Mitigating Gender Bias in Neural Retrieval Systems

by

Shirin Seyedsalehi

M.Sc. in Electrical Engineering

Amirkabir University of Technology, Tehran, Iran, 2019

B.Sc. in Electrical Engineering

Amirkabir University of Technology, Tehran, Iran, 2016

A dissertation

presented to Toronto Metropolitan University

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Program of

Electrical and Computer Engineering

Toronto, Ontario, Canada, 2025

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A DISSERTATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Toronto Metropolitan University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Toronto Metropolitan University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

Mitigating Gender Bias in Neural Retrieval Systems

Doctor of Philosophy 2025

Shirin Seyedsalehi

Electrical and Computer Engineering

Toronto Metropolitan University

Abstract

With the widespread adoption of neural ranking models in modern information retrieval (IR) systems, concerns about fairness—particularly regarding gender bias—have gained urgency. While these models achieve state-of-the-art performance, they also risk learning and amplifying societal stereotypes embedded in the data and representation spaces. This thesis presents a comprehensive investigation into the sources and mitigation of gender bias in neural IR systems, focusing on three critical components: training strategies, embedding representations, and data sampling. By addressing bias across all levels of the learning pipeline, this work aims to promote the development of equitable, high-performance IR systems.

The first major contribution of this thesis introduces a set of bias-aware training strategies that incorporate fairness directly into the model's optimization objectives. By modifying traditional loss functions to penalize biased documents and reward unbiased ones, the learning process is steered toward equitable ranking behavior. A bias penalty term is defined based on document-level gender bias signals, which dynamically adjusts the influence of each training instance during learning. Experimental results demonstrate that this regularization approach not only reduces gender bias in retrieval outputs but also improves the fairness-performance trade-off compared to existing mitigation methods.

The second contribution targets the embedding space where gender bias is often implicitly encoded. A novel disentangled representation learning framework is proposed to isolate gender-specific and content-relevant components within neural embeddings of query-document pairs. This framework employs a dual-objective training strategy that simulta-

neously optimizes a ranking loss and a gender classification loss, ensuring that only the content-relevant component contributes to ranking decisions. By decoupling gender information from the relevance signal, the model avoids propagating stereotypical associations, thereby achieving fairness without compromising retrieval effectiveness. The method shows strong generalization across datasets and embedding architectures.

The third contribution addresses the bias present in training data through a curriculum-inspired sampling strategy. Motivated by the principles of curriculum learning, the proposed method structures the training process such that the model is initially trained on gender-neutral samples before gradually incorporating more biased instances. This staged exposure allows the model to first learn robust and unbiased relevance patterns, reducing the risk of bias propagation from the outset of training. A dynamic sampling mechanism governs the introduction of biased examples over time, guided by statistical estimates of document-level gender bias. The results confirm that such a curriculum-aware approach can significantly reduce bias in IR systems, especially when integrated with other mitigation strategies.

Together, these contributions form a unified framework for mitigating gender bias in neural IR systems through principled interventions at multiple stages of the pipeline. Comprehensive experiments on real-world datasets demonstrate the effectiveness of the proposed methods in reducing harmful stereotypes while preserving or even enhancing retrieval performance. This thesis thus contributes novel methodologies and insights toward the development of socially responsible, fair, and high-quality information retrieval technologies.

Acknowledgements

There are many people to whom I owe my deepest gratitude for supporting me throughout this journey. First and foremost, I would like to express my sincere appreciation to my supervisor, Dr. Ebrahim Bagheri. He has been far more than an academic supervisor. His unwavering support, both intellectually and emotionally, has guided me through every step of my Ph.D. journey. I have learned not only about research and scholarship from him, but also about life, integrity, and the importance of valuing others. He is a true "teacher" in every sense of the word. I truly feel indebted to him. I am also deeply grateful to my co-supervisor, Dr. Morteza Zihayat, for his constant support and kindness. His mentorship has been invaluable, and the lessons I've learned from him will remain with me for life.

My heartfelt thanks go to my friends and labmates, who made this journey so much more joyful and memorable. I have been incredibly fortunate to share these years with such inspiring and supportive colleagues.

To my family—my mother and father—thank you for your boundless love, unconditioned support, and selfless sacrifices, which have been the foundation of everything I have achieved. To my sister, your love and encouragement have always warmed my heart. I am also grateful to my extended family for their continued support and love.

Finally, to my dear partner, Amin, thank you for always believing in me. Your pure love, and constant encouragement have given me the courage to persevere through the most challenging moments. You have stood by me with patience and compassion, cheering me on even when I doubted myself. Your presence has been my greatest source of peace and motivation.

Shirin Seyedsalehi

Dedication

To my beloved mom and dad,

Elham and Nickan

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Information Retrieval (IR) systems are fundamental to the digital era, and crucial for navigating the vast data landscape of today's world. From simple web searches to sophisticated data analytics in corporate environments, IR systems are integral to modern life and provide the tools necessary for personal and professional decision-making. IR systems do not just facilitate over 1.2 trillion searches per year on a platform like Google [65] but also significantly impact various sectors such as:

- Healthcare. In healthcare, IR systems manage extensive patient records and research databases, enabling medical professionals to access vital information swiftly. For instance, databases like PubMed offer access to medical research, facilitating better patient care and fostering the rapid development of medical knowledge [92].

- Finance and Banking. Financial sectors utilize IR to analyze market trends and monitor transactions. Tools provided by Bloomberg and Reuters help professionals sift through large datasets to find critical information on market developments, economic reports, and investment analytics, supporting quick and informed financial decisions [79, 109].

- Legal. IR systems such as LexisNexis and Westlaw are indispensable in the legal arena. They allow legal professionals to efficiently search through vast quantities of legal documents, case law, and statutes, essential for case preparation, conducting due diligence, and ensuring comprehensive legal research [77, 132].

- Academic Research. IR systems are also crucial in academia, where platforms like Google Scholar and JSTOR enable researchers to navigate through countless scholarly articles and publications. This access supports various academic disciplines, enhancing research capabilities and fostering educational advancement [57, 66].

Such systems have deep impacts on different aspects of society. The economic implications of IR systems are vast, influencing sectors from e-commerce to online advertising. They drive consumer behavior, facilitate transactions, and are instrumental in strategic business decisions, impacting billions in daily commerce. Technological advancements in IR have paralleled the rapid evolution of computing power and data science methodologies. Today's IR systems employ sophisticated algorithms and machine-learning techniques to improve accuracy and user experience. Furthermore, IR systems profoundly shape societal interactions and access to information, influencing education, politics, and social dynamics. In education, IR systems provide students and academics access to a wide array of resources, transforming how knowledge is acquired and shared. The availability of digital libraries and online courses has democratized education, making learning more accessible globally. Politically, IR systems play a critical role in shaping public opinion and electoral outcomes by controlling the flow of news and information. Their ability to highlight or suppress information can alter perceptions and influence decisions on a large scale. Culturally, IR systems facilitate the global exchange of ideas and values, promoting cross-cultural understanding and cooperation [121]. They have become platforms for cultural expression and identity exploration, contributing to the global cultural mosaic.

1.2 Motivation

Information retrieval (IR) systems have become integral to critical aspects of modern life, influencing decisions in domains such as healthcare, recruitment, judicial processes, education, and financial services. These systems, often designed to assist humans in making complex decisions, carry the responsibility of operating fairly and inclusively. However, the presence of gender biases in IR systems poses a significant threat to their reliability and societal impact. If not addressed, these biases can lead to prejudiced outcomes, reinforce harmful stereotypes, and exacerbate systemic inequalities, undermining the very purpose of these technologies.

Consider the example of a company employing an AI-powered system to shortlist candidates for a "data scientist" position. If the system prioritizes male applicants due to inherent biases in its design or training data, the consequences extend far beyond a single hiring decision. Highly qualified female candidates may be overlooked, depriving the company of valuable talent and diverse perspectives. This perpetuates gender discrimination and limits the inclusivity of the workplace, ultimately leading to suboptimal outcomes for the organization. Furthermore, by reinforcing stereotypes that men are more suited to technical roles, such biases discourage women from pursuing careers in STEM fields, perpetuating a cycle of inequality.

The risks associated with biased IR systems are even more pronounced in the healthcare sector. AI-driven treatment recommendation systems, which are increasingly relied upon by physicians, can fail to account for differences in how diseases manifest across genders. For example, women often exhibit different symptoms for conditions like heart disease, yet many medical datasets predominantly represent male patients. A biased system trained on such data could misdiagnose or inadequately treat conditions affecting women, leading to delayed care, poorer health outcomes, and even preventable fatalities. The consequences of such biases are not just personal but societal, as they undermine trust in AI-driven healthcare technologies and exacerbate existing disparities in access to quality care.

Biases in IR systems also raise significant ethical concerns in the judicial system. Algorithms used for risk assessment or sentencing recommendations can exhibit biases not only against genders but also against intersecting attributes such as race or socioeconomic status. If a system unjustly evaluates women as being less "trustworthy" or disproportionately flags individuals from marginalized communities, it risks perpetuating systemic discrimination and undermining public trust in the fairness of the justice system. These biases do not exist in isolation; they ripple through society, amplifying inequalities and eroding confidence in institutions.

Even in education, where IR systems are used to recommend learning materials and career opportunities, biases can have profound long-term consequences. If such systems disproportionately promote STEM-related resources to male students while steering female students toward other fields, they reinforce traditional gender roles and limit the diversity of representation in certain industries. This, in turn, affects the career trajectories of entire generations, perpetuating inequalities that extend far beyond the classroom.

These examples collectively highlight the urgent need to address biases in IR systems. The consequences of unchecked bias are far-reaching, influencing individual lives, organizational outcomes, and societal structures. Fair and inclusive systems are not only a moral and ethical necessity but also essential for the effective and equitable application of AI technologies. Systems that are free from bias foster trust, enhance decision-making, and unlock opportunities for individuals across all genders. Moreover, they promote diversity, which has been shown to drive innovation and improve outcomes across various domains.

The motivation for this research lies in the recognition that fair and unbiased IR systems are foundational to a just and inclusive society. By systematically identifying and mitigating gender biases in these systems, we can ensure they fulfill their potential as tools for empowerment rather than mechanisms of inequity. Addressing these challenges is not merely a technical endeavor; it is a commitment to building a future where technology serves all individuals equitably, contributing positively to societal growth and progress.

The primary objective of this doctoral research is to develop methods that promote fairness and mitigate gender biases in ranking systems while preserving their effectiveness. The proposed approaches encompass three main strategies: data-driven methods, Bias mitigation in neural embeddings, and bias-aware training strategies. These strategies aim to address gender bias at different stages of the ranking process. To achieve this, the following research questions (RQs) have been defined:

1.3 Research Question 1 (RQ1)

Incorporating Bias Awareness into Training Strategies. The first research question examines how bias-aware training strategies, particularly through modifications to the loss function, can mitigate gender biases. The question is: **Can penalizing or rewarding the relevance scores of biased and unbiased documents during training result in a less biased ranking system?** The loss function plays a critical role in model training as it directly informs the model about performance and influences weight updates [56]. If the loss function can account for biases and guide the model toward fairness, it can significantly reduce the impact of gender bias. To explore this, I propose two scenarios: 1) Penalizing Biased Documents: where, biased documents are assigned a penalty by increasing the loss associated with them. This approach makes the model aware of the negative impact of biased documents and discourages their prioritization in rankings. 2) Rewarding Unbiased Documents:Here, unbiased documents are rewarded by reducing the loss associated with them. This incentivizes the model to prioritize unbiased documents, reinforcing fairness in rankings.

While dense retrievers have demonstrated strong retrieval effectiveness, they have also been shown to exhibit gender bias in their ranking results [10, 18, 116, 73, 71]. Studies have consistently reported that queries with neutral intent, such as "How to become an engineer?" or "Best tips for parenting," often return ranked lists that disproportionately

emphasize one gender over another [74]. These biases manifest as an overrepresentation of gendered contexts in the retrieved documents, which can reinforce existing stereotypes. Such skewed results undermine the fairness of information retrieval systems and highlight the inherent challenges of ensuring equity in search. Rekabsaz et al. [107] conducted an extensive study to measure the societal biases present in neural rankers, revealing significant gender disparities in retrieval outputs. Their analysis demonstrated that neural ranking models not only propagated biases present in training data but also amplified these biases in their ranking results. They found that certain queries with professional or neutral wording led to disproportionately gendered representations, with stereotypical roles being reinforced. This systematic evaluation provided concrete evidence of the biased behavior of neural rankers and emphasized the need for frameworks to address these issues systematically.

The biases exhibited by dense neural rankers can be attributed to several interrelated factors [40, 39]. First, the training data used for pretraining and fine-tuning these models often reflects societal stereotypes and historical inequities, which the models inadvertently encode and amplify in their embeddings. For instance, Bolukbasi et al. [21] demonstrated how word embeddings derived from large-scale corpora exhibit gendered associations, such as linking "programmer" with male-oriented terms and "nurse" with female-oriented ones. Second, the contextual representations generated by neural models, although *semantically rich*, can lack granularity in distinguishing between neutral and biased attributes, further contributing to skewed ranking results [3]. Third, the optimization objectives of dense retrievers prioritize relevance, often measured through *ranking metrics*, without explicitly incorporating fairness considerations [116]. This tendency can exacerbate biases when rankers disproportionately emphasize features linked to gendered patterns in the data.

Among the three issues identified earlier, the focus of our work in this research question is on the optimization objective of dense retrievers as it directly shapes the behavior of the ranker. This work aims to design a *fair ranker* that balances relevance and fairness in retrieval results. A fair ranker ensures that the ranked documents maintain *high retrieval*

effectiveness while mitigating biases, such as disproportionate gender representations, that can reinforce harmful stereotypes. However, dense retrievers are not naturally incentivized to address fairness due to their optimization objectives, which are narrowly focused on maximizing *relevance*. These objectives are typically implemented through loss functions, which guide the training process [11]. One potential solution to this problem is to regularize the loss function by incorporating *fairness constraints*. This modification may enable the training process to jointly optimize for both *relevance* and *fairness*, ensuring that dense retrievers produce results that are not only effective but also equitable.

In particular, we propose a framework to address biases in dense retrievers by explicitly regularizing the loss function to incorporate fairness constraints. The rationale behind this approach lies in the dual role of the loss function: it serves as the optimization objective, guiding the ranker's training process, and as a mechanism to encode priorities such as *relevance* and *fairness*. By augmenting the standard loss with fairness-specific regularization terms, the ranker can be trained to balance these objectives without compromising *retrieval effectiveness*. Specifically, we extend the traditional relevance-based loss function with additional terms that penalize rankings exhibiting higher levels of bias. These terms are derived from measures such as the degree of gender bias in retrieved documents and are incorporated into both *pointwise* [32, 70] and *pairwise* [64, 29] ranking frameworks. For *pointwise rankers*, the regularization adjusts the predicted relevance scores based on the bias level of individual documents, while for *pairwise rankers*, the adjustment ensures that fairness is enforced in the relative ordering of document pairs. A hyperparameter controls the trade-off between *relevance* and *fairness* during training, allowing the ranker to adapt to specific fairness requirements. This formulation ensures that the ranker simultaneously optimizes for *effectiveness* and *fairness*.

1.4 Research Question 2 (RQ2)

Mitigating Bias in Neural Embeddings. The second research question focuses on the intermediate vector representations generated by neural embeddings, asking: **Can disentangling gender information from the vector representation of query-document pairs reduce gender bias in neural rankers?** This research builds on the hypothesis that separating gender-related information from content-related information in vector representations can lead to a reduction in bias. Specifically, I propose disentangling gender information from the embeddings and calculating query-document relevance solely based on the content-related portion of the representation. By excluding gender information from the ranking process, the model learns to determine relevance purely based on content, preventing it from associating gender with relevance. This disentanglement ensures that the ranking system does not propagate or amplify stereotypical gender biases.

To demonstrate biases in training data, we analyzed 10,000 evenly distributed and randomly selected male and female affiliated queries from the MS MARCO passage ranking task. Results showed that only about 16% of male queries returned documents predominantly affiliated with females, while over 22% of female queries returned predominantly male-affiliated documents, indicating a 6% higher likelihood for female queries to retrieve male-affiliated content. This suggests that neural rankers, once trained, exhibit a preference for male-affiliated documents, irrespective of the query's gender affiliation. Furthermore, male-affiliated queries consistently show a higher similarity score (BM25) with their relevant documents than female-affiliated queries, as depicted in Figure 1.1, making them practically easier to retrieve and satisfy.

These observations, alongside earlier research [105, 15, 116, 68] is the foundation for the work presented in this paper. The *main objective* is to mitigate gender biases in neural rankers while preserving their retrieval effectiveness. We hypothesize that isolating gender as an attribute from query and document representations could reduce bias. If gender is not encoded in the representations, it cannot intensify gender biases, given the observed

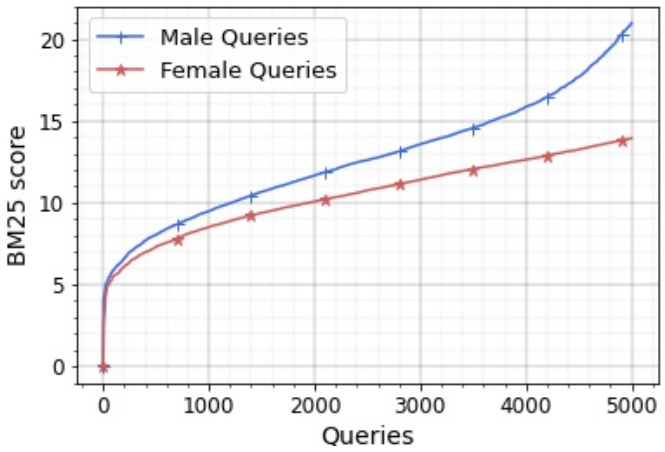

Figure 1.1: The distribution of queries and their BM25 scores calculated with their relevance judgement.

disparities in relevance judgements and varying similarity scores across different gender affiliations. To this end, we propose a neural architecture that disentangles gender from content semantics when encoding a query or a document. In the disentanglement process, the neural representation is broken down systematically into two distinct and independent components, one of which captures the semantics of the content, while the other encapsulates the gender affiliation of the query or the document.

1.5 Research Question 3 (RQ3)

Addressing Bias in Training Data. The third research question aims to optimize fairness by exploring data sampling techniques. The question is: **Does the order of presenting samples to the model affect the fairness of the IR system?** This idea draws inspiration from curriculum learning approaches, which hypothesize that the sequence of presenting training samples can significantly influence model performance. In traditional curriculum learning, easier samples are shown to the model in the early stages, allowing it to build a strong foundational understanding before tackling harder samples, ultimately improving performance [129, 6]. Applying this concept to fairness, I hypothesize that showing unbiased samples to the model during the initial stages of training, followed by biased samples later, can reduce gender bias.

The rationale is that in the early phases of training, the model focuses on learning the fundamental concept of query-document relevance without being influenced by biased data. By delaying exposure to biased samples, the model can avoid associating gender bias with relevance. To operationalize this, I will propose a sampling strategy where unbiased samples are given a higher probability of being selected in the early training stages, and biased samples are introduced gradually. This approach enables the model to prioritize fairness during learning.

Existing approaches to mitigating bias in neural rankers largely focus on either data-level debiasing [12, 14] or modifying the model's learning objectives [105, 116, 137], but both methods come with inherent limitations. Data-level debiasing strategies aim to reduce explicit biases by removing or modifying biased samples before training. While this can help in mitigating biases in the input data, it risks discarding valuable information, which may compromise relevance due to enforced modifications. Additionally, these strategies do not account for the dynamic nature of model training, where bias exposure ideally needs to be controlled and adapted progressively rather than predetermined at the outset. Another group of methods modify the model's training objectives, typically by introducing regularization terms or bias-specific penalties in the loss function [105, 116, 137]. Although these methods integrate bias mitigation directly within the training process, they fundamentally alter the nature of relevance learning. By requiring the model to optimize simultaneously for fairness and relevance, such strategies may lead to trade-offs that can degrade ranking performance [47, 100]. As such, this can prevent the model from distinguishing between relevance and bias in a structured manner, potentially embedding biased patterns into the relevance criteria itself.

In order to understand how biases, more specifically in this case gender biases, are distributed in the training datasets, we visualize gender bias distribution within the MS MARCO passage ranking dataset [89] in Figure 1.2. The figure adopts the gender bias metric in [107] to quantify bias in each document of the training dataset. According to the figure,

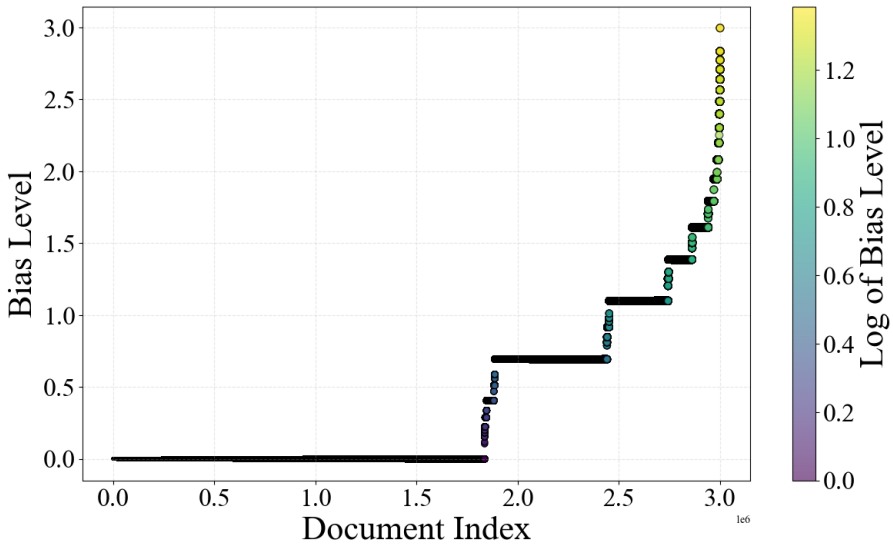

Figure 1.2: Distribution of gender bias scores in the MS MARCO passage ranking dataset.

a significant portion of documents exhibit minimal or no bias, while only a subset displays higher bias levels. This observation introduces an alternative paradigm for mitigating bias in neural ranking models: rather than modifying the data itself to reduce bias or altering the loss function to enforce fairness constraints, we can instead control the order and exposure of biased documents during training. That is, by initially training the model on low-bias documents, we enable it to establish a robust baseline of relevance learning that is relatively unaffected by bias. As the training progresses, we gradually introduce higher-bias samples, allowing the model to incrementally adapt without embedding biased characteristics into its core relevance function.

While this paradigm offers a promising alternative to traditional debiasing methods [19, 14, 18, 137], it introduces several critical challenges. *First*, designing a curriculum that sequences biased and unbiased samples effectively is non-trivial; the curriculum must ensure that relevance learning is prioritized, even as the model gradually encounters more biased samples. If the sequence is not carefully structured, there is a risk that the model may internalize biases prematurely or fail to establish a robust baseline for relevance. Additionally, implementing a dynamic sampling strategy that adapts bias exposure throughout training is essential. This strategy must modify the exposure to bias in a way that aligns

with the model's progress and ensures that training stability and convergence are maintained while controlling for bias integration.

