

[247] M. Zehlike and C. Castillo. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference 2020*, pages 2849–2855, 2020. (Cited on pages 16 and 39.)

[248] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. FA*IR: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578, 2017. (Cited on pages 15, 16, 36, 39, and 47.)

[249] M. Zehlike, K. Yang, and J. Stoyanovich. Fairness in ranking: A survey. *arXiv preprint arXiv:2103.14000*, 2021. (Cited on pages 13, 15, and 17.)

[250] M. Zehlike, T. Sühr, R. Baeza-Yates, F. Bonchi, C. Castillo, and S. Hajian. Fair top-k ranking with multiple protected groups. *Information Processing & Management*, 59(1):102707, 2022. (Cited on pages 15, 16, and 39.)

[251] M. Zehlike, K. Yang, and J. Stoyanovich. Fairness in ranking, part I: Score-based ranking. *ACM Computing Surveys*, 55(6):1–36, 2022. (Cited on page 38.)

[252] M. Zehlike, K. Yang, and J. Stoyanovich. Fairness in ranking, part II: Learning-to-rank and recommender systems. *ACM Computing Surveys*, 55(6):1–41, 2022. (Cited on page 38.)

[253] G. Zerveas, N. Rekabsaz, D. Cohen, and C. Eickhoff. Mitigating bias in search results through contextual document reranking and neutrality regularization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2532–2538, 2022. (Cited on pages 38, 39, 46, and 51.)

[254] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 334–342, 2001. (Cited on page 26.)

[255] J. Zhang, Q. Sun, J. Liu, L. Xiong, J. Pei, and K. Ren. Efficient sampling approaches to Shapley value approximation. *Proceedings of the ACM on Management of Data*, 1(1):1–24, 2023. (Cited on page 63.)

[256] W. Zhang, M. Aliannejadi, Y. Yuan, J. Pei, J.-h. Huang, and E. Kanoulas. Towards fine-grained citation evaluation in generated text: A comparative analysis of faithfulness metrics. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 427–439, 2024. (Cited on pages 5, 88, and 92.)

[257] Z. Zhang, K. Rudra, and A. Anand. Explain and predict, and then predict again. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 418–426, 2021. (Cited on page 60.)

[258] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021. (Cited on page 58.)

[259] J. Zhu, J. Wang, M. Taylor, and I. J. Cox. Risk-aware information retrieval. In *Advances in Information Retrieval: 31th European Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Proceedings 31*, page 17–28. Springer-Verlag, 2009. (Cited on page 38.)

Summary

As artificial intelligence systems become increasingly integrated into our daily lives, automated advice-giving systems are transforming how we access information and make decisions. These systems span diverse applications: hiring assistants that filter job applications, medical search engines that retrieve medical information based on patient symptoms, and conversational AI that provides advice through natural language interactions. While these advances create unprecedented opportunities for information access, they also raise critical questions about responsible deployment. This thesis addresses two fundamental challenges in responsible advice-giving: ensuring fairness across individuals and groups, and enhancing explainability to make systems more interpretable for both developers and users.

Part I: Fairness in ranking systems. The first part of this thesis examines fairness in ranking models, challenging two key assumptions that undermine current approaches.

First, we address the widespread assumption that user exposure (i.e., user attention) over ranked documents is known and predictable. In reality, certain ranking configurations, for example those containing visual outliers, produce unpredictable exposure patterns. Since including such ranked lists in the probabilistic ranking policy would prevent us from making meaningful fairness guarantees, we develop a method that reduces the probability of presenting a ranked list with unknown exposure distribution in the ranking policy. Our approach significantly outperforms existing baselines in minimizing rankings with unknown exposure distribution, while maintaining system utility.

Second, we tackle the assumption that document relevance or merit can be perfectly estimated. In practice, there is always some uncertainty that the model exhibits in its predictions. We demonstrate that by allowing ranking adjustments within model confidence bounds, reordering documents based on ranking scores while respecting uncertainty levels, we can substantially reduce bias in top-ranked results with minimal utility loss. This work reveals the potential of leveraging model uncertainty to balance user utility with responsible AI objectives such as fairness, diversity, and reduced bias.