1.6 Contributions

The concrete contributions of this research are:

1. We propose a *customizable framework* for fairness-aware ranking that is applicable to both *pointwise* and *pairwise* ranking models, incorporating fairness constraints directly into their loss functions to address biases in the ranked results. We design and implement *fairness-aware regularization terms*, enabling the optimization process to balance *relevance* and *fairness* effectively. We operationalize fairness by introducing *penalty* and *reward mechanisms* that adjust relevance scores based on *document-level* and *list-level fairness criteria*, ensuring the ranker mitigates biases while preserving relevance. We evaluate our framework on *two benchmark datasets* consisting of *gender-neutral queries* that are not expected to exhibit gender biases in their retrieval results. We further evaluate our proposed framework on rankers trained on different *large language models* to show its generalizability and against a host of strong *state-of-the-art baselines* to demonstrate its effectiveness.

2. We propose a neural ranking architecture that disentangles content semantics from gender affiliation information and offers two independent representation components that encode each of these aspects separately; Given the disentangled representations for queries and documents, we propose to use the content semantics component of the disentangled representation to rank-order documents in relation to the query and minimize the influence of gender and any associated biases in the ranking process; Our extensive experiments demonstrate that: (1) the disentanglement process significantly reduces stereotypical gender biases in retrieved documents; (2) this reduction does not compromise retrieval effectiveness but rather enhances it; and (3) disentangling

neural representations improves performance parity across various gender affiliations and query subsets.

3. We define the problem of using curriculum learning as a structured approach for managing bias exposure in neural rankers. Our approach leverages the sequence and progression of training samples to manage bias effectively while preserving relevance learning. We introduce a novel bias-aware curriculum that sequences training samples based on bias scores, prioritizing low-bias documents in the initial training phases and gradually incorporating higher-bias samples. We propose a dynamic sampling strategy that adjusts the probability of sampling documents based on their bias scores. The sampling strategy facilitates gradual bias exposure without compromising the model's convergence. We conduct extensive experiments to evaluate the effectiveness of our curriculum-based approach. These experiments demonstrate the utility of the proposed approach and validate its effectiveness in comparison to existing bias mitigation techniques.

1.7 Structure of the Thesis

This thesis is organized as follows:

- Chapter 2 - Literature Review: This chapter provides a comprehensive review of prior research in the fields of gender biases in Machine learning, natural language processing, and information retrieval.

- Chapter 3 - Problem Definition: This chapter introduces the problem statement and defines the foundational concepts required for the rest of the work.

- Chapter 4 - Loss Function Regularization: The focus of this chapter is to present a set of bias-aware training strategies that embed fairness considerations directly into the model's optimization process. This is achieved by altering standard loss functions

to discourage biased content and encourage fair representations, thereby guiding the model toward more equitable ranking outcomes. A key component of the approach is a bias penalty term, derived from document-level gender bias indicators, which dynamically modulates the impact of each training sample. Through experimental evaluation, the chapter validates whether this regularization framework effectively reduces gender bias in retrieval results comparing to the existing bias mitigation techniques.

- Chapter 5 - De-biasing Neural Embeddings: In this chapter, we propose a method for addressing gender bias in neural ranking models by targeting the embedding space where such bias is often implicitly encoded. The core of this method is a disentangled representation learning framework that aims to separate gender-related signals from content-relevant information in query-document embeddings. To achieve this, we design a dual-objective training strategy that simultaneously optimizes for ranking performance and gender attribute separation—ensuring that only the content-relevant components influence ranking decisions. By explicitly decoupling gender information from relevance signals, this approach is intended to mitigate the propagation of biased associations, enabling fairer retrieval outcomes without degrading effectiveness. We have conducted experiments to evaluate the performance of the proposed method.

- Chapter 6 - Bias-aware Curriculum Sampling: This chapter studies the mitigation of bias in training data through a curriculum-inspired sampling strategy. Drawing on the principles of curriculum learning, it proposes a novel training methodology that structures the learning process to begin with gender-neutral examples, gradually incorporating biased instances as training progresses. The goal is to allow the model to first acquire stable and unbiased relevance signals, thereby minimizing early exposure to harmful bias patterns. A dynamic sampling mechanism is introduced to regulate this progression, leveraging statistical estimates of document-level gender bias to inform sample selection over time. This chapter argues that such a curriculum-aware training

pipeline not only reduces the propagation of bias but also complements other mitigation techniques, leading to more fair and robust information retrieval systems.

- Chapter 7 - Conclusions: This chapter summarizes the findings of the thesis, discusses the conclusions drawn from the research, and outlines potential future work.

1.8 Publications

Publications Arising From This Thesis

Journal Papers

- Understanding and Mitigating Gender Bias in Information Retrieval Systems, Shirin Seyedsalehi, Amin Bigdeli, Negar Arabzadeh, Batool AlMou-sawi, Zack Marshall, Morteza Zihayat and Ebrahim Bagheri, Foundations and Trends® in Information Retrieval (Impact Factor: 10.4), accepted, 2024.

- Gender Disentangled Representation Learning in Neural Rankers, Shirin Seyedsalehi, Sara Salamat, Negar Arabzadeh, Sajad Ebrahimi, Morteza Zihayat and Ebrahim Bagheri, Machine Learning Journal (Impact Factor: 4.3), *accepted, 2024.*

- A Generalizable Framework for Bias Mitigation in Dense Neural Rankers, Shirin Seyedsalehi, Morteza Zihayat, Ebrahim Bagheri, Machine Learning Journal (Impact Factor: 4.3), *Submitted, 2025.*

Conference Papers

- Bias-aware Curriculum Sampling for Fair Ranking, Shirin Seyedsalehi, Hai son Le, Morteza Zihayat and Ebrahim Bagheri, *The 48TH International ACM SIGIR Conference on Research and Development in Information Retrieval; SIGIR2025 (Core Rank: A*).*

- De-Biasing Relevance Judgements for Fair Ranking, Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Bhaskar Mitra, Morteza Zihayat and Ebrahim Bagheri, *45th European Conference on Information Retrieval; ECIR2023 (Core Rank: A).*

- Addressing Gender-related Performance Disparities in Neural Rankers, Shirin Seyedsalehi, Amin Bigdeli, Negar Arabzadeh, Morteza Zihayat, Ebrahim Bagheri, *The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval; SIGIR 2022 (Core Rank: A*).*

- Bias-aware Fair Neural Ranking for Addressing Stereotypical Gender Biases, Shirin Seyedsalehi, Amin Bigdeli, Negar Arabzadeh, Bhaskar Mitra, Morteza Zihayat, Ebrahim Bagheri, *25th International Conference on Extending Database Technology; EDBT2022 (Core Rank: A).*

- On the Orthogonality of Bias and Utility in Ad hoc Retrieval, Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Morteza Zihayat and Ebrahim Bagheri, *The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval; (2021 (Core Rank: A*).*

- A Light-weight Strategy for Restraining GenderBiases in Neural Rankers, Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Morteza Zihayat and Ebrahim Bagheri, *44th European Conference on Information Retrieval; ECIR2022 (Core Rank: A)*

Other Publictions During Ph.D.

- Reinforcement Learning for Effective Few-Shot Ranking, Shiva Soleimani, Sajad ebrahimi, Shirin Seyedsalehi, Fattane Zarrinkalam, and Ebrahim Bagheri, *The 48TH International ACM SIGIR Conference on Research and Development in Information Retrieval; SIGIR2025 (Core Rank: A*).*

- Neural Disentanglement of Query Difficulty and Semantics, Sara Salamt, Negar Arabzadeh,

Shirin Seyedsalehi, Amin Bigdeli, Morteza Zihayat, ebrahim Bagheri, *32nd ACM International Conference on Information and Knowledge Management; CIKM2023 (Core Rank: A*)

- Don't Raise Your Voice, Improve Your Argument: Learning to Retrieve Convincing Arguments, Sara Salamat, Negar Arabzadeh, Amin Bigdeli, Shirin Seyedsalehi, Morteza Zihayat and Ebrahim Bagheri, *45th European Conference on Information Retrieval; ECIR2023 (Core Rank: A*).

- A Neural Approach to Forming Coherent Teams in Collaboration Networks, Shirin Seyedsalehi, Radin Hamidi Rad, Mehdi Kargar, Morteza Zihayat, Ebrahim Bagheri, *25th International Conference on Extending Database Technology; EDBT2022 (Core Rank: A*).

- Matches Made in Heaven: Toolkit and Large-Scale Datasets for Supervised Query Reformulation, Negar Arabzadeh, Amin Bigdeli, Shirin Seyedsalehi, Morteza Zihayat, Ebrahim Bagheri, *30th ACM International Conference on Information & Knowledge Management; CIKM2021 (Core Rank: A).*

Chapter 2

Literature Review

2.1 Gender Bias in Machine Learning

Gender bias in machine learning (ML) has emerged as a critical concern in recent years, particularly as ML models are increasingly deployed in high-stakes applications such as hiring, healthcare, and criminal justice. Gender bias in ML arises primarily from biased data, algorithmic decision-making, and reinforcement of societal stereotypes. Data bias stems from historical inequalities, underrepresentation of specific genders, and skewed distributions of attributes within datasets. Algorithmic bias occurs when the design of the ML model or the training objective amplifies existing biases in the data. For instance, algorithms designed to optimize accuracy may inadvertently prioritize the majority group at the expense of fairness [24]. Moreover, feedback loops in dynamic systems can exacerbate bias over time, as biased outputs influence future training data [53].

Gender bias manifests across various domains of machine learning. In the field of computer vision, gender bias becomes starkly apparent in facial recognition systems. Research by Buolamwini and Gebru [24] highlighted that commercial facial recognition algorithms exhibit significantly higher error rates for women, particularly women of color, compared to lighter-skinned men. Subsequent analyses by Raji et al. [103] have confirmed these findings,

noting that such disparities often stem from the underrepresentation of diverse demographic groups in training datasets. More recent research by Drozdowski et al. [45] points to similar issues in biometric authentication systems, where demographic imbalances lead to disproportionately high false rejection rates for women and minorities. These biases can lead to serious implications, including misidentification and the reinforcement of societal prejudices.

Recommendation systems are another area where gender bias is prevalent. Algorithms designed to optimize user engagement often amplify existing disparities by disproportionately recommending stereotypical content to users. Research has highlighted that job recommendation systems can inadvertently perpetuate gender biases, steering women away from roles in male-dominated industries like technology and engineering. For instance, a study by Rus et al. [112] demonstrated that without controlling for bias, women are recommended jobs with significantly lower salaries, indicating a systemic issue in job recommendation algorithms. Mehrotra et al. [23] reveal that these systems inadvertently marginalize users from underrepresented genders, limiting their access to diverse opportunities. Their study explores how feedback loops in recommendation algorithms can amplify popularity bias, leading to decreased aggregate diversity and homogenization of user experiences. Notably, the research highlights that the impact of feedback loops is more pronounced for users in minority groups, thereby perpetuating exclusion cycles for marginalized communities.[82]

In healthcare, gender bias can have life-threatening consequences. Machine learning models trained on historical medical data often underperform for women, as demonstrated in diagnostic tools for cardiovascular diseases [86]. This underperformance is largely attributed to the historical underrepresentation of women in clinical studies. Obermeyer et al. [91] further highlighted that algorithmic decisions in healthcare often reflect systemic biases, leading to disparities in treatment recommendations and outcomes. Recent work by Cheong et al. examined gender bias in mental health prediction models, identifying both data and algorithmic biases. Their study also evaluated various mitigation strategies across different stages of model development to enhance gender fairness [31].

Addressing gender bias in machine learning requires interventions at multiple levels, including the data, algorithms, and evaluation frameworks. At the data level, techniques such as data augmentation and preprocessing have been widely explored. Data augmentation involves generating synthetic examples to increase the representation of underrepresented groups, as discussed by Zhao et al. [139]. Another approach is bias detection and quantification, which uses tools like the Word Embedding Association Test (WEAT) to identify stereotypical associations in datasets [27]. Additionally, preprocessing methods such as the debiasing of embeddings, pioneered by Bolukbasi et al. [21], aim to neutralize gender associations before they are propagated into downstream tasks. In Computer Vision, GANs have been used to generate synthetic images, enhancing the diversity of the training data [38, 60, 76, 93, 98, 134]. Alongside GANs, neural style transfer and adversarial training are also employed to generate synthetic data. Neural style transfer manipulates the style of images while preserving their content, allowing the creation of diverse and visually varied training samples [98].

At the algorithmic level, adversarial training has gained traction as a robust method for mitigating bias. This technique involves designing models to disentangle sensitive attributes, such as gender, from task-relevant features, as demonstrated by Fleisig et al. [52]. Another prominent strategy is the incorporation of fairness constraints into optimization objectives, ensuring that models achieve equitable performance across demographic groups. Research by Zafar et al. [135] illustrates the efficacy of this approach in creating classifiers that respect demographic parity. Regularization methods, such as those proposed by Madras et al. [81], penalize models for learning biased representations, thereby balancing accuracy and fairness. Recent studies by, Bhattacharjee et al. [8] introduced MulT, an end-to-end Multitask Learning Transformer framework designed to simultaneously learn multiple high-level vision tasks, including depth estimation, semantic segmentation, reshading, surface normal estimation, 2D keypoint detection, and edge detection. Their approach leverages a shared attention mechanism to model dependencies across tasks, demonstrating the scalability and

effectiveness of multi-task learning in computer vision applications.

Evaluation-level interventions play a critical role in ensuring that mitigation efforts are effective. Fairness-aware metrics, such as demographic parity and equalized odds, provide quantitative measures of bias in model outputs [61]. These metrics are vital for assessing whether models treat individuals from different demographic groups equitably. Additionally, statistical measures like disparate impact and calibration error have been employed to capture nuanced biases in predictive models, as outlined by Mehrabi et al. [86]. Tools for interpretability and explainability, such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), enable practitioners to analyze model decisions and uncover biased patterns [110, 80]. These tools facilitate the identification of features or decisions that disproportionately affect certain groups, thereby guiding corrective measures. Recent work by Hooker et al. [62] emphasizes the importance of integrating interpretability tools into the development pipeline to foster transparency and accountability. Hooker's research highlights the interplay between model complexity and the efficacy of explanation techniques, suggesting that simpler models may provide more actionable insights into bias mitigation strategies. Moreover, Ghai [54] presents a human-centered approach to enhancing explainability and fairness in AI systems, emphasizing the importance of transparency and human control. This approach aligns with the growing demand for ethical AI systems that prioritize user trust and inclusivity.

2.2 Gender Bias in Natural Language Processing

In natural language processing (NLP), gender bias is evident in the way large-scale language models like GPT and BERT propagate stereotypes from their training data into downstream applications. Caliskan et al. [27] and Zhao et al. [141] revealed that these models embed gendered associations such as linking "he" with occupations like "doctor" and "she" with "nurse." Such biases also surface in translation tasks, where gender-neutral sentences

from languages like Turkish are often translated into gendered English equivalents, reflecting stereotypical roles. More recent studies, such as those by Sheng et al. [119], show that conversational agents powered by language models frequently generate biased responses, further entrenching stereotypes.

In an interesting study [26], the authors find that a significant majority of the most frequent words in the training data of the widely used embeddings are associated with men. For example, 77% of the top 1,000 most frequent words in the GloVe [97] embeddings are associated with men. This pattern holds true across different frequency ranges, with similar trends observed in fastText embeddings [20], albeit to a slightly lesser extent. In addition, their part-of-speech tagging analysis showed that male-associated words are more likely to be verbs, reflecting stereotypes of men as active and agentic. Female-associated words, on the other hand, are more likely to be adjectives and adverbs, suggesting a perception of women that requires additional description or explanation. Also, clustering analysis reveals that male-associated words often relate to domains such as big tech, engineering, sports, and violence. In contrast, female-associated words frequently pertain to appearance, sexual content, and kitchen-related terms. The unequal representation of both genders within the training data of neural embeddings can cause the model to develop biased representations of men and women.

One notable impact of pre-training on large pre-trained language models (PLMs) is how the training data influences gender biases. For example, the training corpus for ELMo [99], the One Billion Word Benchmark, contains a significant gender skew: male pronouns (e.g., "he" "his" and "him") occur three times more frequently than female pronouns (e.g., "she" "her") [139]. Specifically, the dataset shows approximately 5.3 million occurrences of male pronouns compared to 1.6 million occurrences of female pronouns. This imbalance not only reflects societal biases but also propagates these biases into the trained embeddings, leading to a higher likelihood of male-biased associations in downstream tasks. Moreover, male pronouns co-occur more frequently with occupation words, regardless of whether those

occupations are stereotypically male or female. For instance, in the training corpus for ELMo, male pronouns co-occur with occupation words 170,000 times, whereas female pronouns co-occur with occupation words only 36,000 times [139]. These statistics underscore the need for balanced training datasets and pre-training debiasing strategies to mitigate inherent gender biases and ensure fairer and more accurate embeddings.

More recently, [72] transitioned from studying pre-trained language models to investigating large language models and how they perpetuate gender stereotypes, particularly in the context of occupational roles. They found that LLMs are 3-6 times more likely to choose occupations that stereotypically align with a person's gender. These choices align more closely with societal perceptions than with official job statistics, indicating that LLMs amplify existing biases beyond what is reflected in reality. Additionally, the study revealed that the behaviour of LLMs correlates more closely with human judgments about gender stereotypes than with actual labour statistics. This suggests that the training data for these models reflect societal biases rather than objective realities.

Many de-biasing strategies have been proposed to address gender bias in neural embeddings, and large language models. The method presented in [139] involves augmenting the training corpus with gender-swapped variants of sentences i.e., creating a parallel corpus where gender-specific words are swapped, ensuring an equal representation of male and female entities in the training data. For example, every instance of 'he' is swapped with the opposite gender version i.e., 'she' and vice versa. This ensures that the model is exposed to a balanced representation of gendered entities during training. Data augmentation has been shown to significantly reduce bias in downstream tasks such as coreference resolution.

Debiasing neural embeddings through masking gender indicators is another method [101] aimed at reducing gender bias in text classification tasks by explicitly removing gender-specific terms from the training data. The first step involves identifying and listing gender-specific terms such as pronouns titles, and other gendered words that are likely to introduce bias into the embeddings. Once the gender-specific terms are identified, they are systemati-

cally removed or replaced in the training data. This process, known as 'scrubbing' ensures that these terms do not influence the embeddings. The objective is to prevent the model from learning gender associations that could lead to biased outcomes in downstream tasks. The embeddings are then trained on this modified corpus.

In [22], the authors propose a debiasing approach through regularizing the loss function for training the embeddings. In particular, they proposed a regularization loss term for the language model that minimizes the projection of encoder-trained embeddings onto an embedding subspace that encodes gender. This method aims to reduce the influence of gender bias in the learned embeddings. The regularization method is effective in reducing gender bias up to an optimal weight assigned to the loss term.

The debiasing approach in [142] is focused on training debiased word embeddings from scratch by modifying existing word vectors. The authors proposed Gender-Neutral Global Vectors by altering the loss function of the GloVe [97] model to concentrate most of the gender information in the last coordinate of each vector. This allows for the use of word representations that exclude the gender coordinate. They achieve this by using two groups of male/female seed words and encouraging words from different groups to differ in their last coordinate. Additionally, they ensure that the representation of gender-neutral words (excluding the last coordinate) is orthogonal to the gender direction.

2.3 Gender Bias in Information Retrieval

2.3.1 Source of Gender Bias in Information Retrieval Systems

Information retrieval systems consist of three primary components: (1) input query, (2) retrieval method, and (3) gold standard documents. Ideally, the objective of a search engine is to retrieve the gold standard documents using one or more retrieval methods given a specific search query. However, if any of these components harbor biases, the list of retrieved documents presented to users will reflect these biases. Consequently, users are exposed to

biased content, which can negatively influence their perceptions and judgments.

In this section, we explore the presence of gender biases within each of the aforementioned components and demonstrate the evidence of gender bias in each. Additionally, we provide a comprehensive analysis of how each component can serve as a source of gender bias, ultimately contaminating the fairness and accuracy of the information retrieval system.

Gender Bias in Input Query

The input query marks the initial point of interaction between the user and the information retrieval system. While users' search queries may appear gender-neutral, they can inherently carry social biases that affect the search engine's responses. This section will cover how these biases manifest and the implications of query reformulation on gender bias in search results.

Algorithmic Query Reformulation. Imagine a user entering the query 'top scientist'. On the surface, this query seems unbiased and straightforward. However, due to intrinsic societal biases, search engines might prioritize documents that highlight male scientists, assuming that scientists are predominantly male. This subtle bias can significantly impact the user's perception and the visibility of female scientists. When users submit queries, these can carry hidden social biases that influence the retrieval process. If the queries are socially problematic, they introduce biases into the retrieved documents. Therefore, understanding the biases in initial queries is crucial for developing fair and unbiased information retrieval systems.

Query reformulation methods, such as Pseudo-Relevance Feedback (PRF), can exacerbate these biases. The RM3 PRF method, for instance, enhances the original query by incorporating terms from the top-ranked documents retrieved by the initial query, assuming these documents are relevant. However, if the top-ranked documents are biased, the expanded query may reflect and amplify these biases. Consider the example of the query 'top scientists'. If the initial top-ranked documents are biased towards male scientists, the RM3 PRF method might expand the query to include terms like 'top scientists men male', leading

to a biased set of results. This process illustrates how reformulated queries can perpetuate gender stereotypes and underscores the need for analyzing and addressing these biases.

To investigate the extent to which pseudo-relevance feedback methods, such as RM3, introduce biased terms, the authors in [18] targeted a set of non-gendered news-related queries from different TREC corpora, including Robust04 [126], Gov2 [34], ClueWeb09 [28], and ClueWeb12 [35]. They used the *ARaB* metric [107] to examine the level of bias among the top 10 documents retrieved by BM25 and the PRF model. The authors found that the reformulated queries included gender-specific terms not present in the original queries. For example, the query 'Cult Lifestyles' was reformulated to include terms such as 'student Krishna Kim Chilton lifestyle **she mother** cult **her** car pension,' or 'american muslim mosques schools' changed to 'mosque muslim american wahhabi my saudi america religion Islam school **he**' by the PRF model. These additions resulted in increased bias in the retrieved documents, demonstrating how PRF methods can inadvertently introduce and amplify gender bias in search results.

The findings highlight the importance of critically evaluating query reformulation methods to ensure they do not perpetuate existing biases. Researchers and developers must consider the potential for these methods to introduce bias and work towards creating fairer, more balanced information retrieval systems. This includes exploring alternative approaches to query reformulation that mitigate bias and developing metrics to measure and address bias in search results effectively.

User-Driven Query Reformulation. In an observational study [102], a large-scale search log data from Bing explored how users reformulate their queries to include gender-specific terms, a process called gender-specializing query reformulations (GSQR). The study identified approximately 4.7 million pairs of consecutive queries where the second query was a GSQR of the first. For instance, consider a user initially searching for 'NCAA scores.' If their interest lies specifically in women's basketball, they might reformulate the query to 'NCAA women's scores.' This simple modification demonstrates how users can introduce

gender-specific terms to refine their search results. The study aimed to understand the contexts in which users reformulate their queries to include gender-specific terms and the impacts of these reformulations on search results.

The authors defined a query reformulation as specializing if the reformulated query contained all terms from the original query in the same order and included additional contiguous terms related to gender. They calculated the overall frequency of GSQRs and categorized the original queries by topic, revealing higher rates of GSQRs in categories such as shopping and fashion. This categorization provided insights into the contexts where users are more likely to introduce gender-specific terms.

The study also explored the timing and method of query reformulation. Time differences between the original and reformulated queries were analyzed to infer user behavior, showing a median time of 19 seconds between queries, with men-related reformulations occurring slightly more quickly. Additionally, the study examined how users entered their reformulated queries. It was found that most gender-specific query reformulations (GSQRs) were made either by editing the original query directly in the search bar at the top of the search results page (SERP) or by selecting one of the recommended queries provided by the search engine.

Furthermore, the study investigated the genderedness of original queries using average GloVe [97] embeddings of terms. This analysis revealed instances where GSQRs corrected the under-representation of a gender or reinforced existing gender representations. For example, the query 'ADHD symptoms' might be reformulated to 'ADHD symptoms for women' to obtain gender-specific information, while a query like 'NCAA basketball score' might be reformulated to 'NCAA men's basketball score' to emphasize men's results.

These findings highlight the active role users play in shaping search results through their query reformulations. Understanding user behavior in this context is crucial for designing search systems that accommodate user needs while mitigating the introduction of bias. It also underscores the importance of providing users with tools and options to refine their searches in a way that promotes fairness and diversity in the retrieved results.

Gender Bias in Retrieval Methods

Retrieval methods play a pivotal role in retrieving relevant documents in response to user input queries. However, when these methods lack awareness of bias, they may inadvertently retrieve biased document sets, especially in response to sensitive queries. The algorithms and techniques employed by retrieval methods to match the input query with relevant documents can be a source of gender biases. These methods are often based on Pre-trained Language Models (PLMs) trained on large datasets that contain human-generated content such as online forums, books, articles, web pages, etc. If the training data itself contains gender biases, the retrieval algorithms capture those biases from data as a form of association and likely reproduce and even amplify these biases. For example, if a search engine's algorithm has been trained on data that overrepresents male achievements in science and underrepresents female achievements, it will preferentially retrieve documents about male scientists, even when the query is a gender-neutral one.

Gender Bias in Embedding Models. In the field of Natural Language Processing (NLP), studies have investigated the presence of gender biases in embedding representations of word-embedding models [21, 27] such as Word2Vec [87] and pre-trained language models [84, 139] like ELMo [83] and BERT [42]. For instance, the authors in [21] provided significant insights into gender biases within Word2Vec and GloVe word embeddings. They introduced a geometric framework to identify gender in the embedding space and evaluated whether the embeddings of occupations exhibit stereotypical gender biases. Additionally, they examined if the embeddings generate analogies that humans perceive as reflecting gender stereotypes. To investigate these, they created a gender subspace by considering 10 pairs of female and male words such as ('she', 'he') and ('mother', and 'father') and subtracting the embedding representation of each pair's word embeddings to form matrics of ten vector embeddings and finally applied singular value decomposition to obtain *she-he* gender subspace.

Followed by that, this gender subspace was projected into male-stereotypic and female-stereotypic occupations and it has been shown that it is strongly correlated with the anno-

tations of ten crowd-workers who were asked to annotate the occupations into male, female, or neutral. Examples of extreme male and female occupations are captain and nurse, respectively. To investigate if analogies also reflect stereotypes, the authors modified the standard analogy task to generate pairs of words, allowing the systematic creation of analogies based on seed words like 'he' and 'she.' Using a scoring metric based on cosine similarity and a semantic coherence threshold, they identified top analogous pairs. Finally, crowd workers evaluated these pairs to determine if they made sense and reflected gender stereotypes. The number of workers identifying a stereotype in each analogy quantified the degree of bias. The results of the experiment showed that 72 out of 150 analogies were considered gender-appropriate, while 29 exhibited gender stereotypes.

In addition to static word embeddings, contextualized embeddings also demonstrate stereotypical gender biases, as evidenced by the authors in [139]. In their study, 400 sentences were selected from the OntoNotes 5.0 dataset [131], each containing at least one gendered word (e.g., 'he' or 'she'). Subsequently, they created a gender-swapped version for each sampled sentence and analyzed the disparities in ELMo embeddings among the occupation words in these paired sentences. Principal component analysis (PCA) was then applied to these differences.

The results revealed the existence of two principal components related to gender in ELMo embeddings: one representing contextual gender information and the other representing gender inherent in the occupation word itself. The first principal component segregates male and female contexts, while the second component clusters male-associated and female-associated occupation words. This analysis underscores the nuanced depiction of gender in ELMo embeddings, encompassing both contextual and lexical gender information.

The Impact of Neural Embeddings on Retrieval Methods. With the emergence of neural embeddings, retrieval methods shifted from term-frequency-based methods to neural-based retrieval methods that leverage different variations of static word embeddings and contextualized word embeddings for the task of retrieving relevant documents given a search

query. For instance, retrieval methods such as KNRM [133], Conv-KNRM [41], DRMM [59], DUET [88], and MatchPyramid [94] use Word2Vec or GloVe word embeddings for matching query and documents representation.

Despite the improved performance of neural-based retrieval methods, they often rely on static or contextualized embedding models that have been found to contain and amplify gender biases [21, 139]. Moreover, these embedding representations, already biased, are fine-tuned based on gold standard collections to learn query-document semantic mapping. During the fine-tuning process, there is a risk of further reinforcing gender biases. For instance, if a relevant document pair for a query is biased towards a specific gender attribute, the model may capture and reflect this biased association during the ranking process.

To investigate how gender biases within static and contextualized embeddings manifest at the ranking stage by retrieval methods, several studies have measured the level of gender biases among the ranked lists of documents for queries using different retrieval methods [51, 107, 105].The authors in [51] employed term-frequency-based retrieval methods such as TF-IDF and BM25, as well as neural-based retrieval methods including DRMM and MatchPyramid, to compare the level of gender bias among the retrieved lists of documents for a subset of queries. Based on their experimental results, traditional lexical models such as TF-IDF and BM25 exhibited low gender stereotypes, whereas semantic models based on biased word embeddings tended to reinforce gender stereotypes, even with IDF-inspired weighting schemes.

The authors in [107] released a set of gender-neutral queries consisting of 1,765 queries annotated by annotators as being non-gendered, as outlined in Section 3.3. Using this gender-neutral set of queries, BM25, static and contextualized embedding retrieval methods such as BERT, KNRM, and MatchPyramid are employed to rank the top 1000 documents retrieved by BM25. They then measured the level of bias within the ranked lists of documents from each of these retrieval methods using term-frequency-based and boolean-based variations of their proposed metric, namely *ARaB*. The results of the comparison, in terms of *ARaB* met-

rics, demonstrated that all retrieval methods, especially neural models and BERT, showed a bias towards male concepts, with neural models consistently increasing retrieval gender bias compared to BM25. Furthermore, the use of pre-trained word embeddings in neural ranking models tends to increase gender bias.

In another study [105], the authors investigated the level of fairness within the retrieved list of documents ranked by different retrieval methods across $MSMARCO_{FAIR}$ and $TRECDL19_{FAIR}$ queries introduced in Section 3.2. They proposed NFaiRR metric to measure fairness across the top-k ranked list of documents for the two sets of fair queries by different retrieval methods. They evaluated the fairness metric across ranked lists of documents by retrieval methods such as BM25, KNRM, MatchPyramid, BERT-Tiny, and BERT-Mini. Based on the results, ranker-agnostic document sets reveal that in the $MSMARCO_{FAIR}$ collection, the NFaiRR of SetTop200 (top-200 documents retrieved by BM25) is slightly lower than SetAll (top-1000 documents retrieved by BM25), indicating higher gender bias in the top retrieved documents compared to the entire collection. This suggests that $MSMARCO_{FAIR}$ queries tend to pull documents from biased subspaces. Conversely, in the $TRECDL19_{FAIR}$ collection, SetTop200 is more fair than SetAll, approaching ideal fairness, suggesting that $TRECDL19_{FAIR}$ queries lead to balanced gender representation. For ranking models in $MSMARCO_{FAIR}$, classical retrieval models such as BM25 exhibit the lowest fairness, while neural models show significantly higher fairness scores, with BERT rankers achieving the best results.

In conclusion, the investigation into language models and retrieval methods has highlighted significant gender biases that can affect the relevance and fairness of retrieved document sets. These biases originate from the training data and are further amplified by the algorithms and pre-trained language models used in retrieval systems. Studies have shown that neural-based retrieval methods, despite their improved performance, tend to increase gender bias compared to traditional models like BM25. This is evident in models that use both static word embeddings and contextualized embeddings, which often capture and per-

petuate gender stereotypes. Consequently, there is a pressing need to apply de-biasing methods to retrieval algorithms. Implementing such methods can mitigate these biases, ensuring that retrieval systems provide fairer and more balanced results, particularly in response to gender-neutral queries.

Gender Bias in Gold Standard Datasets

One of the main, if not the main, sources of gender biases in both embedding models and retrieval methods used for ranking relevant documents in search queries is the use of biased training datasets. Numerous NLP research studies have statistically analyzed the ratio of male and female-related terms within the datasets used for training neural models. These studies have shown that such training datasets contain considerable gender bias, leading models to learn and reproduce associations between occupations or other entities with gender. For example, in [44], the authors analyzed six dialogue datasets for gender bias and found a significant bias towards males. Specifically, the LIGHT text adventure world dataset [125] was identified as the most biased, with male bias reaching 73%. This high level of bias is attributed to the dataset's multiple potential sources of bias, its crowdsourced nature, and its medieval, fantasy setting, which may reflect the gender biases of the crowdworkers.

In another study [139], the authors targeted the One Billion Word Benchmark corpus [30] and calculated the occurrences of male pronouns (he, his, him) and female pronouns (she, her) in the corpus, as well as the co-occurrence of these pronouns with occupation words. The set of occupation words and their gender assignments were based on the WinoBias corpus [141]. The findings revealed a significant gender skew in the Billion Word corpus as it could be observed that male pronouns tend to appear three times more frequently than female pronouns, and male pronouns were more commonly associated with occupation words.

Investigating gender biases in gold-standard training datasets for information retrieval methods is imperative. Quantifying these biases is crucial because neural-based retrieval methods are often trained on such data and biased training data can transfer biases into

the algorithmic and representational aspects of retrieval methods. This bias can ultimately affect the ranking process by causing models to associate gender with query needs for sensitive queries, leading to biased documents being ranked higher and increasing their exposure to users.

Gold standard datasets for training retrieval methods are typically obtained from crowdworkers' annotations, where workers are tasked with identifying relevant documents to queries. Gender bias within these datasets can arise from two primary sources:

1. **Perceived bias in human judgments:** Crowdworkers assigned to annotate relevant documents may inadvertently inject their own biases into the dataset. These biases can be influenced by various factors such as cultural background, personal beliefs, or societal stereotypes. For instance, a crowdworker's cultural background might influence their interpretation of what constitutes relevance, potentially leading to the inclusion or exclusion of certain documents based on gender-related assumptions or stereotypes.

2. **Inherent biases in the content:** The material available for annotation might inherently contain biases due to the nature of its source or the domain it represents. This could encompass biases in language usage, representation of specific demographics, or the framing of topics. For instance, documents sourced from certain domains or publications might predominantly focus on specific gender-related issues, thus skewing the dataset towards particular gender perspectives.

In this section, we thoroughly investigate each of these potential sources of bias to understand their implications within the context of gender biases in information retrieval datasets.

Impact of Perceived Bias on Human Judgements. Information retrieval models are typically trained and evaluated using a collection of relevance judgments determined by human assessors. If these documents display biases toward a particular gender, such biases have the potential to manifest in the retrieved list of documents presented to users. It is plausible that biases ingrained within individuals' mental frameworks may influence their

decisions regarding the relevance or irrelevance of information [49, 50, 120].

In this section, we explore the critical question of whether these perceived biases impact human decision-making during the document judgment process, based on the research work in [74]. The experimental setup of this study employs the Grep-BiasIR dataset, introduced in Section 3.3.4, which includes bias-sensitive queries and documents with varying gender indications (male, female, and neutral). The experiments are conducted in two distinct settings: gender-specific (where the gender of the participants is known) and gender-agnostic (where the participants' gender is unknown).

The queries in this study span six categories: Appearance, Career, Domestic Work, Child Care, Cognitive Capabilities, and Physical Capabilities. Each query is paired with relevant and non-relevant documents presented in both male and female versions. Participants from the Amazon Mechanical Turk platform were tasked with rating the relevance of these query-document pairs on a scale from non-relevant to perfectly relevant. Relevance judgments were collected from 50 participants for the gender-agnostic setting and from 10 male and 10 female participants for the gender-specific setting.

The study aimed to investigate three primary hypotheses. First, it examined whether relevant documents aligned with expected gender stereotypes received higher relevance scores (H1). Second, it assessed whether non-relevant documents reflecting gender stereotypes also influenced relevance scores (H2). Lastly, the research explored whether participants' gender influenced their judgments of gender-biased content (H3). The design of the study allowed the researchers to explore these hypotheses comprehensively within the controlled experimental setup.

In the gender-agnostic experiments, findings revealed that participants generally rated documents aligning with expected gender stereotypes higher in relevance. For example, in the Domestic Work category, relevant documents with female-indicating content scored higher than those with male content. This trend aligns with societal stereotypes where women are often associated with domestic responsibilities. However, statistical significance was limited,

indicating a subtle but observable influence of biases. For non-relevant documents, the results were mixed, with some categories showing stereotype-disconfirming content being rated as more relevant, potentially due to the surprising nature of such information. For instance, in the Appearance category, male-indicating non-relevant documents were rated higher, possibly reflecting a perceived novelty or unexpectedness in this context.

The gender-specific experiments aimed to assess whether participants' gender impacted their judgments of gender-biased documents. Results showed no significant interaction between participant gender and their relevance ratings for either relevant or non-relevant documents. For instance, male and female participants similarly rated queries such as "how to build muscles" and "what is considered plus size" with no significant deviations linked to their gender. This suggests that gender bias in perceived relevance judgments is not directly influenced by the annotator's gender under the study's experimental conditions.

The study's hypotheses regarding stereotype confirmation (H1), stereotype-confirming effects for non-relevant documents (H2), and the impact of participant gender (H3) were only partially supported. While relevance scores often aligned with stereotypes, the effect sizes were small, and statistical significance was generally not observed. An exception was observed in the Appearance category for non-relevant documents, where stereotype-disconfirming male content scored higher than female content, potentially influenced by how unexpected or surprising the content seemed.

The study is not without its limitations, which constrain the generalizability and interpretability of its findings. First, the sample size, particularly for the gender-specific experiments, was relatively small, reducing the statistical power to detect significant effects. Additionally, the study relied on binary gender classifications (male and female), which limits the scope of the findings and excludes insights from individuals who do not identify within these categories. Another limitation arises from the controlled nature of the experimental setup, where participants evaluated isolated query-document pairs without additional real-world contextual factors, such as document ranking, source credibility, or temporal relevance.

Bias in the Content of Gold Standard Datasets. The presence of gender biases within the content of relevance judgment collections is examined in this section. To achieve this, a three-staged methodological approach, as proposed in [19], is employed. The first stage involves identifying and labeling queries based on their gender. To facilitate this, the authors in [107], introduced a gender-annotated dataset, which included queries labeled as female, male, or neutral. Various models were trained to classify query gender, encompassing both dynamic embeddings (such as BERT, DistilBERT, RoBERTa, and XLNet) and static embeddings (including fastText and Word2Vec). The performance of these classifiers was rigorously evaluated using a 5-fold cross-validation strategy. The uncased fine-tuned BERT model demonstrated superior performance, with high accuracy and F1 scores across all gender classes. Subsequently, this model was employed to label the entire MS MARCO development query set for gender, resulting in a comprehensive dataset.

In the second stage, the authors measured various psychological characteristics of the relevance judgment documents associated with gendered queries. Using the Linguistic Inquiry and Word Count (LIWC) toolkit [96], they quantified affective processes, cognitive processes, drives, and personal concerns within gold standard documents associated with development set labeled queries from each of the male, female, and neutral categories. This stage was crucial for determining whether the psychological expressions within the documents aligned with known findings from psychological literature or exhibited stereotypical biases. This analysis helped to uncover implicit biases embedded in the content of the documents, providing a detailed understanding of how these biases manifest in relevance judgments.

The third stage involved reporting findings on gender stereotypical biases in the gold standard relevance judgments. The results demonstrated that documents related to female queries exhibited higher degrees of negative emotions such as anxiety and sadness, while male query-associated documents showed more anger. In terms of cognitive processes, documents associated with female queries demonstrated higher cognitive complexity. For drives, male query-related documents expressed more affiliation, achievement, and power, whereas

female query-related documents emphasized reward and risk avoidance. Regarding personal concerns, documents linked to male queries had a higher focus on work and leisure. This stage revealed that the relevance judgment documents indeed reflected stereotypical gender biases, consistent with some psychological research findings while highlighting unexpected biases in the datasets.

These findings underscore the necessity of exploring de-biasing methods to mitigate prevalent gender biases in gold-standard datasets used for training retrieval methods.

2.3.2 Eliminating Gender Bias in Information Retrieval

Recent research has focused on identifying and reducing gender biases in neural rankers. Rekabsaz et al. [107] were among the first to identify this issue in such systems. Their paper investigates the presence and extent of gender bias in neural ranking models. The authors develop a framework to measure gender bias, introducing two metrics to quantify gender bias in the ranked list of retrieved documents. Their work offers a dataset of gender-neutral queries and employs it to evaluate various models, including the baseline BM25 method and several neural ranking models. Their findings show that all models exhibit a male bias, but neural models, especially those using contextualized embeddings like BERT, significantly amplify this bias. The study also reveals that transfer learning with pre-trained embeddings tends to increase gender bias in neural rankers. The work by Rekabsaz can be considered a pioneering work that highlights the need to reduce gender biases in neural rankers while maintaining their retrieval effectiveness.

Subsequently, Bigdeli et al. [19] investigate the presence of gender biases in gold standard relevance judgment datasets used for training and evaluating neural rankers. Since these relevance judgment datasets greatly influence how neural rankers learn the concept of relevance, the authors focused on quantifying and analyzing gender biases in relevance judgments. The authors use a fine-tuned BERT model to label a large collection of queries within the MS MARCO dataset [89], which were then used to assess the associated docu-

ments for their psychological characteristics using the Linguistic Inquiry and Word Count (LIWC) toolkit [95]. Their findings showed that stereotypical biases are common in relevance judgment collections, particularly with regards to affective and cognitive processes, as well as personal concerns and drives. Bigdeli et al advocate for the need for unbiased gold standard relevance judgement datasets that can avoid training biased neural rankers. Based on the findings from Rekabsaz et al. [106] and Bigdeli et al. [14] that showed neural rankers are susceptible to intensifying gender biases, follow up work has been focused on developing methods that can reduce or eliminate such biases. The authors in [106] attempt to eliminate gender information from the intermediate vector representation produced by BERT. Their proposed architecture comprises a BERT encoder and two classifier heads. One of the classifiers acts as an adversarial network, designed to discourage BERT from encoding gender information in its internal representations. This adversarial network is trained to predict gender, while BERT's encoder is trained to minimize this prediction accuracy by making its internal representations less informative. Using the adversarial framework, the network aims to maximize relevance prediction while minimizing the prediction of gender labels. This approach allows the encoder to gradually exclude gender information from the intermediate vector representation, preventing the gender classifier head from being able to predict gender from the vector representation.

Another relevant work [18] explores the commonly held belief that reducing bias in ranker systems comes at the cost of utility (retrieval effectiveness). The authors propose a bias-aware pseudo-relevance feedback framework that aims to revise input queries to maintain or improve retrieval utility while significantly reducing bias. The paper demonstrates that it is possible to reduce bias without compromising retrieval effectiveness. This work challenges the traditional view of bias and utility as competing aspects and suggests that they can be addressed concurrently. Although the method is effective in reducing gender biases while maintaining the performance, the method is limited to non-neural models such as BM25.

The work by Zerveas et al. [138] introduced a novel approach to mitigate bias in neural

rankers through an end-to-end differentiable, transformer-based framework called COntextual Document Embedding Reranking (CODER), which optimizes document relevance scores while simultaneously imposing neutrality regularization. CODER uses a transformer query encoder that scores a set of candidate documents collectively rather than in isolation, achieving contextual ranking. For bias mitigation, a regularization loss penalizes high-scoring documents that deviate from neutrality with respect to gender. The neutral ranking objective is achieved by comparing the distribution of scores against ideal, unbiased ranking scores. The authors show that CODER provides a smoother and more predictable bias mitigation process.

Different from other work that focus on adjusting the training model or the model architecture, the recent work by Bigdeli et. al. [14] addresses the problem of gender bias in neural retrieval models by proposing a simple and effective training data sampling strategy. The authors suggest incorporating the degree of gender bias when sampling documents for training neural rankers, allowing these models to maintain retrieval effectiveness while reducing gender biases. This strategy involves a systematic negative sampling approach that exposes neural rankers to biased documents, teaching them to avoid gender biases without architectural changes to neural rankers. This approach is notable for its simplicity and efficacy, offering a practical solution for reducing gender biases while being applicable to a range of neural rankers.

These existing methods for mitigating gender bias in information retrieval systems can be broadly categorized into three main classes: *Data-Driven Debiasing*, *Loss Function Regularization*, and *Adversarial Training*. Each category represents a unique approach to addressing biases, focusing on different aspects of the model development and training process, and collectively, they form a comprehensive framework for tackling this pressing issue.

Data-Driven Debiasing strategies focus on addressing biases at their source: the training data. The method proposed by Bigdeli et al. [14] exemplifies this approach. By incorporating a systematic negative sampling strategy, this method ensures that training data exposes

neural rankers to biased documents, teaching them to minimize reliance on gender-biased patterns without requiring architectural modifications. The strength of data-driven debiasing lies in its simplicity and versatility, as it can be applied to a wide range of neural ranking models. By carefully curating and sampling data, these methods can preemptively reduce the propagation of biases in downstream tasks, laying a strong foundation for equitable information retrieval systems.

Loss Function Regularization methods, on the other hand, directly incorporate bias mitigation objectives into the model's training process. These methods can be further divided into two subcategories: the introduction of interpolated losses and the application of explicit regularization terms. Zerveas et al. [138] present an exemplary interpolated loss approach through their CODER framework. By integrating a neutrality regularization loss alongside the primary ranking objective, CODER ensures that high-scoring documents align with unbiased ranking distributions. This balanced optimization enables the model to achieve contextual relevance while maintaining neutrality.

Adversarial Training represents a more nuanced approach, wherein the model is explicitly trained to disentangle sensitive attributes, such as gender, from its internal representations. Rekabsaz et al. [105] introduced one of the pioneering adversarial methods for bias mitigation in neural rankers. Their approach employs a dual-objective training framework, where a gender classifier is adversarially trained against the ranker's encoder. This process encourages the encoder to produce representations that are uninformative about gender, effectively neutralizing gender biases in the intermediate vector representation. While adversarial training offers a powerful mechanism for achieving fairness, it often requires careful tuning to balance the competing objectives of relevance and neutrality.

These three categories collectively illustrate the diversity of strategies available for addressing gender bias in neural rankers. While data-driven approaches target biases at the data level, loss function regularization and adversarial training focus on embedding fairness into the model's learning process. Each method has its strengths and trade-offs, but together,

they provide a holistic framework for mitigating bias. By advancing these approaches, researchers can ensure that information retrieval systems operate equitably, fostering trust and inclusivity in their application across critical domains.