Part II: Explaining advice-giving processes. The second part of this thesis shifts focus to the interpretability of advice-giving systems.

We investigate how ranking models can be made more interpretable by extending the notion of feature attribution explanations, which are commonly used in other domains, to listwise feature attribution for ranking systems. We introduce RankingSHAP as a concrete instantiation of this approach, demonstrating competitive performance with existing methods while offering greater flexibility in examining specific aspects of ranking decisions.

We then address the interpretability in retrieval-augmented generation (RAG) systems achieved through citations that are generated along with large language model-generated answers, aiming to explain the source of a generated piece of information. We introduce the concept of citation faithfulness, requiring that cited sources actually influence answer generation, and distinguish it from citation correctness, which merely requires sources to contain relevant information. Through empirical analysis of a state-of-the-art RAG model, we demonstrate the prevalence of unfaithful citations and reveal concerning post-rationalization behaviors where models retrospectively justify their

outputs.

Conclusion and future directions. This thesis advances responsible advice-giving across multiple dimensions, yet significant challenges remain. Our work highlights the critical need for thorough evaluation of proposed approaches and careful examination of prevailing assumptions in current practices. While we have made meaningful progress on individual components, achieving truly trustworthy and reliable advice-giving systems will likely require a more holistic perspective that integrates fairness, interpretability, and other dimensions of responsible AI deployment.

Samenvatting

Naarmate kunstmatige intelligentiesystemen steeds meer geïntegreerd raken in ons dagelijks leven, veranderen geautomatiseerde systemen die advies geven de manier waarop we toegang krijgen tot informatie en beslissingen nemen. Deze systemen omvatten verschillende toepassingen: systemen in werving en selectie die sollicitaties filteren, medische zoekmachines die medische informatie ophalen op basis van de symptomen van een patiënt, en conversationele AI die advies geeft via natuurlijke taalinteracties. Hoewel deze ontwikkelingen ongekende mogelijkheden scheppen voor informatietoegang, roepen ze ook kritische vragen op over verantwoorde implementatie. Dit proefschrift behandelt twee fundamentele uitdagingen bij verantwoord advies: het waarborgen van eerlijkheid tussen individuen en groepen, en het verbeteren van de uitlegbaarheid om systemen beter interpreteerbaar te maken voor zowel ontwikkelaars als gebruikers.

Deel I: Eerlijkheid in rangschikkingssystemen. Het eerste deel van dit proefschrift onderzoekt eerlijkheid in rangschikkingsmodellen en stelt twee belangrijke aannames ter discussie die de huidige benaderingen ondermijnen.

Ten eerste bespreken we de wijdverbreide aanname dat blootstelling van geordende lijstje van documenten aan gebruikers (en dus gebruikersaandacht) bekend en voorspelbaar is. In werkelijkheid produceren bepaalde rangschikkingsconfiguraties, bijvoorbeeld die met visuele uitschieters, onvoorspelbare blootstellings- en aandachtspatronen.

Omdat het opnemen van dergelijke geordende lijsten in bestaande probabilistische manieren om documenten te ordenen ons beletten om zinvolle garanties over eerlijkheid te geven, ontwikkelen we een nieuwe methode die de kans verkleint dat een geordende lijst met een onbekende verdeling van blootstelling (en dus aandacht) wordt gepresenteerd. Onze aanpak presteert aanzienlijk beter dan bestaande benaderingen in het minimaliseren van geordende lijsten met resultaten waarvan de verdeling van blootstelling en aandacht onbekend is, terwijl de nuttigheid van het systeem behouden blijft.

Ten tweede pakken we de aanname aan dat de relevantie of waarde van documenten perfect kan worden geschat. In de praktijk vertoont een model om documenten te ordenen altijd enige onzekerheid in zijn voorspellingen. We tonen aan dat door aanpassingen toe te staan in de ordening van een model die binnen de betrouwbaarheidsgrenzen van het model vallen, en documenten te herordenen op basis van scores met inachtneming van de onzekerheidsniveaus, we de bias in de hoogst geordende resultaten aanzienlijk kunnen verminderen met minimaal verlies aan nuttigheid. Deze bijdrage onthult de mogelijkheden van het benutten van modelonzekerheid om gebruikersnut in evenwicht te brengen met verantwoorde AI-doelstellingen zoals eerlijkheid, diversiteit, en verminderde bias.