Chapter 3

Problem Definition

3.1 Gender Fairness in Ranking

Fairness in ranking can be defined in various ways, depending on the perspective from which the problem is analyzed [136, 43, 9]. In this manuscript, we focus on fairness in terms of gender equality. There are two broad definitions for fairness, namely *Group Fairness*, and *Individual Fairness*. In this section, we define gender fairness in information retrieval as a subset of *group* and *individual* fairness.

Definition 1. ***Group Fairness****: Group fairness aims to ensure that groups of individuals with different protected sensitive attributes receive comparable treatments statistically.*

One of the most common definitions of group fairness is **statistical parity** [48]. Statistical parity requires the prediction y to be independent of the sensitive attribute s, denoted as $y \perp s$. This can be mathematically represented for binary classification and binary attributes as follows:

$$P(y = 1 \mid s = 0) = P(y = 1 \mid s = 1) \tag{3.1}$$

To measure statistical parity, we can use the difference in probabilities:

$$\Delta_{SP} = |P(y = 1 \mid s = 0) - P(y = 1 \mid s = 1)| \tag{3.2}$$

A lower value of Δ_{SP} indicates a fairer classifier. Statistical parity can be extended to multi-class and multi-category sensitive attributes by ensuring that the prediction y is independent of s.

One approach to conceptualize group fairness in IR argues that fairness may be achieved when the probability of a document being retrieved for a query is independent of its gender attribute, particularly for gender-neutral queries. This interpretation is rooted in the idea that documents with identical content but differing gender attributes should have equal chances of being retrieved, as their relevance to the query remains unaffected by gender. For instance, given a gender-neutral query q_n, and two documents d_m and d_f with identical content but associated with male and female attributes respectively, the probability of retrieving either document should not differ significantly.

This perspective aligns with the principle of *statistical parity*, which can be mathematically expressed to measure the degree of fairness in the retrieval process. Statistical parity ensures that the likelihood of retrieval is balanced across gender attributes, reducing disparities introduced by implicit or explicit biases:

$$\Delta_{SP} = |P(d_g, q_n) \mid g = m) - P(d_g, q_n) \mid g = f)| \tag{3.3}$$

where, $P(d_g, q_n)$ is the probability of the document d_g being the top-retrieved document for the gender-neutral query q_n. Therefore, group fairness in information retrieval is addressed if and only if:

$$\Delta_{SP} \to 0 \tag{3.4}$$

Definition 2. ***Individual Fairness****: Individual fairness focuses on ensuring that similar individuals receive similar algorithmic outcomes.*

Individual fairness for gender focuses on ensuring that algorithmic outcomes treat similar entities equitably, regardless of their gender-related attributes. In information retrieval, this means that documents or queries with similar intrinsic relevance or difficulty should receive similar treatment by the system, independent of their association with a particular gender. This approach addresses disparities that may arise from systemic biases, ensuring that gender does not influence the outcomes for otherwise comparable items. By grounding fairness in unbiased metrics of similarity or merit, individual fairness for gender promotes equitable treatment at the level of individual entities while avoiding overgeneralization based on group characteristics.

An example of individual fairness is the amortized attention framework [9] that ensures that items with similar relevance receive equitable exposure over a series of rankings, making it a valuable approach for addressing gender bias in information retrieval. This framework can help mitigate imbalances in attention, such as clicks or visibility, between gender-associated documents or content. For example, in a scenario where equally relevant documents feature female-associated content (e.g., biographies of women) and male-associated content (e.g., biographies of men), systemic biases in ranking algorithms might lead to disproportionately higher exposure for male-associated content. By applying the amortized attention framework, the cumulative exposure of female- and male-associated documents can be aligned with their relevance across multiple rankings, ensuring balanced representation and preventing the consistent underrepresentation of any gender.

Another example of individual fairness frameworks is the Expected Exposure Model by **(author?)** [43], which evaluates fairness in rankings by examining how attention (exposure) is distributed across items of the same relevance grade, making it particularly effective for addressing gender bias in information retrieval. The model ensures that documents or items associated with different characteristics (in this case gender) but of equal relevance receive comparable exposure across stochastic rankings. For instance, in response to a query like 'top scientists in history,' documents about male scientists may consistently rank higher than

those about female scientists due to systemic biases, resulting in unequal exposure. The expected exposure model addresses this by aligning exposure with relevance, ensuring that female-associated content receives exposure comparable to equally relevant male-associated content, thereby fostering fairness in gender representation.

Given the definitions of group fairness, and individual fairness in information retrieval, a fair information retrieval system satisfies both group and individual fairness to ensure that the model has fair behaviour towards different demographic groups, and also the individuals with different genders.

With the fairness definitions in mind, we are going to take a deeper look into the neural rankers, and the existing gender biases in them. Rekabsaz et al. are among the first researchers who discovered that neural rankers not only exhibit gender biases but also reinforce the existing biases [107]. During their extensive experiments, they revealed that neural rankers exhibit gender biases in the sense that for a gender-neutral query, most of the retrieved documents exhibit inclination toward males. This contradicts group fairness defined earlier.

On the other hand, the authors in [117] revealed another category of gender biases in information retrieval systems. Their research shows that the neural rankers perform better when applied to male queries, as compared to the female queries. This contradicts individual fairness that states every individual should be treated the same by an information retrieval system, regardless of their gender attribute. This research led to the following research question in terms of the origin of these gender biases as well as different datasets, and metrics for measuring gender bias in information retrieval systems.

3.2 Benchmarking Gender Fairness

Gender orientation of queries and documents can be defined in multiple ways, depending on the perspective from which gender is analyzed. One key perspective is the lexical definition

of gender, where gender is inferred based on the explicit or implicit presence of gendered words in the text. Explicit indicators include gendered pronouns ('he,' 'she'), gendered nouns ('man,' 'woman'), or professions explicitly marked by gender ('policeman,' 'policewoman'). For example, a query like 'female CEO leadership strategies' or a document discussing 'the contributions of male nurses' clearly signals gender through direct lexical cues. However, lexical gender orientation can also be implicit, where certain words or topics inherently suggest a specific gender without using explicitly gendered terms. For instance, a query about 'pregnancy nutrition tips' implicitly points towards a female subject, as pregnancy is strongly associated with women. Beyond lexical cues, subject-based gender orientation refers to the gender of the entities or individuals discussed in a document, such as the main character in a novel, the subject of a biography, or the individuals featured in a news article. A biography titled 'The Life of Marie Curie' would be considered female-gendered due to its subject. Another dimension is the author's gender, where the gender is determined by the identity of the author or creator of the text. For example, a memoir written by a female author might be considered female-gendered, regardless of its content. These definitions highlight that gender orientation in text can emerge from linguistic choices, subject focus, or authorial perspective, and these facets are not mutually exclusive. In this manuscript, given the current predominant focus on the literature (albeit not the sole focus), we review the lexical definition of gender, emphasizing how the textual content of queries and documents — both explicitly and implicitly — reflects gender orientation. As such and in this manuscript, a query or document is considered gendered if it includes explicit or implicit gendered words. In the absence of such indicators, it is classified as neutral. This approach provides a measurable foundation for analyzing gender bias in information retrieval systems, while acknowledging that other dimensions of gender representation remain valuable for broader contextual analyses.

This approach to defining gender orientation is limited by its reliance on textual cues, which may overlook the complexities of gender representation and its intersectionality with

other social factors. Implicit gender orientation, for instance, is context-dependent and may vary across cultures or individual interpretations, making it challenging to generalize. Additionally, focusing on lexical definitions risks oversimplifying gender as a binary concept, excluding non-binary and fluid identities. We acknowledge that while this framework provides a measurable starting point, it does not capture the full scope of gender representation, which may extend beyond textual content to societal, historical, and cultural contexts.

3.3 IR Datasets

Several datasets have been proposed to identify and measure gender biases in NLP and information retrieval systems [67, 111, 114]. These datasets incorporate queries labeled with gender information, which can serve multiple purposes, such as analyzing search engine behavior, studying gender bias in human judgment, and analyzing human query generation. In this section, we will provide an in-depth explanation of these datasets, including their samples, statistical characteristics, and limitations. An overview of the datasets is included in Table 3.1.

3.3.1 Gendered Queries

This dataset is proposed by Rekabsaz et. al. [107]. The creation of the gender-annotated queries dataset involved several key steps aimed at ensuring the inclusion of queries that do not contain any gender-specific elements. This process was essential to accurately measure the gender bias present in the retrieval models. Here is a detailed explanation of how the dataset was created:

1. **Query Selection**: The queries were selected from the test set of the MS MARCO Passage Retrieval collection [90], a dataset comprising 8,841,822 passages and a large set of informational question-style queries from Bing's search logs. The initial selection focused on queries whose ranked list of documents displayed the highest inclinations towards gender,

determined by the retrieval results of seven ranking models.

For all the Information Retrieval (IR) models, the retrieval gender bias of each test set query was calculated using the Term Frequency (TF) gender magnitude measure and the Rank Bias (RaB) approach at a cutoff of 10 [108]. This process generated two separate lists of queries for each of the seven IR models studied: one list for queries biased toward females and another for queries biased toward males. Consequently, this resulted in a total of 14 lists of sorted queries. A pooling method introduced in [69] was applied to these sorted lists with a cutoff of 500, leading to a total of 3,924 unique queries. This method ensures a comprehensive selection of queries, capturing various degrees of gender bias as perceived by different models.

2. **Human Annotation**: The next step involved human annotation to categorize the queries accurately. Three Amazon Mechanical Turk workers were tasked with classifying each query into one of four categories:

- *Non-gendered*: Queries that do not refer to any specific gender.
- *Female*: Queries containing words or phrases related to female concepts (e.g., queen, pregnant).
- *Male*: Queries containing words or phrases related to male concepts (e.g., king, father).
- *Other or Multiple Genders*: Queries that refer to other genders or multiple genders (e.g., transgender, references to both male and female).

The detailed descriptions and guidelines for these categories were provided to the annotators to ensure consistency and accuracy. Based on the annotations, each query was assigned to a category using the majority vote of the annotators. Queries that did not reach an unambiguous majority decision (i.e., each annotator chose a different category) were removed from the dataset. This step was crucial to maintain the reliability of the dataset. The details of the dataset are included in Table 3.1.

3.3.2 MSMARCOFair: Gender-neutral Queries

The process of creating this dataset involves several detailed steps aimed at identifying and annotating fairness-sensitive queries related to gender equality. **(author?)** [105] began with an initial selection of queries from two prominent datasets: the TREC Deep Learning Track 2019 Passage Retrieval (TRECDL19) [37] and the development set of the MSMARCO Passage Re-ranking collection[90]. Specifically, they selected 1,765 non-gendered queries from the MSMARCO collection, which had been previously annotated by Amazon Mechanical Turk workers.

Next, the researchers employed three Amazon Mechanical Turk workers, all native English speakers, to annotate the queries from TRECDL19 in a similar manner. This crowdsourced annotation ensured consistency across both datasets. Following this, a meta-annotation process was carried out to verify the initial annotations and identify queries where the presence of gender bias in retrieval results would be socially problematic.

During the meta-annotation process, the researchers evaluated each query on two criteria: whether it was non-gendered and whether gender bias in its retrieval results would be socially problematic. Socially problematic queries were identified based on their potential to reinforce existing gender norms and promote gender inequality. The researchers focused on domains such as education, career, health, violence, exploitation, social inequality, and politics. For example, a query like 'how important is a governor?' was marked as fairness-sensitive because bias in this context could reinforce career stereotypes. Another query, 'When do babies start eating whole foods?' was identified as problematic due to its potential to reinforce the stereotype of 'women as caretakers', thereby impacting career choices and perpetuating gender norms.

The final step involved compiling the datasets, ensuring only those queries agreed upon by both meta-annotators were included. This resulted in the MSMARCOFair dataset containing 215 queries and the TRECDL19Fair dataset including 30 queries. These datasets were designed to serve as benchmarks for studying fairness in retrieval results, enabling research

Table 3.1: Overview of the datasets for gender bias in information retrieval.

Query set	queries	neutral	male	female	other	human annotator
Gendered Queries	3,750	1,765	1,202	742	41	
$MSMARCOFair$	215	215	-	-	-	
$TRECDL19_{Fair}$	30	30	-	-	-	
BERT Gendered Queries	51,827	48,200	2,222	1,405	-	
Grep-BiasIR	118	-	-	-	-	

on fairness alongside utility in information retrieval models.

3.3.3 BERT-annotated Gendered Queries

To begin, the authors in [10] employed a publicly available gender-annotated dataset provided by Rekabsaz et al. (2020), which includes queries labeled as non-gendered (neutral), female, male, or other/multiple genders, as mentioned in section 3.3.

On this basis, they trained classifiers using both dynamic and static embeddings to predict the gender of queries. The performance of these classifiers was evaluated using a 5-fold cross-validation strategy. The fine-tuned uncased BERT model outperformed others, showing the highest accuracy and F1 scores for gender identification: 0.856 accuracy, 0.816 for female, 0.872 for male, and 0.862 for neutral queries.

Using the fine-tuned BERT model, the researchers labeled all 51,827 queries in the MS MARCO Dev set, resulting in 48,200 neutral queries, 2,222 male queries, and 1,405 female queries. To create a balanced dataset, they retained all 1,405 female queries and randomly selected 1,405 male and 1,405 neutral queries. These labeled queries, along with their associated relevant judgment documents, were used to investigate the presence of stereotypical gender biases.

3.3.4 Grep-BiasIR dataset

The Grep-BiasIR dataset [73], designed to investigate gender representation bias in information retrieval systems, comprises 118 bias-sensitive queries and 708 associated documents. The creation process began with the categorization of queries into seven gender-related stereotypical concepts based on the gender role dimensions introduced by Behm-Morawitz

and Mastro. These categories include Career, Domestic Work, Child Care, Cognitive Capabilities, Physical Capabilities, Appearance, and Sex & Relationship. Each category contains around 15 queries, resulting in a well-rounded dataset that addresses a variety of gender-related topics.

For each query, the dataset includes one relevant and one non-relevant document. The relevant documents were identified by submitting the queries to the Google search engine and selecting documents that fully addressed the query's information need. Non-relevant documents were either taken from the same search results or created by the authors to ensure they did not match the search query. Each document is provided in three variations: male, female, and neutral. The variations maintain the same content, with gender-indicating words modified accordingly. For instance, male indications include words like 'man' and 'he', while female indications use 'woman' and 'she'. Neutral terms like 'person' and 'they' were used to create gender-neutral versions. This thorough and systematic approach ensures that the dataset can effectively facilitate the study of gender biases in information retrieval systems.

The dataset underwent rigorous auditing by two post-doctoral researchers who reviewed each query and document for quality. They judged the items as high, medium, or low quality and only high-quality items were included in the final dataset. This review process also involved checking for ambiguous content and ensuring that gender-neutral documents were properly formulated, such as using only surnames to avoid gender-specific references. Additionally, the reviewers assessed the expected stereotypes for each query based on anticipated gender characteristics and behaviors. This meticulous process of data collection and auditing ensures that the Grep-BiasIR dataset is a reliable and valuable resource for investigating gender representation biases in IR systems.

3.4 Problem Formulation

When considering the issue of gender, user queries can be broadly categorized in two classes, namely (i) *gender-neutral queries*, and (ii) *gender-specific queries*. Gender-neutral queries are those queries, which seek information that can be answered independently of gender considerations and include examples such as 'what happened in cabo shooting' and 'what is early childhood studies'. In contrast, gender-specific queries will need to take gender as a consideration when effectively addressing the query. Examples of such queries include 'when can you feel signs of pregnancy' (female-affiliated query) and 'what is age for prostate cancer' (male-affiliated query). Table 3.2 provides further examples of gender-specific and gender-neutral queries as provided by Rekabsaz et al [108]. The objective of our work is to ensure that a neural ranker Φ is fair when dealing with these two different types of queries. We adopt the definition of gender fairness as laid out by earlier work [118, 13, 75], and formulate them as follows:

Fairness for gender-neutral queries. A neural ranker would be deemed fair when processing gender-neutral queries if the retrieved ranked list of documents would not exhibit any predispositions towards any specific gender. This principle is predicated on the understanding that a query devoid of gender-related cues should yield a set of documents whose relevance is determined independently of gender implications [71, 108, 19]. For instance, the query 'how can one become an engineer?' is gender-neutral, as the path to becoming an engineer is independent of the individual's gender. In such a case, one would expect to receive a ranked list of documents that does not carry any preconceived notions of gender preference in relation to the engineering profession. In order to quantitatively assess the fairness of a ranked list of documents, denoted as R_q, in relation to a gender-neutral query q, researchers have assumed that a function $\Psi(R_q)$ can be formulated for measuring the extent of gender bias manifested by R_q [107, 106, 1] where lower values of $\Psi(R)$ depict increased degrees of fairness. For a ranking R_q to be considered fair in response to a gender-neutral

Table 3.2: Sample queries and their gender affiliations from [107].

Query Gender Affiliation	**Query**
Female	actress who born at litrhuania what does it mean when you bleed before period
Male	who is king philip of spain is king arthur real or legend?
Neutral	where is kobenhavn what is hemianopsia

query, the ideal outcome would be:

$$\Psi(R_q) \to 0 \tag{3.5}$$

This symbolizes the expectation that the ranked list of documents for a gender-neutral query q should approach a state of gender parity, where *ideally* no discernible bias in favor of any gender is observable.

Fairness for gender-specific queries. When processing gender-specific queries, a neural ranker would be deemed fair if its ability to effectively rank documents does not vary based on the gender of the query. This concept posits that the performance of a neural ranker, can be quantitatively assessed using a performance metric $\lambda(Q)$, where a higher $\lambda(Q)$ indicates superior model performance on the query set Q. For a neural ranker to be considered fair under this definition, it must exhibit comparable performance levels for queries belonging to different gender affiliations, such as male-affiliated queries (Q_m), or female-affiliated queries (Q_f), essentially satisfying the following condition:

$$\lambda(Q_m) \approx \lambda(Q_f) \tag{3.6}$$

This definition emphasizes the need for promoting a system that treats all queries equally without bias towards any gender association.

In summary, a fair ranker should not exhibit stereotypical biases toward both *gender-neutral* and *gender-specific* queries. For gender-neutral queries, the goal is to reduce, and ideally remove, biases in the document retrieval process, as shown in Equation 3.5: $\Psi(R_q) \to 0$. For gender-specific queries, the ranker should demonstrate comparable retrieval effectiveness across different gendered queries, as detailed in Equation 3.6: $\lambda(Q_m) \approx \lambda(Q_f)$.

Chapter 4

Loss Function Regularization

4.1 Methodology

4.1.1 Proposed Framework

In this work, we propose a systematic approach to mitigating gender biases in information retrieval systems by introducing bias-aware training into the neural ranker's optimization process. The training procedure is regularized to jointly optimize for document-query relevance and reduce the influence of biases present in the ranked results. Biases are treated as systematic distortions that may need to be minimized to align with the dual objectives of fairness and effectiveness. The loss function, as the primary objective guiding the optimization process, plays a critical role in shaping the training dynamics of neural rankers. By incorporating fairness constraints into the loss function, the model is trained to explicitly account for both relevance and bias mitigation during optimization. This ensures that the ranker produces more balanced outputs while maintaining retrieval performance.

The training process of a neural ranker can be framed probabilistically. Let $\mathcal{T} = \{(q, d_i)\}_{i=1}^{N}$ represent the training dataset, where q is a query and d_i is a document. For each query-document pair, the model $\Phi(q, d_i)$ predicts a relevance score. Additionally, let $\mathcal{Y} = \{y_i\}_{i=1}^{N}$ be the corresponding ground truth relevance labels, where y_i indicates the true

relevance of document d_i to query q. The likelihood of observing the correct ranking outcomes can be expressed as the posterior probability $P(\Phi \mid \mathcal{T}, \mathcal{Y})$. The training objective is to maximize this probability. Equivalently, minimizing the negative log-likelihood gives the loss function:

$$\mathcal{L} = -\mathbb{E}_{\mathcal{T}}\big[\log P(\Phi \mid \mathcal{T}, \mathcal{Y})\big] \tag{4.1}$$

Using Bayes' Rule, the posterior probability can be expanded as:

$$P(\Phi \mid \mathcal{T}, \mathcal{Y}) \propto P(\mathcal{Y} \mid \mathcal{T}, \Phi) \cdot P(\Phi) \tag{4.2}$$

Here, $P(\mathcal{Y} \mid \mathcal{T}, \Phi)$ is the conditional probability of observing the labels $\mathcal{Y}$ given the training data $\mathcal{T}$ and the model predictions Φ, and $P(\Phi)$ is a prior distribution over the model parameters. Assuming a uniform prior $P(\Phi)$, the objective simplifies to:

$$\mathcal{L} = -\mathbb{E}_{\mathcal{T}}\big[\log P(\mathcal{Y} \mid \mathcal{T}, \Phi)\big] \tag{4.3}$$

In this work, we aim to extend standard ranking objectives by incorporating fairness constraints to address gender biases. By integrating bias-awareness into the loss function, we enable the neural ranker to optimize not only for relevance but also for fairness, ensuring that the system produces rankings that are both effective and unbiased. This dual objective is fundamental to our approach, as it allows the ranker to account for and mitigate the harmful effects of biases during training. To achieve this, we introduce two functions, Ψ and ζ, which measure the bias and fairness of the training samples, respectively. The function Ψ quantifies the degree of gender bias in a training sample. A higher Ψ-value for a document d_i in a training pair (q, d_i) indicates a stronger inclination toward a specific gender. Conversely, ζ evaluates the fairness of training samples, where a higher ζ-value indicates that the document d_i exhibits a balanced representation of different genders. We incorporate these measures into the loss function under two distinct scenarios:

1. **Bias Penalty**: Penalizing training samples based on the level of bias measured by Ψ. This approach modifies the log-likelihood loss to create a *bias-aware loss function*, denoted as $\mathcal{L}_{\text{Penalty}}$:

$$\mathcal{L}_{\text{Penalty}} = -\mathbb{E}_{\mathcal{T}}\big[\log P(\mathcal{Y} \mid \mathcal{T}, \Phi, \Psi)\big] \tag{4.4}$$

2. **Fairness Reward**: Rewarding training samples for their fairness based on the values measured by ζ. This approach introduces a *fairness-aware loss function*, denoted as $\mathcal{L}_{\text{Reward}}$:

$$\mathcal{L}_{\text{Reward}} = -\mathbb{E}_{\mathcal{T}}\big[\log P(\mathcal{Y} \mid \mathcal{T}, \Phi, \Psi, \zeta)\big] \tag{4.5}$$

By incorporating these loss functions into the training process, the neural ranker is guided to balance relevance and fairness objectives, potentially reducing gender biases in the ranking process while maintaining retrieval effectiveness. In the following, we outline how $\mathcal{L}_{\text{Penalty}}$ and $\mathcal{L}_{\text{Reward}}$ can be effectively operationalized in practice to penalize and/or reward training samples to obtain a fair neural ranker.

Penalizing *Irrelevant* Documents

In ranking systems, irrelevant documents are expected to be positioned farther from the query in the embedding space, resulting in low relevance scores. However, when irrelevant documents exhibit high levels of bias, their proximity to the query can inadvertently influence the ranking process, potentially propagating biased content. This is particularly problematic as it conflicts with the objectives of fairness and effectiveness by introducing unintended biases into the system's output. To address this, it is necessary to penalize irrelevant documents based on their level of bias. By explicitly discouraging the model from associating biased irrelevant documents with the query, we aim to deprioritize these documents and reduce their influence on the final rankings. Hence, we incorporate bias-awareness into the loss function by adjusting the relevance score of irrelevant documents according to their bias levels.

We propose that the relevance score $\Phi(q, d^-)$ for a query q and an irrelevant document d^- is modified using a bias penalty to produce a bias-aware relevance score $\Phi_B(q, d^-)$, defined as:

$$\Phi_B(q, d^-) = \alpha(\Phi(q, d^-) + \lambda\Psi(d^-)) \tag{4.6}$$

Here, α represents an activation function applied to the adjusted score, while $\Phi_B(q, d^-)$ denotes the bias-adjusted relevance score for the query q and the irrelevant document d^-. The parameter λ controls the influence of the bias penalty, and $\Psi(d^-)$ quantifies the level of bias in the document. By incorporating this bias-aware adjustment, the model is signaled to push biased irrelevant documents farther from the query in the embedding space, effectively reducing their relevance scores. This adjustment ensures that such documents contribute minimally, if they are biased, to the ranking process, supporting the objective of a fairer and more balanced ranking system.

Penalizing *Relevant* Documents

Relevant documents are typically expected to have high relevance scores and should be ranked higher for a given query. However, when these documents exhibit bias, their high relevance can inadvertently amplify unfair patterns in the ranking process. This is particularly concerning because it can perpetuate biased content while still appearing relevant, undermining the fairness of the ranking system. To address this, we propose to adjust the ranking of biased relevant documents, reducing their influence on the final output without diminishing their relevance. Since there are often multiple relevant documents for a single query, it becomes crucial to prioritize documents with lower bias and demote those exhibiting higher levels of bias. The aim is not to discard relevant documents but to ensure that those with less bias are given higher prominence, fostering fairness in the rankings. To implement this, we modify the relevance score of each relevant document based on its bias level, incorporating this adjustment into the loss function. The bias-aware relevance score for a relevant

document d^+ can be defined as:

$$\Phi_B(q, d^+) = \alpha(\Phi(q, d^+) + \lambda\Psi(d^+)) \tag{4.7}$$

where α is an activation function, $\Phi_B(q, d^+)$ represents the bias-adjusted relevance score for the query q and document d^+, λ is a parameter that controls the penalty's strength, and $\Psi(d^+)$ quantifies the bias in the relevant document. This approach ensures that biased relevant documents are penalized in the ranking process, reducing their prominence while preserving the overall relevance of the documents.

Penalizing Both *Irrelevant* and *Relevant* Documents

In this approach, we combine the strategies from the first two scenarios to address bias in both irrelevant and relevant documents. The goal is to ensure that the model accounts for bias across all types of documents, improving fairness throughout the ranking process. For irrelevant documents (d^-), the objective is to move them farther from the query in the embedding space, minimizing their relevance score. To achieve this, we penalize these documents by artificially increasing their relevance score based on their bias level. This adjustment encourages the model to position biased irrelevant documents farther from the query, thus lowering their influence on the final ranking.

For relevant documents (d^+), the goal is to prevent biased relevant documents from being ranked too close to the query. These documents may already have high relevance scores, but their bias can cause them to disproportionately dominate the ranking. To prevent this, we introduce a penalty by increasing their relevance score in proportion to their bias level, signaling the model to avoid overemphasizing them. Therefore, the relevance scores for both irrelevant and relevant documents are adjusted according to the bias measures, as described in Equations 4.6 and 4.7. This combined approach ensures two key outcomes:

1. Biased irrelevant documents are deprioritized, as their relevance scores are minimized

and their position relative to the query is moved farther away.

2. Biased relevant documents are not overemphasized, as their relevance scores are moderated to prevent them from dominating the ranking.

Rewarding *Irrelevant* Documents

In this approach, the goal is to enhance fairness by rewarding irrelevant documents (d^-) that exhibit higher fairness scores. The key assumption is that encouraging fairness in the ranking of irrelevant documents will help the model position them appropriately without amplifying biases. Typically, for irrelevant documents, the aim is to minimize their relevance score $\Phi(q, d^-)$. However, to prevent overly penalizing fair irrelevant documents, we adjust their relevance score by decreasing it in proportion to their fairness score $\zeta(d^-)$. This adjustment prevents the model from excessively pushing fair irrelevant documents away from the query in the embedding space, helping to maintain a more equitable ranking. The adjusted relevance score for an irrelevant document is given by:

$$\Phi_R(q, d^-) = \alpha \left(\Phi(q, d^-) - \lambda \zeta(d^-) \right) \tag{4.8}$$

where α is an activation function applied to the relevance score, Φ_R is the fairness-adjusted relevance score for the query q and irrelevant document d^-, λ is a coefficient controlling the magnitude of the fairness reward, and $\zeta(d^-)$ represents the fairness of document d^-. By incorporating this fairness reward, the model ensures that irrelevant documents that are fair are not penalized too harshly, allowing them to be ranked appropriately within the overall system while avoiding the introduction of bias. This strategy ultimately helps maintain fairness across the ranking process.

Rewarding *Relevant* Documents

In this approach, we aim to encourage fairness in the ranking of relevant documents (d^+) by rewarding those with higher fairness scores. The hypothesis is that prioritizing fairness in relevant documents can help the model rank them appropriately while maintaining overall ranking effectiveness. For relevant documents, the primary objective is to maximize their relevance score $\Phi(q, d^+)$. However, when a relevant document is biased, we adjust its relevance score downward based on its fairness score $\zeta(d^+)$. This adjustment signals to the model that the document is already sufficiently close to the query and that its position should not be further emphasized, ensuring that fairness is prioritized. The reward-adjusted relevance score for a relevant document is defined as:

$$\Phi_R(q, d^+) = \alpha \left(\Phi(q, d^+) - \lambda\zeta(d^+)\right) \tag{4.9}$$

where α is an activation function, Φ_R is the adjusted relevance score for the query q and the relevant document d^+, λ is a hyperparameter controlling the strength of the fairness reward, and $\zeta(d^+)$ quantifies the fairness of the document. By incorporating this adjustment into the ranking, the model is encouraged to prioritize relevant documents that exhibit fairness, ensuring that they are ranked in a way that is not only relevant but also equitable.

Rewarding Both *Irrelevant* and *Relevant* Documents

Simultaneously rewarding both irrelevant and relevant documents based on their fairness scores promotes fairness throughout the ranking system while preserving relevance. Adjusting relevance scores according to fairness ensures that fair documents are prioritized appropriately, regardless of their relevance category, and minimizes the impact of biased documents. For irrelevant documents (d^-), the relevance score is decreased by an amount proportional to their fairness score $\zeta(d^-)$. This rewards fair irrelevant documents by signaling that their relevance is already sufficiently low, and further adjustment is unnecessary.

On the other hand, for relevant documents (d^+), the relevance score is adjusted downward in proportion to their fairness score $\zeta(d^+)$, guiding the model to prioritize those that are fair and pushing them closer to the query in the embedding space. By adjusting the relevance scores of both irrelevant and relevant documents as shown in Equations 4.8 and 4.9, we ensure that fairness is promoted across the entire ranking process. This adjustment helps the model learn to balance fairness and relevance, leading to a more equitable ranking system that treats bias as an important factor in determining the final outputs. As a result, the model is better equipped to rank documents in a manner that reflects both relevance and fairness, enhancing the quality and inclusiveness of the ranking system.

In the next section, we will concretely demonstrate how these six strategies—penalizing and rewarding both irrelevant and relevant documents—can be incorporated into the loss function of pointwise [32, 70] and pairwise [64, 29] neural ranking models. Pointwise models, which focus on predicting the relevance of individual documents for a given query, provide a simple yet effective approach for incorporating fairness adjustments on a per-document basis. They are well-suited to scenarios where the focus is on ensuring that each document is fairly ranked relative to a given query. Pairwise models, on the other hand, consider the relative ordering between document pairs, making them particularly effective in scenarios where ranking accuracy and fairness need to be optimized in terms of the relative positioning of documents. By comparing pairs of documents and learning the correct order, pairwise models can directly address situations where the fairness of one document should influence the position of another, especially when the documents are similar in relevance but differ in bias or fairness. Overall, Pointwise models are simpler to train and evaluate, while pairwise models tend to provide stronger performance in ranking tasks where the relative order of documents is critical. In the next section, we will outline how to integrate our proposed fairness adjustments into both types of models, providing concrete examples of how each strategy can be operationalized in practice.

4.1.2 Fair Pointwise Neural Rankers

Pointwise neural rankers treat the ranking task as a regression or classification problem by independently predicting the relevance of each query-document pair. The model is trained to assign scores that match the ground-truth relevance labels for each pair, focusing on individual query-document relevance without considering pairwise relationships. One key strength of pointwise models is their simplicity and ease of implementation, making them particularly well-suited for datasets where relevance labels are explicitly provided for individual query-document pairs. This makes them an attractive choice when the task focuses on optimizing relevance scores for each document independently, as is common in ranking tasks with clearly labeled datasets. For a given query q and document d_i, the relevance likelihood can be formulated as:

$$P(\mathcal{Y} \mid \mathcal{T}, \Phi) = \prod_i P(y_i \mid \Phi(q, d_i)) \tag{4.10}$$

Here, $P(y_i \mid \Phi(q, d_i))$ denotes the probability of observing the true relevance label y_i, given the predicted relevance score $\Phi(q, d_i)$. Assuming that the relevance label y_i follows a logistic distribution, we model this probability as:

$$P(y_i \mid \Phi(q, d_i)) = \frac{1}{1 + e^{-\Phi(q,d_i)}} \tag{4.11}$$

In this context, the relevance score $\Phi(q, d_i)$ is interpreted as the logit of the relevance score. The log-likelihood for the pointwise loss function is then:

$$\mathcal{L}_{\text{pointwise}} = -\sum_i \log P(y_i \mid \Phi(q, d_i)) = \sum_i \log \left(1 + e^{-y_i \cdot \Phi(q,d_i)}\right) \tag{4.12}$$

This formulation can be simplified by using the Mean Squared Error (MSE) loss, which penalizes the squared difference between the predicted relevance score $\Phi(q, d_i)$ and the ground

truth label y_i. The MSE-based formulation is:

$$\mathcal{L}_{\text{pointwise}} = \sum_i \left(\frac{1}{1 + e^{-\Phi(q,d_i)}} - y_i \right)^2 \tag{4.13}$$

The simplicity of pointwise models makes them an ideal starting point for exploring fairness in ranking tasks. By adjusting the predicted relevance score based on fairness measures, we directly incorporate the impact of gender biases or fairness issues into the model's optimization process. This modification enhances the model's ability to not only optimize relevance but also improve fairness in ranking results.

Penalizing Documents

To incorporate fairness into the pointwise loss function, we introduce a *unified bias-aware penalty framework* that can be applied to irrelevant documents, relevant documents, or both, depending on the specific scenario. The bias-adjusted loss function is defined as:

$$\mathcal{L}_{\text{Penalty}} = \sum_i \left[\left(\frac{1}{1 + e^{-(\Phi(q,d_i) + (1-y_i)\cdot\Psi(d_i) + y_i\cdot\Psi(d_i))}} \right) - y_i \right]^2 \tag{4.14}$$

In this formulation, y_i represents the ground-truth relevance label for the document d_i with respect to the query q, where $y_i = 1$ indicates a relevant document and $y_i = 0$ indicates an irrelevant document. The term $\Psi(d_i)$ quantifies the bias of the document d_i, serving as a measure of how strongly the document deviates from fairness. The two terms, $(1 - y_i) \cdot \Psi(d_i)$ and $y_i \cdot \Psi(d_i)$, selectively apply the penalty based on the relevance label y_i. Specifically, $(1 - y_i) \cdot \Psi(d_i)$ applies the penalty to biased irrelevant documents ($y_i = 0$), ensuring that such documents are deprioritized in the ranking. Similarly, $y_i \cdot \Psi(d_i)$ applies the penalty to biased relevant documents ($y_i = 1$), discouraging their overemphasis in the ranking. This unified formulation supports three distinct scenarios:

1. For *penalizing irrelevant documents* (Section 4.1.1), the loss function applies a bias-aware penalty only to documents labeled as irrelevant ($y_i = 0$), encouraging the model

to reduce their prominence in the ranking while accounting for their bias.

2. For *penalizing relevant documents* (Section 4.1.1), the loss function applies the penalty only to documents labeled as relevant ($y_i = 1$), preventing biased relevant documents from being overly emphasized in the ranking.

3. For *penalizing both irrelevant and relevant documents* (Section 4.1.1), the loss function applies penalties to all documents based on their bias levels, regardless of their relevance labels.

To demonstrate the effect of regularizing the loss function, we analyze the impact of incorporating a bias-aware penalty term in the case of penalizing irrelevant documents. The gradient of the loss function with respect to the relevance score $\Phi(q, d_i)$ is calculated as follows:

$$\frac{\partial \mathcal{L}_{\text{Penalty}}}{\partial \Phi(q, d_i)} = 2\big[\sigma(\Phi(q, d_i) + \Psi(d_i)) - y_i\big] \cdot \sigma(\Phi(q, d_i) + \Psi(d_i)) \cdot \big(1 - \sigma(\Phi(q, d_i) + \Psi(d_i))\big), \quad (4.15)$$

where σ is the sigmoid activation function. The penalty term $\Psi(d_i)$ modifies the input to the sigmoid function and plays a critical role in shaping the gradient. The impact of $\Psi(d_i)$ on the gradient can be interpreted as follows. First, the penalty term $\Psi(d_i)$ shifts the input $\Phi(q, d_i)$ to $\Phi(q, d_i) + \Psi(d_i)$. A positive bias score ($\Psi(d_i) > 0$) increases the effective relevance score, causing the sigmoid output $\sigma(\Phi(q, d_i) + \Psi(d_i))$ to move closer to one. Conversely, a negative bias score ($\Psi(d_i) < 0$) reduces the effective relevance score, pushing the sigmoid output closer to zero. Second, the gradient is sensitive to the magnitude of $\Psi(d_i)$, ensuring that highly biased documents ($\Psi(d_i) > 0$) induce a stronger adjustment in $\Phi(q, d_i)$. This encourages the model to correct for biases by appropriately adjusting relevance predictions during backpropagation.

Algorithm 1 outlines the training procedure for a ranking network using a bias-aware pointwise loss function in the case of penalizing irrelevant documents. The process begins by initializing the model parameters (θ, b) randomly (Line 2). Over a specified number

Algorithm 1 Training of the Ranking Network with the Bias-Aware Pointwise Loss.

1: *Data:* $\{(q, d, y)\}$, number of training iterations T.
2: *Initialize:* θ, b randomly. **for** $t = 1$ to T **do**
   each sample (q, d, y) in the batch
3: $E \leftarrow \text{encoder}(q \oplus d)$
4: $s \leftarrow \sigma(\theta E + b)$
5: $\Psi(d) \leftarrow y \cdot \Psi(d)$
6: $\mathcal{L}^{\text{neg}}_{\text{Penalty}} \leftarrow \sum_i \left[\left(\frac{1}{1+e^{-(\Phi(q,d_i)+(1-y_i)\cdot\Psi(d_i))}}\right) - y_i\right]^2$
7: $\frac{\partial \mathcal{L}_{\text{Penalty}}}{\partial \Phi(q,d_i)} \leftarrow 2\big[\sigma(\Phi(q,d_i)+\Psi(d_i)) - y_i\big] \cdot \sigma(\Phi(q,d_i)+\Psi(d_i)) \cdot \big(1-\sigma(\Phi(q,d_i)+\Psi(d_i))\big)$
8: $\theta^{(t+1)} = \theta^{(t)} - \eta \frac{\partial \mathcal{L}_{\text{Penalty}}}{\partial \Phi(q,d_i)} \frac{\partial \Phi(q,d_i)}{\partial \theta}$
9: $b^{(t+1)} = b^{(t)} - \eta \frac{\partial \mathcal{L}_{\text{Penalty}}}{\partial \Phi(q,d_i)} \frac{\partial \Phi(q,d_i)}{\partial b}$
10:
11:

of iterations T (Line 3), the model processes each query-document pair (q, d, y) from the training batch (Line 4). For each pair, the query and document are encoded into embeddings E (Line 5), which are used to compute the relevance score s via a sigmoid activation function parameterized by θ and b (Line 6). The bias term $\Psi(d)$ is updated based on the relevance label y to modulate its impact on the loss (Line 7). The bias-aware penalty loss $\mathcal{L}^{\text{neg}}_{\text{Penalty}}$ is then calculated (Line 8). The gradient of this penalty loss with respect to $\Phi(q, d_i)$ is derived (Line 9) using the sigmoid's derivative to capture the sensitivity of relevance scores to parameter updates. Finally, the model parameters θ and b are updated via gradient descent (Line 10), weighted by the learning rate η and the gradient of the bias-aware penalty. This iterative process optimizes relevance predictions while mitigating bias, resulting in a fairer ranking system.

Rewarding Documents

To incorporate fairness rewards into the pointwise loss function, we introduce a *unified bias-aware reward framework* that applies to irrelevant documents, relevant documents, or both, depending on the specific case. The reward-adjusted loss function is defined as:

$$\mathcal{L}_{\text{Reward}} = \sum_i \left[\left(\frac{1}{1+e^{-(\Phi(q,d_i)-(1-y_i)\cdot\zeta(d_i)-y_i\cdot\zeta(d_i))}}\right) - y_i\right]^2 \tag{4.16}$$

In this formulation, y_i represents the ground-truth relevance label for the document d_i, where $y_i = 1$ for relevant documents and $y_i = 0$ for irrelevant documents. The term $\zeta(d_i)$ measures the fairness of the document d_i, serving as a reward for documents with higher fairness. The expression $(1 - y_i) \cdot \zeta(d_i)$ ensures that fairness rewards are applied only to irrelevant documents ($y_i = 0$), encouraging the model to deprioritize biased irrelevant documents while recognizing fairness. Similarly, the term $y_i \cdot \zeta(d_i)$ applies fairness rewards to relevant documents ($y_i = 1$), helping the model prioritize fair relevant documents over biased ones.

This unified framework integrates the following cases:

1. When *rewarding irrelevant documents* (Section 4.1.1), the loss function applies fairness rewards exclusively to documents labeled as irrelevant ($y_i = 0$). This signals the model to deprioritize biased irrelevant documents while maintaining fairness in the ranking process.

2. When *rewarding relevant documents* (Section 4.1.1), the fairness rewards are applied exclusively to documents labeled as relevant ($y_i = 1$). This encourages the model to rank fair relevant documents higher, ensuring that fairness is prioritized in relevant document rankings.

3. When *rewarding both irrelevant and relevant documents* (Section 4.1.1), the fairness rewards are applied simultaneously to all documents regardless of their relevance label. This ensures a comprehensive approach to mitigating biases across all document types, balancing fairness and relevance in the ranking system.

4.1.3 Fair Pairwise Neural Rankers

Unlike the pointwise approach, pairwise neural rankers focus on the relative ordering of documents for a given query. The primary objective is to train the model to ensure that the predicted relevance score for a relevant document is higher than that of an irrelevant

document. This method is particularly well-aligned with the goals of ranking tasks, where the relative ordering of documents often carries more importance than their absolute relevance scores. Rather than predicting the relevance of individual documents, the pairwise approach models the relative preference between a relevant document d^+ and an irrelevant document d^- for a given query q. The conditional probability for a pairwise ranking approach is expressed as:

$$P(\mathcal{Y} \mid \mathcal{T}, \Phi) = \prod_{(i,j)} P\big((y_i, y_j) \mid (\Phi(q, d_i), \Phi(q, d_j))\big), \tag{4.17}$$

where $P\big((y_i, y_j) \mid (\Phi(q, d_i), \Phi(q, d_j))\big)$ denotes the likelihood of document d_i being more relevant than document d_j. This likelihood is modeled using a sigmoid function as follows:

$$P\big((y_i, y_j) \mid (\Phi(q, d_i), \Phi(q, d_j))\big) = \sigma\big(\Phi(q, d_i) - \Phi(q, d_j)\big), \quad \sigma(z) = \frac{1}{1 + \exp(-z)}. \tag{4.18}$$

The training objective for pairwise ranking is to maximize this probability, which translates into minimizing the negative log-likelihood:

$$\mathcal{L}_{\text{pairwise}} = -\sum_{(i,j)} \log P\big((y_i, y_j) \mid (\Phi(q, d_i), \Phi(q, d_j))\big). \tag{4.19}$$

Expanding this expression gives the following loss function:

$$\mathcal{L}_{\text{pairwise}} = \sum_{(i,j)} \log\big(1 + \exp(-(\Phi(q, d_i) - \Phi(q, d_j)))\big). \tag{4.20}$$

This formulation directly optimizes the ranking order by penalizing cases where a relevant document d^+ is not scored higher than an irrelevant document d^-. To further enforce a significant margin between the relevance scores of relevant and irrelevant documents, we incorporate a hinge-like loss function. This loss penalizes situations where the difference in

scores is less than a predefined margin m:

$$\mathcal{L}_{\text{margin}} = \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - \Phi(q, d^+) + \Phi(q, d^-)). \tag{4.21}$$

Here, N^+ and N^- represent the number of relevant and irrelevant documents, respectively, and m is a hyperparameter that defines the minimum desired gap between scores. The hinge-like behavior of this loss encourages the model to prioritize clear distinctions between the scores of relevant and irrelevant documents, thereby improving the robustness of the ranking system.

Penalizing Documents

To incorporate fairness into pairwise loss functions, we propose a *unified penalty framework* that applies to biased irrelevant documents, biased relevant documents, or both. The penalty-adjusted pairwise loss function is defined as:

$$\mathcal{L}_{\text{Penalty}} = \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - (\tanh(\Phi(q, d^+)) + \lambda\Psi(d^+)) + (\tanh(\Phi(q, d^-)) + \lambda\Psi(d^-))), \tag{4.22}$$

where $\Phi(q, d^+)$ and $\Phi(q, d^-)$ represent the predicted relevance scores for the relevant and irrelevant documents, respectively, and $\Psi(d^+)$ and $\Psi(d^-)$ denote the bias scores for the relevant and irrelevant documents. The hyperparameter λ controls the strength of the bias penalty, and m defines the desired margin between the relevance scores of the two document types. The loss penalizes pairs where the relevance score gap between d^+ and d^- is insufficient due to the presence of bias. This formulation unifies the three cases where we may apply penalties to documents:

1. For *penalizing biased irrelevant documents* (Section 4.1.1), the term $\lambda\Psi(d^-)$ increases the relevance score of biased irrelevant documents, encouraging the model to push their scores further down and create a clearer distinction from relevant documents.

Algorithm 2 Training of the Ranking Network with the Bias-aware Pair-wise Loss.