Deel II: Adviesprocessen uitleggen. Het tweede deel van dit proefschrift verschuift de aandacht naar de interpreteerbaarheid van adviessystemen.

We onderzoeken hoe rangschikkingsmodellen beter interpreteerbaar kunnen worden gemaakt door het verklaringen van kenmerkattributie, die vaak in andere domeinen worden gebruikt, uit te breiden naar lijstsgewijs kenmerkattributie voor ordeningssystemen. We introduceren RankingSHAP als een concrete invulling van deze aanpak, waarbij we resultaten behalen die concurrerend zijn met bestaande methoden en tegelijkertijd meer

flexibiliteit verkrijgen om specifieke aspecten van beslissingen over rangschikkingen te sturen en onderzoeken.

Vervolgens behandelen we interpreteerbaarheid in *retrieval-augmented generation* (RAG)-systemen die verkregen door middel van citaties die worden gegenereerd samen met antwoorden die zijn geproduceerd door grote taalmodellen, met als doel de bron van een gegenereerd stuk informatie te verklaren. We introduceren het concept van citatiegetrouwheid, waarbij de geciteerde bronnen daadwerkelijk van invloed (moeten) zijn op de generatie van antwoorden, en onderscheiden dit van citatiecorrectheid, waarbij bronnen alleen relevante informatie moeten bevatten. Door middel van een empirische analyse van een *state-of-the-art* RAG-model tonen we de prevalentie van ontrouwe citaties aan en onthullen we relevant post-rationalisatiegedrag waarbij modellen hun uitkomsten retrospectief rechtvaardigen.

Conclusie en toekomstige richtingen. Dit proefschrift verlegt de grenzen van systemen die op een verantwoorde manier advies geven op meerdere vlakken, maar er blijven aanzienlijke uitdagingen bestaan. Ons werk benadrukt de dringende noodzaak van een grondige evaluatie van voorgestelde benaderingen en een zorgvuldige analyse van de heersende aannames in de huidige praktijk. Hoewel we op individuele onderdelen aanzienlijke vooruitgang hebben geboekt, zal het realiseren van echt betrouwbare systemen die advies geven waarschijnlijk een holistischer perspectief vereisen dat eerlijkheid, interpreteerbaarheid, en andere dimensies van verantwoorde AI-implementatie integreert.

Zusammenfassung

Mit der zunehmenden Integration von Systemen der künstlichen Intelligenzn (KI) in unser tägliches Leben verändern automatische Beratungssysteme die Art und Weise, wie wir auf Informationen zugreifen und Entscheidungen treffen. Diese Systeme umfassen verschiedene Anwendungen: Einstellungsassistenten, die Stellenbewerbungen filtern, medizinische Suchmaschinen, die medizinische Informationen auf der Grundlage von Patientensymptomen abrufen, und konversationelle KI, die mithilfe von natürlicher Sprachinteraktion Ratschläge erteilt. Während diese Fortschritte nie dagewesene Möglichkeiten des Informationszugangs schaffen, werfen sie auch kritische Fragen zum verantwortungsvollen Einsatz auf. Diese Arbeit befasst sich mit zwei grundlegenden Herausforderungen bei der verantwortungsvollen Erteilung von Ratschlägen: der Gewährleistung von Gerechtigkeit zwischen Einzelpersonen und Gruppen und der Verbesserung der Erklärbarkeit, um Systeme sowohl für Entwickler als auch für Benutzer besser interpretierbar zu machen.

Teil I: Fairness in Rankingsystemen. Der erste Teil dieser Dissertation befasst sich mit der Gerechtigkeit in Ranking-Modellen (Modellen zur Sortierung von Dokumenten entsprechend ihrer Relevanz), und stellt zwei zentrale Annahmen in Frage, die aktuelle Ansätze untergraben.

Zunächst diskutieren wir die weit verbreitete Annahme, dass die Sichtbarkeit geordneter Dokumentlisten für Nutzer (und damit deren Aufmerksamkeit) bekannt und vorhersehbar sei. In der Realität führen bestimmte Ranking-Konfigurationen, beispielsweise solche mit visuellen Ausreißern, zu unvorhersehbaren Sichtbarkeits- und Aufmerksamkeitsmustern. Da die Einbeziehung solcher geordneten Listen in bestehende probabilistische Methoden zur Ordnung von Dokumenten uns daran hindert, aussagekräftige Garantien hinsichtlich der Fairness zu geben, entwickeln wir eine neue Methode, die die Wahrscheinlichkeit verringert, dass eine geordnete Liste mit einer unbekannten Verteilung der Aufmerksamkeit präsentiert wird. Unser Ansatz übertrifft bestehende Ansätze bei der Minimierung geordneter Dokumentlisten, bei denen die Verteilung von Sichtbarkeit und Aufmerksamkeit unbekannt ist, erheblich, während gleichzeitig die Qualität des Systems erhalten bleibt.

Zweitens befassen wir uns mit der Annahme, dass die Relevanz oder der Wert von Dokumenten perfekt geschätzt werden kann. In der Praxis gibt es immer eine gewisse Unsicherheit, die das Modell in seinen Vorhersagen aufweist. Wir zeigen, dass wir ungewünschte thematische Färbungen der Top-Ranking-Ergebnisse bei minimalem Qualitätsverlust erheblich reduzieren können, durch Anpassungen der Ordnung der Dokumente innerhalb der Unsicherheit des Modells. Dies zeigt das Potenzial der Nutzung von Modellunsicherheiten auf, um den idealen Nutzen für den Benutzer mit verantwortungsvollen KI-Zielen wie Gerechtigkeit, Vielfalt und der reduzierter unerwünschter Inhalte zu kombinieren.

Teil II: Erklärung der Prozesse der Ratschlagserteilung. Im zweiten Teil dieser Dissertation wird der Fokus auf die Interpretierbarkeit von ratgebenden Systemen gelegt.

Wir untersuchen, wie Ranking-Modelle interpretierbarer gemacht werden können, indem wir den Begriff der Wichtigkeit von Merkmalen zur listenleise Wichtigkeit für Ranking-Systeme ausweiten. Wir stellen RankingSHAP als eine konkrete Anwendung dieses Ansatzes vor, die mit bestehenden Methoden konkurrieren kann und

gleichzeitig mehr Flexibilität bei der Untersuchung spezifischer Aspekte von Ranking-Entscheidungen bietet.

Anschließend befassen wir uns mit der Interpretierbarkeit in *Retrieval-Augmented Generation* (RAG)-Systemen, die durch Zitate erreicht wird, die zusammen mit großen Sprachmodell-generierten Antworten generiert werden, mit dem Ziel, die Quelle einer generierten Information zu erklären. Wir führen das Konzept der Zitattreue ein, das voraussetzt, dass zitierte Quellen tatsächlich die Generierung von Antworten beeinflussen, und unterscheiden es von der Zitatkorrektheit, die lediglich voraussetzt, dass Quellen relevante Informationen enthalten. Durch die empirische Analyse eines modernen *state-of-the-art* RAG-Modells demonstrieren wir die Prävalenz von untreuen Zitaten und zeigen die Tendenz des Modells zur nachträglichen Rationalisierung, bei der sie ihre Ergebnisse rückwirkend durch Zitate gerechtfertigt werden.

Schlussfolgerung und zukünftige Richtung. Diese Dissertation bringt die Grenzen verantwortungsvoller Beratungssysteme in mehreren Dimensionen voran, dennoch bleiben weiterhin erhebliche Herausforderungen. Diese Arbeit unterstreicht die Notwendigkeit einer gründlichen Evaluierung der vorgeschlagenen Ansätze und einer sorgfältigen Prüfung der vorherrschenden Annahmen in der derzeitigen Praxis. Obwohl wir in Bezug auf einzelne Komponenten bedeutende Fortschritte erzielt haben, erfordert das Erreichen wirklich vertrauenswürdiger und zuverlässiger Beratungssysteme voraussichtlich eine ganzheitlichere Perspektive, die Gerechtigkeit, Interpretierbarkeit und andere Dimensionen eines verantwortungsvollen KI-Einsatzes berücksichtigt.