Data: $\{(q, d, y)\}$, number of training iterations T.
Initialize: θ, b randomly.
for $t = 1$ **to** T **do**
each sample (q, d, y) in the batch
$E^+ \leftarrow \text{encoder}(q \oplus d^+)$
$E^- \leftarrow \text{encoder}(q \oplus d^-)$
$\Phi(q, d^+) \leftarrow \sigma(\theta E^+ + b)$
$\Phi(q, d^-) \leftarrow \sigma(\theta E^- + b)$
Calculate $\Psi(d^-)$
$\mathcal{L}^{\text{neg}}_{\text{Penalty}} = \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - \tanh(\Phi(q, d^+)) + (\tanh(\Phi(q, d^-) + \lambda\Psi(d^-)))$
$\frac{\partial \mathcal{L}^{\text{neg}}_{\text{Penalty}}}{\partial \Phi(q,d)} \leftarrow \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \mathbb{1}_{z>0} \cdot (1 - \tanh^2(\Phi(q, d) + \lambda y \Psi(d)))$
$\theta^{(t+1)} = \theta^{(t)} - \eta \frac{\partial \mathcal{L}_{\text{Penalty}}}{\partial \Phi(q,d_i)} \frac{\partial \Phi(q,d_i)}{\partial \theta}, \quad b^{(t+1)} = b^{(t)} - \eta \frac{\partial \mathcal{L}_{\text{Penalty}}}{\partial \Phi(q,d_i)} \frac{\partial \Phi(q,d_i)}{\partial \theta}$

2. For *penalizing biased relevant documents* (Section 4.1.1), the term $\lambda\Psi(d^+)$ increases the relevance score of biased relevant documents, discouraging the model from over-ranking them in comparison to irrelevant documents.

3. For *penalizing both types of documents* (Section 4.1.1), both $\Psi(d^+)$ and $\Psi(d^-)$ are incorporated into the loss, simultaneously addressing biases in both relevant and irrelevant documents.

We show how the gradient of the penalty-adjusted loss function is computed when only applying it to irrelevant documents. We calculate the gradient of the loss function as follows:

$$\frac{\partial \mathcal{L}^{\text{neg}}_{\text{Penalty}}}{\partial \Phi(q, d)} = \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \mathbb{1}_{z>0} \cdot (1 - \tanh^2(\Phi(q, d) + \lambda y \Psi(d))) \tag{4.23}$$

The bias term $\Psi(d^-)$ shifts the gradient of $\Phi(q, d^-)$ through the tanh function, directly influencing how strongly the model adjusts the relevance score for biased irrelevant documents. This mechanism ensures that documents with higher bias are penalized more during backpropagation, promoting fairness in the ranking process.

Algorithm 2 outlines the training procedure for a ranking network using the bias-aware pairwise loss. The process begins with initializing the model parameters θ and the bias

term b randomly (Line 2). Training proceeds over T iterations, during which each query-document pair (q, d^+, d^-) in the batch is processed iteratively (Line 4). For each sample, the encoder generates embeddings E^+ and E^- for the relevant document d^+ and the irrelevant document d^-, concatenated with the query q (Lines 5–6). These embeddings represent the query-document pairs in a latent semantic space. The algorithm then computes the relevance scores $\Phi(q, d^+)$ and $\Phi(q, d^-)$ for the relevant and irrelevant documents, respectively, using a sigmoid function parameterized by θ and b (Lines 7–8).

The bias score $\Psi(d^-)$ for the irrelevant document is then calculated, which captures the degree of bias associated with d^- (Line 9). Using these scores, the bias-aware pairwise loss $\mathcal{L}^{\text{neg}}_{\text{Penalty}}$ is computed (Line 10). This loss ensures that the margin between the relevance scores of relevant and irrelevant documents is adjusted to account for the bias in d^-, applying a penalty proportional to $\Psi(d^-)$. The gradient of the loss function with respect to the relevance scores is computed (Line 11), incorporating the derivative of the tanh function and the bias term $\lambda\Psi(d^-)$. This allows the model to dynamically adjust its parameters based on both relevance and bias. Finally, the model parameters θ and b are updated using gradient descent, with the learning rate η controlling the step size for each update (Line 12). This iterative process ensures that the ranking network learns to prioritize relevance while mitigating the influence of bias, resulting in a fairer and more effective ranking system.

Rewarding Documents

To incorporate fairness into the pairwise loss function, we propose a *unified reward framework* that addresses fair irrelevant documents, fair relevant documents, or both. The reward-adjusted pairwise loss function is defined as:

$$\mathcal{L}_{\text{Reward}} = \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - (\tanh(\Phi(q, d^+)) - \lambda\zeta(d^+)) + (\tanh(\Phi(q, d^-)) - \lambda\zeta(d^-))), \tag{4.24}$$

where $\Phi(q, d^+)$ and $\Phi(q, d^-)$ are the predicted relevance scores for the relevant and ir-

relevant documents, respectively. The fairness scores $\zeta(d^+)$ and $\zeta(d^-)$ reflect the degree of fairness associated with each document. The hyperparameter λ controls the weight of the fairness reward, and m ensures a sufficient margin between the scores of relevant and irrelevant documents. The framework addresses the following cases:

- For *irrelevant documents* (Section 4.1.1), fairness rewards are applied by reducing their relevance scores via $\lambda\zeta(d^-)$. This ensures that fair irrelevant documents maintain their low rank without unnecessary penalization.
- For *relevant documents* (Section 4.1.1), fairness rewards are applied by reducing their relevance scores via $\lambda\zeta(d^+)$. This encourages the model to prioritize fair relevant documents over others.
- When applied to *both relevant and irrelevant documents* (Section 4.1.1), fairness adjustments ensure that fairness considerations span all document types, creating a more balanced ranking system.

The gradients and the algorithms can be written out similarly to those already shown for penalizing documents.

4.1.4 Theoretical Justification

Before providing our empirical findings, we provide a theoretical analysis of the proposed fairness-aware framework. Our goal is to formalize how integrating bias and fairness terms into the loss function shapes the optimization trajectory of neural rankers. We show that the proposed modifications amplify the gradient signal for biased content, encourage demotion of such documents, and achieve a principled trade-off between relevance and fairness.

Lemma 1 (Bias Penalty Gradient Amplification)**.** *Let $\mathcal{L}_{Penalty}$ be the pairwise margin-based*

ranking loss regularized with a bias term:

$$\mathcal{L}_{Penalty} = \frac{1}{n}\sum_{i=1}^{N^+}\sum_{j=1}^{N^-}\max\left(0, m - \left(\Phi(q,d^+) + \lambda\Psi(d^+)\right) + \left(\Phi(q,d^-) + \lambda\Psi(d^-)\right)\right)$$

*Assume $\Psi(d) \geq 0$ and is Lipschitz continuous [**?**]. Then the gradient $\frac{\partial \mathcal{L}_{Penalty}}{\partial \theta}$ increases monotonically with $\Psi(d^-)$, causing the model to push more biased irrelevant documents farther from the query in embedding space.*

Sketch. By differentiating the loss with respect to $\Phi(q, d^-)$ and applying the chain rule, the gradient includes the term $\lambda\frac{\partial\Psi(d^-)}{\partial\theta}$. Since $\Psi(d^-)$ is non-negative and Lipschitz, the gradient magnitude increases with the bias score, leading to stronger penalization of biased documents. □

Theorem 4.1.1 (Fairness-Constrained Optimization Yields Pareto-Optimal Trade-offs)**.** *Let the total loss be defined as $\mathcal{L}_{Reward}(\theta)$ or $\mathcal{L}_{Penalty}(\theta)$, where the objective combines a standard pairwise relevance loss with additional terms based on document-level bias $\Psi(\cdot)$ or fairness $\zeta(\cdot)$. For any $\lambda > 0$, every local minimum θ^* of the modified loss function represents a solution on the Pareto frontier of relevance and fairness.*

Sketch. The loss functions $\mathcal{L}_{\text{Penalty}}$ and $\mathcal{L}_{\text{Reward}}$ can be written as scalarized combinations of two objectives: the pairwise relevance loss $\mathcal{L}_{\text{pairwise}}(\theta)$, and a fairness term involving either $\Psi(d)$ or $\zeta(d)$, which we denote generically as $\mathcal{L}_{\text{fairness}}(\theta)$. That is,

$$\mathcal{L}_{\text{fair}}(\theta) = \mathcal{L}_{\text{pairwise}}(\theta) + \lambda\mathcal{L}_{\text{fairness}}(\theta).$$

Assuming both $\mathcal{L}_{\text{pairwise}}$ and $\mathcal{L}_{\text{fairness}}$ are continuous and differentiable with respect to θ, and that $\lambda > 0$, this formulation constitutes a scalarization of a bi-objective optimization problem. According to classical results in multi-objective optimization theory [**?**], minimizing a weighted sum of continuous objectives yields a solution that lies on the Pareto frontier, provided the objectives are conflicting to some extent (as is the case here, where improving

fairness may slightly compromise relevance). Therefore, any local minimum θ^* of $\mathcal{L}_{\text{fair}}(\theta)$ represents a Pareto-optimal solution in the trade-off space between minimizing relevance error and improving fairness, as encoded via Ψ or ζ. □

Corollary 4.1.1.1 (Bounded Fairness–Relevance Trade-off)**.** *Assume that the cumulative contribution of the bias or fairness regularization term (i.e., based on $\Psi(\cdot)$ or $\zeta(\cdot)$) is bounded above by β, and that the pairwise relevance loss is Lipschitz continuous. Then, for any $\lambda > 0$, the difference in relevance loss between the model trained with a fairness-aware loss $\mathcal{L}_{Penalty}$ or $\mathcal{L}_{Reward}$, and a model trained solely for relevance, is bounded as:*

$$|\mathcal{L}_{pairwise}(\theta^*) - \mathcal{L}_{pairwise}(\theta_0)| \leq \lambda\beta$$

where $\theta_0 = \arg\min \mathcal{L}_{pairwise}$ and $\theta^* = \arg\min \mathcal{L}_{Penalty}$ or $\arg\min \mathcal{L}_{Reward}$.

Sketch. Let $\mathcal{L}_{\text{pairwise}}(\theta)$ be the original relevance-based loss function and let $\mathcal{L}_{\text{Penalty}}(\theta) = \mathcal{L}_{\text{pairwise}}(\theta) + \lambda\mathcal{L}_{\text{bias}}(\theta)$, where $\mathcal{L}_{\text{bias}}(\theta)$ is the additional term based on $\Psi(d)$, or similarly for $\mathcal{L}_{\text{Reward}}(\theta)$ using $\zeta(d)$. Assume $\mathcal{L}_{\text{bias}}(\theta) \leq \beta$ for all $\theta \in \Theta$, and that $\mathcal{L}_{\text{pairwise}}$ is L-Lipschitz continuous. Let $\theta^* = \arg\min \mathcal{L}_{\text{Penalty}}(\theta)$ and $\theta_0 = \arg\min \mathcal{L}_{\text{pairwise}}(\theta)$. By the definition of $\mathcal{L}_{\text{Penalty}}$, we have:

$$\mathcal{L}_{\text{pairwise}}(\theta^*) + \lambda\mathcal{L}_{\text{bias}}(\theta^*) \leq \mathcal{L}_{\text{pairwise}}(\theta_0) + \lambda\mathcal{L}_{\text{bias}}(\theta_0).$$

Rearranging gives:

$$\mathcal{L}_{\text{pairwise}}(\theta^*) - \mathcal{L}_{\text{pairwise}}(\theta_0) \leq \lambda\left(\mathcal{L}_{\text{bias}}(\theta_0) - \mathcal{L}_{\text{bias}}(\theta^*)\right) \leq \lambda\beta.$$

Thus, the increase in relevance loss due to fairness-aware training is upper-bounded by $\lambda\beta$, which completes the proof. □

These results provide a theoretical foundation for our fairness-aware loss design. The bias penalty increases gradient pressure on biased documents, while fairness-constrained

optimization ensures that the model trades off relevance and fairness in a controlled, Pareto-optimal manner. The bounded trade-off further assures that gains in fairness do not disproportionately degrade retrieval effectiveness.

4.2 Experiments

4.2.1 Research Questions

Our experiments are designed to address five key research questions:

- **RQ1:** *Are the proposed fairness-aware loss regularization scenarios effective in reducing gender bias in ranked results?* To evaluate this, we apply the six proposed scenarios—penalizing or rewarding relevant, irrelevant, or both document types—on both pointwise and pairwise ranking loss functions. We assess their effectiveness in mitigating gender bias while maintaining ranking effectiveness.

- **RQ2:** *Is the proposed fairness-aware framework generalizable across different pre-trained language models used as encoders?* To answer this, we conduct experiments using two base language models, *BERT-mini* and *ELECTRA-small*, and evaluate the consistency of results across these encoders.

- **RQ3:** *Which type of loss function—pointwise or pairwise—is more amenable to becoming a fair ranker?* This research question investigates the comparative performance of pointwise and pairwise rankers under the six proposed fairness-aware loss regularization scenarios, focusing on their ability to balance bias mitigation and retrieval effectiveness across diverse datasets.

- **RQ4:** *How does the choice of the regularization coefficient (λ) impact the performance of the fairness-aware framework?* We test the best-performing models from both pointwise and pairwise loss functions with various values of the regularization

coefficient ($\lambda \in \{0.1, 0.5, 1, 2, 5\}$) and analyze its effect on retrieval effectiveness and fairness.

- **RQ5:** *How does the proposed fairness-aware framework compare to state-of-the-art fairness-aware ranking methods?* To explore this, we benchmark our best-performing fairness-aware method against three state-of-the-art approaches:

    1. `AdvBERT`: An adversarial debiasing method applied to ranking models' intermediate layers [105].
    2. `CODER`: A transformer-based model that incorporates neutrality regularization [138].
    3. `Light-Weight Sampling Strategy (LWS)`: A bias-aware negative sampling approach that trains models to mitigate bias [17].

4.2.2 Dataset, and Setup

Datasets and Setup. We conduct our experiments on the MSMARCO passage ranking dataset [89], which consists of approximately 200,000 queries and 8.8 million passages. For training, we use a randomly sampled subset of 3,000,000 query-passage pairs, processed over one epoch with the Adam optimizer and a sigmoid activation function. Our neural rankers are implemented using the OpenMatch framework [78], leveraging its architecture, implementation, and hyperparameter settings to ensure consistency with prior work. To measure document-level bias, we employ the ARaB-TF function defined in [107] to quantify Ψ used in Section 3. Full implementation details and the source code for our work are publicly available on GitHub[1].

Evaluation Queries. To evaluate the reduction of bias and ranking performance, we focus on **gender bias** across two distinct types of query sets:

[1] `https://github.com/fairnesspaper/fairnesspaper`

- **Gender-neutral queries:** These queries are used to assess whether the ranker introduces gender stereotypes in contexts where no explicit gender association is expected. We adopt the query set curated by Rekabsaz et al. [107], consisting of 1,765 queries annotated by three Amazon Mechanical Turk workers. These queries were derived from an initial pool of 55,578 MSMARCO queries selected based on gender association. Ideally, retrieved results for these queries should exhibit no gender preference.

- **Socially sensitive queries:** These queries consist of 215 examples that are more likely to propagate stereotypes or reinforce gender inequality if bias is present in the rankings. These queries are designed to evaluate the ranker's ability to mitigate biases in contexts with inherent societal sensitivity [105].

Evaluation Metrics. We evaluate the models on two key aspects: **ranking effectiveness** and **gender bias**. For ranking effectiveness, we use the **Mean Reciprocal Rank (MRR)**, reporting MRR@10 as the standard benchmark metric for the MSMARCO dataset [89]. To assess gender bias, we employ three complementary metrics:

- **Average Rank Bias (ARaB)** [107]: This metric quantifies the presence of gendered terms in ranked documents using both Term Frequency (TF) and Boolean metrics to capture bias at the document level.

- **NFaiRR** [105]: A document-level fairness metric designed to evaluate ranking fairness, where higher values indicate more equitable rankings with respect to gender-neutral queries.

- **Linguistic Inquiry and Word Count (LIWC)** [95]: This metric examines the frequency of gendered terms in retrieved text by counting references to male and female pronouns, providing insights into the linguistic attributes of retrieved content.

Table 4.1: Performance of the model across the six proposed scenarios using the pairwise loss function with the "BERT-mini" base model on the 215-query dataset [105].

		Cut-off@10					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Penalty	Irrelevant	-3.51	-15.66	-15.02	-13.17	-6.76	1.83
	Relevant	10.72	-60.62	-59.52	-60.76	-28.26	8.22
	Both	8.83	-48.20	-46.31	-45.53	-42.92	11.95
Reward	Irrelevant	-16.00	-96.83	-96.62	-95.06	-39.16	10.62
	Relevant	-5.42	-87.32	-85.55	-84.64	-29.84	8.69
	Both	10.84	-51.00	-47.79	-44.72	-33.47	9.56
		Cut-off@20					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Penalty	Irrelevant	-3.81	-12.64	-13.29	-12.47	-6.72	1.47
	Relevant	8.92	-51.97	-52.06	-54.35	-16.03	5.62
	Both	6.80	-43.53	-40.87	-38.93	-36.14	9.01
Reward	Irrelevant	-15.93	-91.28	-92.19	-92.17	-30.78	7.21
	Relevant	-5.42	-77.59	-78.81	-80.97	-23.02	5.61
	Both	9.10	-44.31	-41.18	-38.93	-25.67	7.12

4.2.3 Findings

Findings for RQ1. Tables 1 and 2 present the results of the six proposed bias mitigation scenarios applied using the pairwise loss function on two datasets: the 215 socially sensitive queries and the 1,765 gender-neutral queries. The findings reveal that scenarios involving penalties, particularly *penalty on relevant documents* and *penalty on both types of documents*, consistently demonstrate strong bias mitigation across bias metrics like *ARaB-TC* and *LIWC*. However, these scenarios often result in modest trade-offs in ranking effectiveness, as indicated by marginal reductions in metrics like *MRR@10*. Rewarding scenarios, such as *rewarding irrelevant documents* or *rewarding both types of documents*, exhibit mixed results: while they achieve substantial bias reduction, their impact on ranking performance varies, with some cases showing significant drops in *MRR@10*. These results suggest that penalty-based strategies tend to achieve more consistent bias mitigation, albeit with slight compromises in ranking effectiveness.

Tables 3 and 4 focus on the pointwise loss function, revealing smaller overall changes in ranking performance compared to the pairwise approach. Similar to the pairwise results, scenarios applying penalties, especially to both relevant and irrelevant documents, exhibit

.

Table 4.2: Performance of the model across the six proposed scenarios using the pairwise loss function with the "BERT-mini" base model on the 1765-query dataset [107].

		Cut-off@10					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Penalty	Irrelevant	-7.71	-26.04	-30.29	-35.92	-3.35	0.90
	Relevant	1.09	-64.96	-64.48	-66.78	-20.14	8.81
	Both	-3.66	-38.15	-37.16	-36.82	-32.95	14.99
Reward	Irrelevant	-28.23	-79.06	-80.10	-79.23	-26.29	11.47
	Relevant	-12.57	-82.27	-86.23	-87.60	-18.70	7.6698
	Both	-2.51	-39.74	-38.37	-37.72	-23.02	10.15
		Cut-off@20					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Penalty	Irrelevant	-7.46	-22.84	-28.13	-35.21	-4.18	1.24
	Relevant	0.92	-60.69	-61.99	-66.22	-17.55	7.27
	Both	-3.40	-32.88	-31.42	-30.16	-32.51	13.17
Reward	Irrelevant	-27.11	-68.94	-70.30	-68.84	-23.98	9.28
	Relevant	-11.87	-78.53	-81.39	-81.21	-14.60	5.49
	Both	-2.41	-35.74	-35.18	-36.10	-22.94	9.01

the strongest bias reduction across all datasets. However, these penalty-based strategies occasionally lead to reductions in fairness metrics like *NFaiRR*, highlighting challenges in balancing fairness and ranking effectiveness. On the other hand, reward-based scenarios, while less effective in bias reduction, sometimes lead to marginal improvements in ranking metrics. For example, scenarios involving *rewards for irrelevant documents* show slight gains in *MRR@10* while achieving moderate bias mitigation. These findings underscore the trade-offs inherent in different mitigation strategies, with penalty-based scenarios being more reliable for bias reduction and reward-based scenarios offering potential ranking benefits in specific contexts.

Findings for RQ2. The objective of the second research question is to investigate whether the behavioral patterns observed on one language model can be generalized to other language models. For this purpose, we repeat our experiments on a second language model, namely Electra-small and report on our findings in Tables 5-8. Similar to the findings with *BERT-mini*, scenarios involving *applying penalty to relevant documents* and *applying penalty to both types of documents* consistently demonstrate strong bias mitigation across datasets and evaluation metrics. These approaches achieve substantial reductions in bias metrics such as *ARaB-TC* and *ARaB-TF*, while either maintaining or slightly improving ranking

Table 4.3: Performance of the model across the six proposed scenarios using the pointwise loss function with the "BERT-mini" base model on the 215-query dataset [105].

		Cut-off@10					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Penalty	Irrelevant	-9.51	-11.61	-10.02	-8.78	-15.03	1.97
	Relevant	-8.55	-16.69	-13.93	-8.11	-2.62	0.58
	Both	-14.88	-15.53	-17.57	-22.42	-16.37	2.00
Reward	Irrelevant	-6.65	-17.05	-21.91	-30.27	17.94	-1.71
	Relevant	-16.96	2.23	2.72	3.60	-0.46	-0.59
	Both	-14.56	-0.75	-3.53	-10.76	16.91	-2.19
		Cut-off@20					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Penalty	Irrelevant	5.73	-4.81	-5.26	-5.07	-21.35	2.07
	Relevant	-8.88	-18.36	-16.86	-13.09	-5.44	0.39
	Both	-14.96	-8.53	-10.66	-13.60	-19.71	1.91
Reward	Irrelevant	-6.48	-27.40	-30.03	-32.82	13.58	-1.88
	Relevant	-16.70	1.21	1.47	2.17	-0.99	-0.27
	Both	-15.44	-11.47	-12.40	-15.23	14.02	-2.28

Table 4.4: Performance of the model across the six proposed scenarios using the pointwise loss function with the "BERT-mini" base model on the 1765-query dataset [107].

		Cut-off@10					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Penalty	Irrelevant	0.41	-16.10	-15.40	-15.37	-11.92	2.99
	Relevant	0.96	-15.54	-11.96	-8.82	-2.13	1.07
	Both	-4.24	-21.81	-22.72	-24.01	-14.94	4.11
Reward	Irrelevant	2.16	-8.67	-11.42	-21.24	12.34	-3.53
	Relevant	-1.63	7.87	10.99	11.09	4.16	-1.49
	Both	1.24	1.54	-3.55	-18.55	15.77	-5.45
		Cut-off@20					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Penalty	Irrelevant	0.68	-13.82	-14.42	-13.81	-13.10	3.44
	Relevant	0.97	-20.96	-19.65	-18.45	0.06	0.47
	Both	-4.08	-23.35	-25.10	-25.98	-14.70	3.92
Reward	Irrelevant	2.31	-15.72	-17.47	-21.88	13.77	-3.48
	Relevant	-1.49	-0.09	1.36	1.67	3.05	-0.90
	Both	-4.08	-1.40	-5.01	-12.51	12.82	-4.56

effectiveness. For instance, *applying penalty to both types of documents* leads to significant fairness improvements, as reflected in higher *NFaiRR* scores, albeit with moderate trade-offs in ranking effectiveness (*MRR@10*). This pattern is consistent with the earlier results on *BERT-mini*, suggesting that penalty-based approaches are robust and generalizable across different pre-trained language models.

In contrast, scenarios involving rewards exhibit greater variability in their performance

Table 4.5: Performance of the model across the six proposed scenarios using the pairwise loss function with the "Electra-small" base model on the 215-query dataset [105].

		Cut-off@10					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Penalty	Irrelevant	-9.36	-64.48	-68.67	-68.55	-15.99	3.12
	Relevant	13.38	-82.91	-75.81	-66.57	-25.60	5.36
	Both	10.23	-80.87	-77.88	-75.19	-31.82	7.45
Reward	Irrelevant	-14.94	-73.27	-71.57	-67.21	-50.83	13.05
	Relevant	-18.38	-94.24	-96.56	-100.04	-18.74	4.70
	Both	-4.46	-129.84	-135.63	-141.12	-31.30	7.15
		Cut-off@20					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Penalty	Irrelevant	-9.13	-56.96	-60.91	-61.69	-16.77	2.77
	Relevant	11.50	-80.78	-75.66	-69.48	-22.07	3.96
	Both	9.47	-79.95	-78.13	-76.57	-31.29	6.12
Reward	Irrelevant	-14.38	-72.55	-70.42	-67.73	-47.21	10.58
	Relevant	-17.01	-84.55	-88.58	-92.14	-14.64	3.23
	Both	-4.41	-68.38	-64.19	-60.31	-30.12	5.71

Table 4.6: Performance of the model across the six proposed scenarios using the pairwise loss function with the "Electra-small" base model on the 1765-query dataset [107].

		Cut-off@10					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Penalty	Irrelevant	-18.13	-74.43	-84.88	-88.74	-13.16	4.15
	Relevant	3.43	-56.66	-71.59	-77.87	-17.15	6.5160
	Both	-3.90	-88.53	-95.07	-94.13	-25.80	11.69
Reward	Irrelevant	-12.44	-74.12	-68.29	-64.40	-38.52	17.63
	Relevant	-25.63	-47.89	-62.06	-67.54	-16.31	5.28
	Both	-6.49	-95.04	-87.38	-85.91	-25.13	10.93
		Cut-off@20					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Penalty	Irrelevant	-17.47	-88.43	-92.38	-91.73	-15.99	5.0068
	Relevant	3.48	-56.39	-66.27	-69.31	-15.99	5.60
	Both	-3.61	-93.16	-96.29	-92.75	-25.89	10.66
Reward	Irrelevant	-11.95	-74.77	-70.19	-67.69	-36.60	15.74
	Relevant	-24.66	-57.19	-65.10	-66.41	-12.63	4.04
	Both	-6.07	-93.86	-88.94	-88.97	-24.60	10.08

across the two language models. For example, while *applying reward to irrelevant documents* and *applying reward to both types of documents* achieve substantial reductions in bias metrics, they often show pronounced trade-offs between bias mitigation and ranking effectiveness. In some cases, *applying reward to irrelevant documents* achieves notable improvements in fairness, indicated by higher *NFaiRR*, but these gains come at the expense of reduced *MRR@10*. The sensitivity of these reward-based scenarios to the underlying encoder is more evident with

Table 4.7: Performance of the model across the six proposed scenarios using the pointwise loss function with the "Electra-small" base model on the 215-query dataset [105].

		Cut-off@10					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Penalty	Irrelevant	1.41	-31.54	-32.03	-33.11	-44.35	3.04
	Relevant	-4.35	-12.29	-12.24	-11.78	-2.37	1.32
	Both	-5.23	-12.17	-13.55	-16.18	-23.57	3.58
Reward	Irrelevant	8.65	-33.11	-38.38	-44.24	20.02	-2.14
	Relevant	-7.87	-1.89	-4.41	-5.95	0.00	-0.32
	Both	-15.55	-46.98	-49.31	-47.83	21.53	-4.01
		Cut-off@20					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Penalty	Irrelevant	1.12	-23.70	-20.48	-15.66	-40.00	2.45
	Relevant	-3.82	-8.22	-6.98	-7.15	-4.13	1.49
	Both	-5.59	-4.06	-3.29	-4.05	-22.26	3.78
Reward	Irrelevant	7.60	-25.12	-27.85	-29.96	15.82	-1.82
	Relevant	-10.39	4.47	5.63	9.13	-1.13	0.19
	Both	-14.85	-32.12	-27.08	-16.63	18.21	-2.68

Table 4.8: Performance of the model across the six proposed scenarios using the pointwise loss function with the "Electra-small" base model on the 1765-query dataset [107].

		Cut-off@10					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Penalty	Irrelevant	-1.37	-6.76	-5.97	-4.37	-13.00	2.55
	Relevant	-2.61	-10.77	-10.61	-9.81	-5.80	1.94
	Both	-3.67	-14.92	-11.96	-7.51	-18.72	5.22
Reward	Irrelevant	1.32	-19.42	-22.13	-24.88	14.51	-4.3581
	Relevant	-2.11	-11.47	-10.66	-9.86	1.34	-1.47
	Both	-11.62	-13.42	-13.64	-14.39	12.13	-4.58
		Cut-off@20					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Penalty	Irrelevant	-0.92	-5.99	-5.08	-3.30	-14.05	3.31
	Relevant	-2.76	-10.02	-9.86	-9.14	-3.60	1.2884
	Both	-3.46	-15.01	-12.20	-7.66	-18.43	5.12
Reward	Irrelevant	1.25	-18.31	-20.21	-22.16	15.97	-4.1863
	Relevant	-2.03	-11.51	-10.74	-10.05	-0.24	-0.44
	Both	-11.33	-12.82	-12.95	-13.46	9.24	-3.41

Electra-small, where the ranking performance sometimes degrades more significantly than with *BERT-mini*. Nevertheless, in specific contexts, *applying reward to irrelevant documents* demonstrates slight gains in ranking effectiveness, suggesting opportunities for optimization to better balance fairness and effectiveness.

Overall, the findings indicate that scenarios involving *applying penalty to relevant documents* and *applying penalty to both types of documents* exhibit consistent and generalizable

behavioral patterns across language models, making them robust options for fairness-aware ranking. In contrast, reward-based scenarios, such as *applying reward to irrelevant documents* and *applying reward to both types of documents*, demonstrate varying effectiveness depending on the base encoder, highlighting their sensitivity to the underlying architecture.

Findings for RQ3. The objective of RQ3 is to evaluate the comparative potential of pointwise and pairwise loss functions in serving as fair rankers by analyzing their ability to balance bias mitigation and ranking effectiveness under the proposed fairness-aware scenarios. The findings in Tables 9 and 10 highlight notable differences in the performance of the pairwise and pointwise loss functions under the proposed bias mitigation scenarios. These observations are summarized as follows: (1) The *pairwise loss function consistently outperforms the pointwise loss function* in ranking effectiveness, as measured by MRR. For both the 215-query and 1765-query datasets, pairwise models demonstrate higher MRR improvements at both cutoff levels (10 and 20), indicating their superior ability to preserve ranking quality while incorporating fairness-aware adjustments. (2) The pairwise models achieve stronger reductions in bias metrics, including *ARaB-TC*, *ARaB-TF*, *ARaB-Bool*, and *LIWC*. Across both datasets and cutoff levels, the pairwise models consistently exhibit larger decreases in these bias measures compared to pointwise models, reflecting their higher effectiveness in mitigating gender bias. (3) Improvements in the fairness metric *NFaiRR* are more pronounced for pairwise models. This indicates that the pairwise approach better promotes fairness across the rankings, achieving consistently higher *NFaiRR* scores than pointwise models on both datasets and cutoff levels. (4) The pairwise loss function demonstrates a better ability to balance fairness and ranking effectiveness. While the pointwise models achieve moderate reductions in bias metrics, these often come at the cost of ranking effectiveness, as evidenced by negative or marginal improvements in MRR. In contrast, the pairwise models successfully maintain or improve ranking effectiveness while achieving greater bias reduction, highlighting their robustness. These findings collectively indicate that the pairwise loss function is more effective in achieving fairness-aware ranking and serves as

Table 4.9: Comparison of the best-performing pairwise and pointwise approaches using the "BERT-mini" base model on the 215-query dataset [105].

			Cut-off@10			
	MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Best pairwise model	10.71	-60.62	-59.52	-60.75	-28.25	8.22
Best pointwise model	-8.54	-16.68	-13.93	-8.10	-2.62	0.57
			Cut-off@20			
	MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Best pairwise model	8.92	-51.96	-52.05	-54.35	-16.02	5.62
Best pointwise model	-8.88	-18.36	-16.85	-13.09	-5.44	0.39

Table 4.10: Comparison of the best-performing pairwise and pointwise approaches using the "BERT-mini" base model on the 1765-query dataset [107].

			Cut-off@10			
	MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Best pairwise model	1.08	-64.95	-64.47	-66.77	-20.13	8.81
Best pointwise model	0.95	-15.54	-11.95	-8.81	-2.13	1.06
			Cut-off@20			
	MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Best pairwise model	0.91	-60.68	-61.99	-66.22	-17.55	7.2717
Best pointwise model	0.97	-20.96	-19.64	-18.45	0.06	0.46

a better foundation for fair rankers compared to the pointwise loss function.

Findings for RQ4. This research question focuses on understanding the influence of the regularization coefficient (λ) on the performance of the fairness-aware framework, particularly its ability to balance bias mitigation and ranking effectiveness. We have chosen the same pairwise and pointwise models reported in Tables 9 and 10 in this research question. As shown in Figure 1, as λ increases, the fairness of the models improves consistently, as indicated by the upward trend in the *NFaiRR* metric across both datasets (1,765 and 215 queries) and both loss functions (pairwise and pointwise). Simultaneously, bias metrics such as *ARaB-TC*, *ARaB-TF*, *ARaB-Bool*, and *LIWC* exhibit substantial reductions, demonstrating the model's enhanced capability to mitigate gender biases with larger regularization coefficients. However, increasing λ introduces a clear trade-off, as reflected in the decline of ranking effectiveness measured by *MRR*. As fairness improves, *MRR* steadily decreases, highlighting the tension between bias mitigation and retrieval effectiveness. This trade-off becomes particularly evident at higher values of λ, where fairness metrics reach their peak,

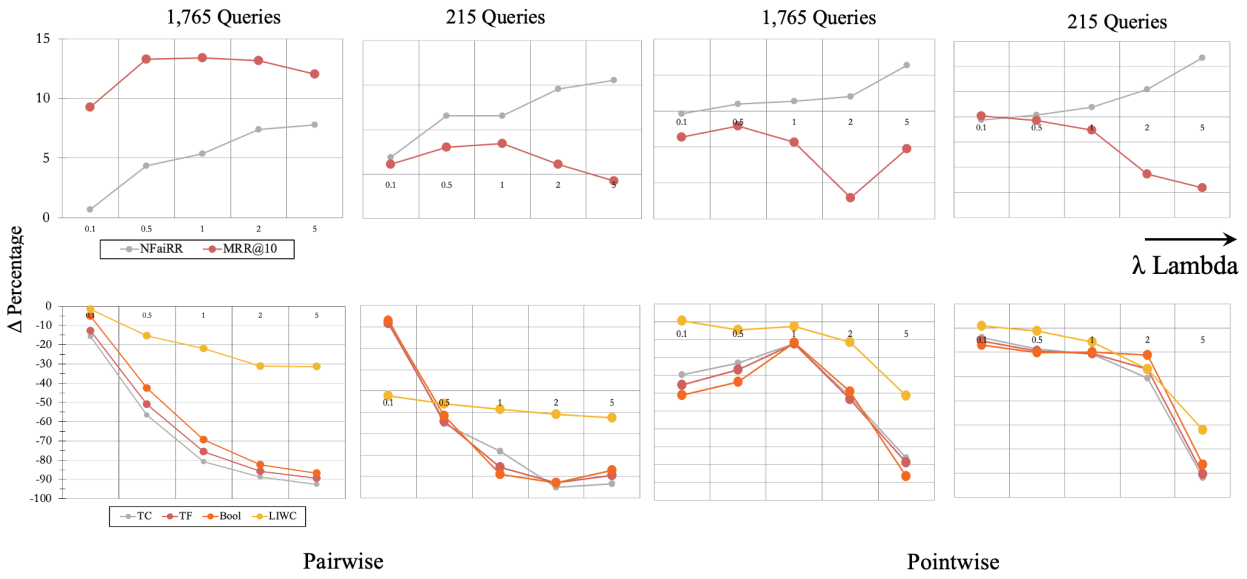

Figure 4.1: The impact of varying the value of λ on the performance of the best 'fair' pointwise and pairwise ranker using the "BERT-mini" base model. These rankers are the same as those reported in Tables 9 and 10.

but *MRR* suffers the most significant drop. These results emphasize the importance of carefully tuning λ to achieve an optimal balance that aligns with the specific goals of the application, whether prioritizing fairness, effectiveness, or a combination of both.

Findings for RQ5. This research question aims to evaluate how the proposed fairness-aware framework compares to three state-of-the-art methods: Light-Weight Sampling (LWS), AdvBERT, and CODER. We adopt the best pairwise variation chosen based on Tables 9 and 10 as representative of our proposed approach. We find that our proposed approach demonstrates superior bias mitigation across all bias metrics when compared to AdvBERT and LWS. On both the 215-query and 1,765-query datasets, the framework achieves more substantial reductions in metrics such as *ARaB-TC* and *ARaB-TF*, reflecting its effectiveness in minimizing gender biases in ranked results. Unlike AdvBERT and LWS, which exhibit inconsistent performance in bias mitigation across datasets, the proposed approach maintains robust bias reduction across all evaluated scenarios.

When compared to CODER, our proposed approach achieves competitive bias reduction while maintaining higher ranking effectiveness. Although CODER demonstrates strong bias

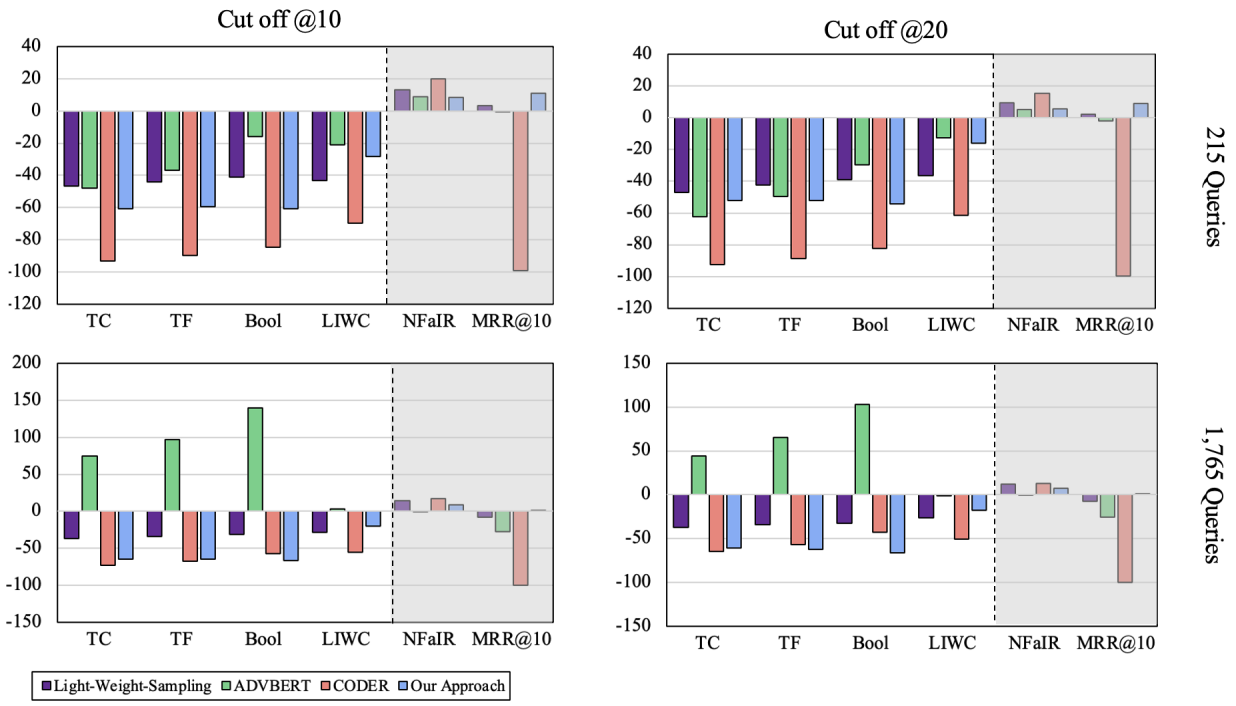

Figure 4.2: Comparison of our proposed approach with the state-of-the-art methods on the 215-query and 1,765-query datasets based on the best 'fair' pairwise ranker using the "BERT-mini" base model. This ranker is the same as the pairwise ranker reported in Tables 9 and 10. We note negative values on the left side of each figure and positive values on the right side are desirable.

reduction capabilities, it significantly compromises ranking performance, as evidenced by a marked decline in *MRR* values. This trade-off limits CODER's practicality in real-world information retrieval systems, where delivering relevant and accurate results remains a primary requirement alongside fairness. In contrast, our proposed approach effectively balances fairness and ranking quality. By integrating fairness constraints into the loss function, our approach achieves notable reductions in bias metrics while maintaining or slightly improving *MRR*. This balanced performance highlights our approach's potential for deployment in practical IR systems, where fairness and relevance are equally critical. Overall, the results affirm that our approach not only surpasses the state-of-the-art methods in bias mitigation but also ensures that fairness enhancements do not come at the cost of retrieval effectiveness.

4.3 Concluding Remarks

In this chapter, we presented a systematic approach to mitigating gender bias in dense neural rankers through fairness-aware loss function regularization. By introducing penalty and reward mechanisms into both pointwise and pairwise ranking frameworks, the proposed method effectively balances retrieval effectiveness and fairness. Comprehensive experiments on benchmark datasets demonstrate the framework's ability to reduce gender bias while maintaining or enhancing ranking performance. Comparisons with state-of-the-art fairness-aware methods further highlight the robustness and competitiveness of our proposed approach.

We believe this work opens avenues for further research. First, while the proposed approach incorporates fairness constraints into the optimization objectives of neural rankers, future research could explore *adaptive regularization strategies* that dynamically adjust the trade-off between relevance and fairness during training based on dataset characteristics and model performance. Such an approach could improve generalizability across diverse queries and datasets by tailoring fairness requirements to specific use cases without manual tuning of hyperparameters. Second, our approach could be extended to address intersectional biases involving multiple sensitive attributes (e.g., gender, race, and age) through *multi-objective optimization techniques* that simultaneously account for fairness across multiple dimensions. This would involve developing multi-task loss functions or joint debiasing mechanisms that minimize the compounded effects of biases while preserving ranking quality.

Chapter 5

De-biasing Neural Embeddings

5.1 Preliminaries

Neural rankers. Given a set of queries, denoted by $Q = \{q_1, q_2, \ldots, q_n\}$, and a corresponding pool of documents represented by $D = \{d_1, d_2, \ldots, d_m\}$, a neural ranker, Φ, employs a neural network architecture with a set of parameters θ to rank documents in D in relation to queries in Q. The neural ranker generates a ranked list R of documents by evaluating the relevance of each document to a given query. This is achieved by calculating a relevance score $s = \Phi(q, d)$ for each query-document pair (q_i, d), where $q_i \in Q$ and $d \in D$. Since neural rankers are supervised methods, during the training process, their parameters θ are optimized to improve the ranker's ability to accurately reflect the relevance of documents in relation to input queries.

Neural ranking architectures. A neural ranker Φ often consists of two components: *(i)* an encoder, and *(ii)* a scoring mechanism. The encoder, which is typically a large language model (LLM), processes the inputs to generate vector representations for queries and documents. Within a cross-encoder architecture [104], the vector representation of the query q and document d are often concatenated, which can be expressed as:

$$E = \text{encoder}(q \oplus d) \tag{5.1}$$

where $\oplus$ denotes the concatenation operator. Subsequently, a multi-layer feedforward network is employed as the scoring mechanism. It takes the vector E and computes the relevance scores used to rank documents in relation to the input query q.

Training neural rankers. A neural ranker Φ is often trained using a pairwise training process [25], which adopts a contrastive learning strategy [144]. This strategy ensures that vectors representing queries are placed closer to those of their relevant documents and placed furthest away from those of their irrelevant documents within the vector space. This objective is achieved through a *marginal ranking loss* function, as follows:

$$L = \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - \Phi(q, {d_i}^+) + \Phi(q, {d_j}^-)) \tag{5.2}$$

where d_i^+ and d_j^- denote relevant and an irrelevant document, respectively, relative to the query q. Furthermore, N^+ and N^- represent the total number of relevant and irrelevant documents, and n is the total number of training samples across all queries. This loss function helps guide the training process, enabling the neural ranker Φ to better distinguish between relevant and irrelevant documents for each query.

5.2 Overview of the Disentanglement Approach

Neural rankers order documents by the similarity between their vector representations and those of user queries. It has been empirically demonstrated [108] that these vectors often encode gender preferences, which can intensify biases. Consequently, these biases are implicitly considered during the ranking process. Therefore, our hypothesis is that by excluding gender information from the vector representations of queries and documents, the neural ranker will be unable to access, even implicitly, any gender data during the ranking process, thereby preventing the intensification of biases.

To this end, we propose to *'disentangle'* query and document vector representations into

discernible components dedicate to gender and semantic sub-vectors. The gender sub-vector would be responsible for capturing possible gender information in the query or document, whereas the semantic component would, independently of the gender information, only represent the content value of the query or document. Disentangled representation learning [5] aims to create factorized representations that isolate underlying factors of input data. In our work, we focus on separating content gender from content semantics. We propose that separating gender-related information from E can debias neural ranking outputs. Thus, we disentangle E into two components: E_r, containing all content semantics and non-gender factors, and E_g, representing gender-related information. Let Γ be a function that disentangles a vector E of size d into two components of sizes m and n such that:

$$E_r, E_g = \Gamma(E, m, n), \quad d = m + n \tag{5.3}$$

Based on Equation 5.3, we propose using E_r for ranking while excluding E_g, potentially reducing the gender biases inherent in neural rankers. By omitting E_g from the ranking process, we suggest that gender will no longer influence query performance or the makeup of the ranked results.

5.3 Neural Architecture for Gender Disentanglement

Figure 5.1 shows our architecture for disentangling gender from content semantics in neural rankers. It includes two distinct networks: the *Ranking Network* and the *Gender Network*. The Ranking Network processes only the semantic subvector, E_r, learning the relevance between queries and documents. The Gender Network refines the gender-specific subvector, E_g, to predict gender attributes accurately. When trained together, these networks separate content into E_r and gender information into E_g, effectively disentangling the two.

The Ranking Network. This network is designed to capture and learn the concept of relevance between queries and documents. It operates by only processing the semantic

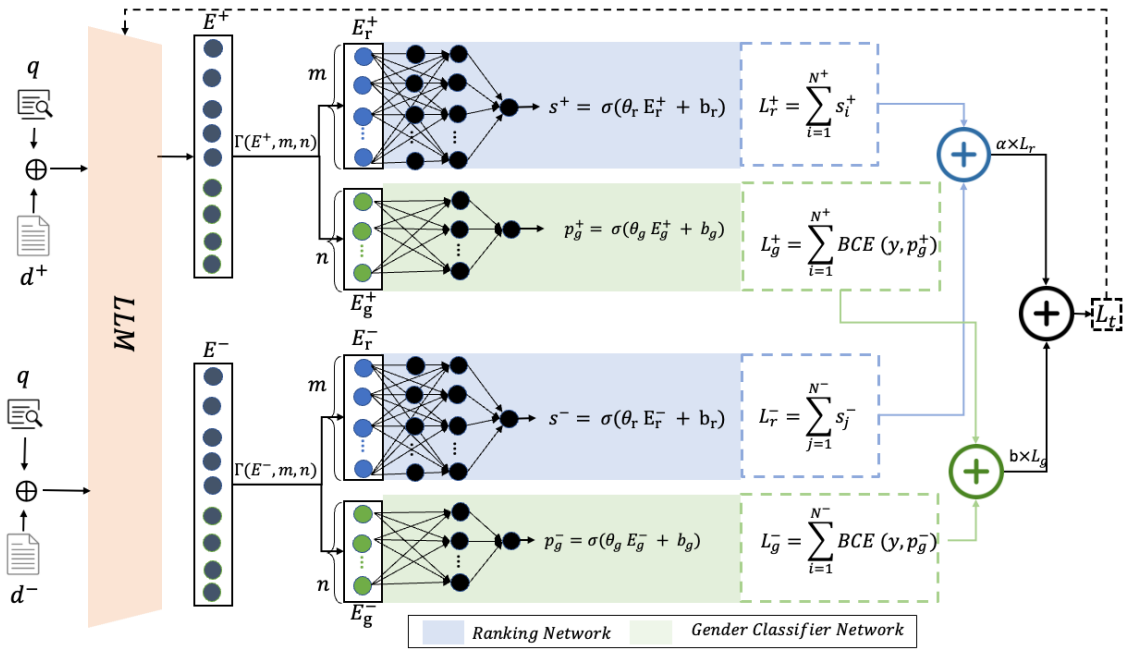

Figure 5.1: Overview of the proposed neural disentanglement architecture.

component of the original vector, E_r, using a Multi-Layer Perceptron (MLP), denoted as MLP_r, to predict relevance scores for query-document pairs. During the training process, this network generates relevance scores for both relevant and irrelevant documents associated with each query in the dataset, as defined in the following:

$$s = \sigma(\theta_r E_r + b_r), \tag{5.4}$$

where σ is the activation function, and $\Theta = \theta_r \cup b_r$ are the parameters of the MLP responsible for predicting the relevance score s. To enhance the model's discrimination capabilities, we embed a contrastive loss function that aims to increase the relevance scores for matches between queries and their corresponding relevant documents while reducing the scores for mismatches with irrelevant documents. The loss function can be formulated as follows:

$\mathrm{L}^+ = \sum_{i=1}^{N^+} \sigma(\theta_r E_{r_i}^+ + b_r), \quad L^- = \sum_{j=1}^{N^-} \sigma(\theta_r E_{r_j}^- + b_r)(5.5)$ Therefore,

$$L_r = \frac{1}{n} \sum \max(0, m - L^+ + L^-) \tag{5.6}$$

Given L_r, θ_r and b_r are updated as follows:

$$\theta_r^{(t+1)} = \theta_r^{(t)} - \eta \frac{\partial L_r}{\partial \theta_r}, \quad b_r^{(t+1)} = b_r^{(t)} - \eta \frac{\partial L_r}{\partial b_r} \tag{5.7}$$

where η is the learning rate, and t denotes the iteration number. This approach allows the network to accurately identify and enhance the relevance of query-document pairs, thus optimizing the performance of the neural ranker.

The Gender Network. This network specifically targets the gender-specific subvector of the neural ranker's intermediate vector, E_g, to predict gender attributes. The network utilizes a Multilayer Perceptron (MLP), denoted as MLP_g, which processes E_g to estimate the probability of document or query gender affiliation as follows:

$$p_g = \sigma(\theta_g E_g + b_g), \tag{5.8}$$

where σ is the activation function, θ_g are the weights, and b_g the bias of the MLP_g.

To obtain accurate gender affiliations, a function Λ is assumed, which can identify the gender affiliation of a text t:

$$g = \Lambda(t) \tag{5.9}$$

Here, g represents the gender affiliation of the text, t. The training of the Gender Network is governed by a Binary Cross Entropy loss function, formulated as:

$$L_g = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(p_{g_i}) + (1 - y_i) \log(1 - p_{g_i})] \tag{5.10}$$

where N is the total number of instances, p_{g_i} is the predicted probability that the i-th

Algorithm 3 Training of the Disentangled Ranking Network

Data: $\{(q, d^+, d^-)\}$, number of training iterations T.
Initialize: θ_r, θ_g, b_r, b_g randomly.
$g^+ \leftarrow \Lambda(q \oplus d^+)$
$g^- \leftarrow \Lambda(q \oplus d^-)$ **for** $t = 1$ **to** T **do**
    each sample (q, d^+, d^-, g^+, g^-) in the batch
$E^+ \leftarrow \text{encoder}(q \oplus d^+)$
$E^- \leftarrow \text{encoder}(q \oplus d^-)$
$E_r^+, E_g^+ \leftarrow \Gamma(E^+, m, n)$
$E_r^-, E_g^- \leftarrow \Gamma(E^-, m, n)$
$s^+ \leftarrow \sigma(\theta_r E_r^+ + b_r)$
$s^- \leftarrow \sigma(\theta_r E_r^- + b_r)$
$p_g^+ \leftarrow \sigma(\theta_g E_g^+ + b_g)$
$p_g^- \leftarrow \sigma(\theta_g E_g^- + b_g)$
$L_r \leftarrow \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - s^+ + s^-)$
$L_g \leftarrow -\frac{1}{N} \sum_{i=1}^{N} [g_i \log(p_{g_i}) + (1 - g_i) \log(1 - p_{g_i})]$
$L_t \leftarrow \alpha \times L_r + \beta \times L_g$
$\theta^{(t+1)} = \theta^{(t)} - \eta \frac{\partial L_t}{\partial \theta}, \quad b^{(t+1)} = b^{(t)} - \eta \frac{\partial L_t}{\partial b}$

instance belongs to a particular gender, and $y_i = \Lambda(q_i \oplus d_i)$ is the true label derived from the function Λ.

Simultaneous training of the ranking and gender networks disentangles gender and semantic information. This dual-training strategy divides the encoder's output, E, into two components: E_r for semantics and E_g for gender information. Consequently, the architecture assesses relevance using E_r and remains unbiased by gender influences from E_g. The total loss function L_t is a linear combination of the ranking loss L_r and the gender classification loss L_g:

$$\begin{aligned} L_t =& \alpha \times \left(\frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - \sigma(\theta_r E_{r_i}^+ + b_r) + \sigma(\theta_r E_{r_j}^- + b_r)) \right) \\ &+ \beta \times \left(-\frac{1}{N} \sum_{i=1}^{N} [y_i \log(p_{g_i}) + (1 - y_i) \log(1 - p_{g_i})] \right) \end{aligned} \tag{5.11}$$

The final formulation of L_t clearly illustrates the model's balance between optimizing ranking performance and promoting gender fairness. Adjustable weights α and β enable fine-tuning to meet specific performance and fairness objectives.

5.4 Model Training

To effectively train the model, we form a dataset comprising of samples $(q, d^+, d^-, gender^+, gender^-)$, where $gender^+$, and $gender^-$ denote the gender affiliation of the relevant, and irrelevant documents, respectively. The training procedure is illustrated in Algorithm 3, which begins by initializing the dataset with query-document pairs (q, d^+, d^-) and settings the number of training iterations T. Initially, the network parameters θ_r, θ_g, b_r, and b_g, are initialized randomly. Subsequently, true gender affiliations are computed for both relevant and irrelevant documents by applying function Λ to the concatenated pairs, resulting in g^+ and g^- as described in Lines 3 and 4 of the algorithm. The primary training loop executes over T iterations. Each iteration processes a batch of samples, which includes both query-document pairs and their corresponding gender affiliations. In each iteration the following steps are taken: **(Step 1)** The query-document pairs are transformed into vector representations E^+ and E^- (Lines 7 and 8). **(Step 2)** The representations are then split into components dedicated to ranking (E_r^+ and E_r^-) and gender (E_g^+ and E_g^-) (Lines 9 and 10). **(Step 3)** Relevance scores are computed from the ranking components (Lines 11 and 12), while gender affiliations are estimated from the gender components (Lines 13 and 14). **(Step 4)** The algorithm computes the ranking loss L_r using a hinge loss in Line 15. The gender loss L_g is calculated using binary cross-entropy (Line 16), and **(Step 5)** The total loss L_t is computed as a linear interpolation of both losses, as shown on Line 17. Finally, the parameters θ_r, θ_g, b_r, and b_g are updated on Line 18 by minimizing the total loss.

5.5 Adversarial Strategy

In this appendix, we explore the question: "What if we further penalize the presence of gender information in the ranking component of the representation to ensure it is entirely gender-neutral?" To address this, we employ an adversarial strategy. We introduce an adversary network specifically designed to detect gender in the ranking representation and

aimed to alter the representation to remove any gender-related information, rendering the adversary network incapable of detecting gender from the ranking part.

To this end, we trained a gender classifier network with parameters Θ_c, which takes the ranking representation (E_r) as input and attempts to classify the gender into two categories: male or female. This process is formalized as follows:

$$p_c = \sigma(\theta_c E_r + b_c), \tag{5.12}$$

where σ is the activation function, and p_c represents the predicted probability that the ranking representation is male. We use a binary cross-entropy loss for training, defined as:

$$L_c = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_{c_i}) + (1 - y_i)\log(1 - p_{c_i})] \tag{5.13}$$

where y_i is the true gender label obtained from the function Λ.

To ensure the ranking representation (E_r) does not contain gender information, we add an adversary loss L_{adv} to the total network loss L_t. This adversary loss maximizes the entropy of the predicted gender probability p_c, making gender information unpredictable:

$$\begin{aligned} L_a dv(\theta_E) &= \mathcal{H}(p_c|E_r;\theta_c), \\ \mathcal{H}(p) &= -\sum_{i \in lables} p_i \log(p_i) \end{aligned} \tag{5.14}$$

By maximizing the entropy of p_c, we modify the ranking representation during training to exclude gender information, making it challenging for the adversary network to predict gender. The total loss L_t is defined as an interpolation of three losses: 1) the ranking loss

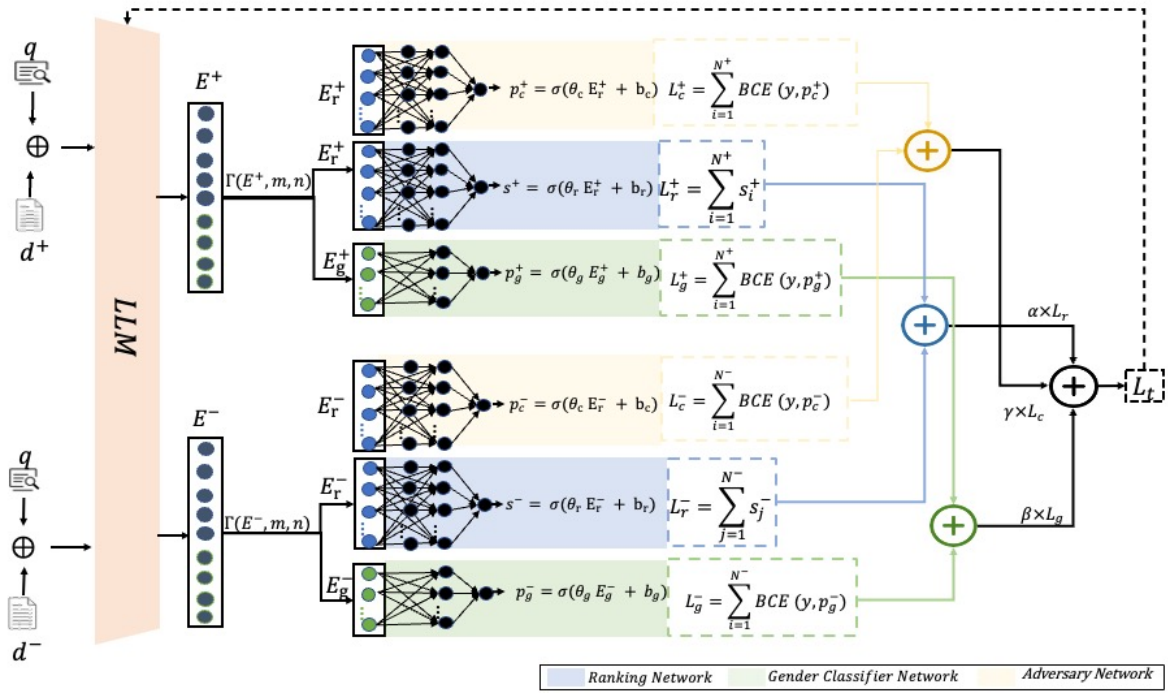

Figure 5.2: Overview of the proposed neural disentanglement architecture with the adversary network.

L_r, 2) the gender classification loss L_g, and 3) the adversary loss L_{adv}:

$$
\begin{aligned}
L_t =& \alpha \times \left(\frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - \sigma(\theta_r E_{r_i}^+ + b_r) + \sigma(\theta_r E_{r_j}^- + b_r)) \right) \\
&+ \beta \times \left(-\frac{1}{N} \sum_{i=1}^{N} [y_i \log(p_{g_i}) + (1 - y_i) \log(1 - p_{g_i})] \right) \\
&- \gamma \times \frac{1}{N} \sum_{i=1}^{N} p_{c_i} \log(p_{c_i})
\end{aligned} \tag{5.15}
$$

The network architecture, including the adversary network, is illustrated in Figure 5.2. During training, we first optimize the adversary network parameters (Θ_c). In this stage, only the adversary network parameters are updated, and the encoder parameters remain unchanged. Then, while optimizing the total loss (L_t), the parameters of the encoder, ranking network, and gender classifier network are updated.

5.6 Experiments

5.6.1 Datasets and Setup

To train our neural rankers , we utilize the MS MARCO passage ranking dataset [90], which contains approximately 200,000 queries and 8.8 million passages. For training purposes, we select 200,000 random samples of query triples $(query, doc^+, doc^-)$ to teach the model effectively. The models are trained for five epochs using the Adam optimizer and a sigmoid activation function. Additional implementation details, as well as the code, are available in our publicly available GitHub repository[1].

To evaluate model performance and measure the proposed model's effectiveness in reducing gender biases we require two sets of queries:

Gender-neutral queries (Q_n): This set allows us to evaluate stereotypical gender biases for gender-neutral queries. In particular to test the condition set out in Equation 3.5. When a gender-neutral query is fed to the model, it is expected that the ranked list does not show any inclination towards male, or female. We use two query sets proposed by Rekabsaz et al. The primary set [105] comprises 1,765 gender-neutral queries, annotated from a pool of 55,578 MS MARCO queries by three Amazon Mechanical Turk workers. The annotators flagged queries with words or phrases related to gendered concepts. The second set, contains 215 socially problematic queries that could potentially reinforce existing gender norms and propagate gender inequality if the search results are biased.

Gender-specific queries (Q_g): This set is employed to evaluate the fairness for the gender-specific queries. In particular this query set is used to evaluate the condition set out in Equation 3.6. We use the dataset labeled by Bigdeli et al.[19]. Their work involved training a BERT classifier with human-annotated queries, which they applied to the MS MARCO passage ranking development set. This labeling effort resulted in two sets: 1,405 male affiliated queries and 1,405 female affiliated queries. We evaluate the model performance on

[1] `https://github.com/genderdisen/genderdisen`

Table 5.1: Gender bias measures for 215 neutral queries with MiniLM base model.

Cut-off 10	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.1602	0.3183	0.1374	0.1101	0.8107	1.023
AdvBert	0.0093	0.0304	0.0022	0.0008	0.9854	0.1134
Bias-aware Penalty	0.167	0.1994	0.0867	0.0689	0.8453	0.8545
CODER	0.0152	0.0254	0.0240	0.0313	0.9336	0.4114
Ours	0.1877	0.0737	0.046	0.0567	0.8664	0.8404
Cut-off 20	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.1658	0.2635	0.1142	0.092	0.8274	0.7966
AdvBert	0.0103	0.0289	0.0028	0.0016	0.9824	0.1101
Bias-aware Penalty	0.1722	0.1674	0.0725	0.0576	0.8564	0.6426
CODER	0.0155	0.0351	0.0263	0.0311	0.9348	0.3491
Ours	0.1941	0.0574	0.035	0.0422	0.8722	0.717

both male affiliated, and female affiliated queries in order to asses whether the model shows comparable performance over different genders.

5.6.2 Baselines and Metrics

To evaluate our work against robust state-of-the-art baselines, we use five distinct methods: 1) **Original Model**: A cross-encoder model trained for the passage re-ranking task, using the OpenMatch implementation [78]. 2) **AdvBert**[106][2]: Utilizes an adversarial strategy to eliminate gender data from the neural rankers' intermediate representations, replicated from their GitHub repository. 3) **Bias-aware Penalty**[116]: Incorporates a direct bias penalty in the neural ranker's loss function to explicitly address gender biases during training. 4) **CODER**[137][2]: A transformer-based framework that evaluates document relevance collectively rather than individually and includes neutrality regularization to penalize deviations from gender neutrality. 5) **Light-weight Sampling**[14]: Employs a negative sampling strategy that selects the most biased documents as negative samples to train the model, teaching it to recognize and reduce bias.

To evaluate model performance, we assess ranking effectiveness and the degree of gender biases: i) **Ranking Effectiveness.** We use Mean Reciprocal Rank (MRR) to gauge the

[2] The numbers reported in the table represent the best results obtained from their implementation. We verified these results with the authors through multiple meetings, during which they confirmed the accuracy of the very low MRR values.

Table 5.2: Gender bias measures for 1765 neutral queries with MiniLM base model.

cut-off 10	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.2673	0.1535	0.0721	0.0611	0.7066	1.5599
AdvBert	0.0081	0.0051	0.0018	0.000267	0.9657	0.2374
Bias-aware Penalty	0.2814	0.0506	0.0256	0.0218	0.7396	1.4368
CODER	0.0021	0.1507	0.0721	0.0663	0.8404	0.7199
Our Approach	0.2969	0.0805	0.0131	0.0178	0.7623	1.4521
cut-off 20	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR ↑	LIWC↓
Original Model	0.2726	0.0721	0.0641	0.0538	0.722	1.3001
AdvBert	0.0099	0.0025	0.0003	0.0014	0.9642	0.2283
Bias-aware Penalty	0.2868	0.0256	0.0192	0.0152	0.7527	1.1809
CODER	0.0025	0.1490	0.0716	0.0663	0.8407	0.6467
Our Approach	0.3023	0.0131	0.0313	0.0052	0.7658	1.2767

baseline models' performance. MRR calculates the average of reciprocal ranks for all queries, focusing on the rank of the first relevant result, with MRR@10 being the standard metric for the MS MARCO passage ranking task [89]. ii) **Measuring Gender Biases.** We use three metrics to quantitatively assess each model's gender biases: a) **Average Rank Bias (ARaB)** [107] measures the presence of gender-specific words in documents, using Term Frequency (TF) and Boolean methods to calculate gendered terms. b) **NFaiRR Metric** [105] evaluates fairness at the document level within ranked lists and across all queries, based on the concept of 'document neutrality', where a higher NFaiRR indicates a fairer ranking. c) **Linguistic Inquiry and Word Count (LIWC)** [95] is employed to determine the gender affiliation of text using the social referents category, specifically the male and female reference subcategories, as outlined in [19].

5.6.3 Ranking Effectiveness and Bias Mitigation Evaluation

In our experiments, we evaluate the effectiveness of our proposed disentanglement approach in reducing stereotypical gender biases in neural rankers. We conduct experiments using two sets of gender-neutral queries, comprised of 215 and $1,765$ queries introduced in Section 6.3, respectively. To demonstrate the generalizability of our approach, we report the results based on the MiniLM [128] language model in Tables 5.1 and 5.2, and BERT-Mini language model [7] in Tables 5.3 and 5.4. Figure 5.3 shows the training loss, and MRR on the

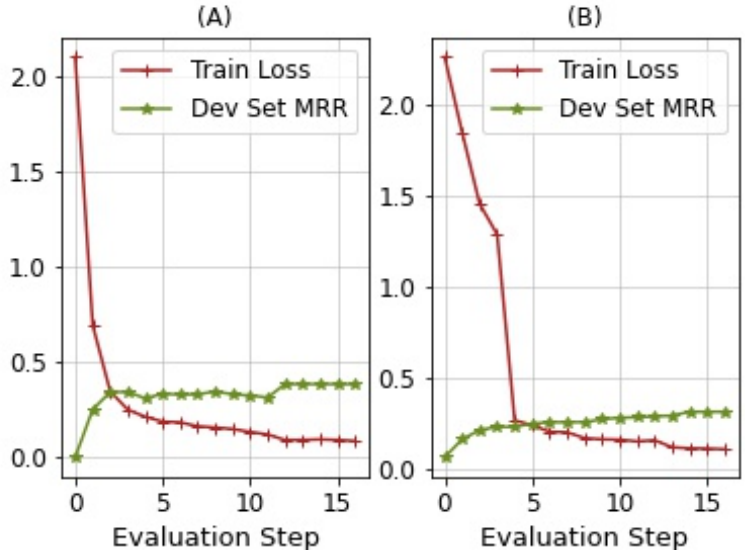

Figure 5.3: The train loss, and MRR of the developement set queries for (A) MiniLM base model, and (B) BERT-Mini base model.

development set queries for both of the base models. We can infer from the figure that as training goes on, the training losses are decreased, while the MRR of the development sets is increases, which shows that the model is trained properly, and it is not overfitted on the training data. As shown in Table 5.1, our model significantly outperforms the original model in the 215 query set, achieving a higher MRR of 0.1877 at Cut-off 10 compared to the original's 0.1602. Our model also shows considerable reductions in ARaB metrics: ARaB-tc decreases from 0.3183 to 0.0737, ARaB-tf from 0.1374 to 0.046, and ARaB-bool from 0.1101 to 0.0567, with the NFaIR score improving from 0.8107 to 0.8664. In contrast, the CODER model, while achieving lower ARaB values, only reaches an MRR of 0.0152, and the AdvBert model, despite lower ARaB values, significantly compromises retrieval effectiveness with an MRR of only 0.0093 at Cut-off 10. This table demonstrates that our approach not only reduces bias but also enhances ranking effectiveness. Furthermore, at Cut-off 20, our model continues to show improvement, with an MRR of 0.1941 compared to the original model at 0.1658. The ARaB-tc value further decreased to 0.0574, and ARaB-tf to 0.035. The NFaIR score increased to 0.8722, again indicating reduced bias. Despite AdvBert's superior bias reduction, it's MRR remained noticeably low at 0.0103, reinforcing the trade-off between

Table 5.3: Gender bias measures for 215 neutral queries with BERT-Mini base model.

cut-off 10	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.1662	0.2544	0.1058	0.0751	0.8273	0.8467
Advbert	0.0431	0.0308	0.0159	0.0155	0.9644	0.2943
Bias-aware Penalty	0.1714	0.2472	0.1025	0.0756	0.8389	0.8217
CODER	0.0014	0.0260	0.0171	0.0205	0.9649	0.2998
Our Approach	0.1399	0.0376	0.0132	0.0075	0.8583	0.6969
cut-off 20	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR	LIWC↓
Original Model	0.1742	0.2318	0.0929	0.0646	0.8457	0.6964
Advbert	0.0487	0.0289	0.0151	0.015	0.9657	0.2474
EDBT	0.181	0.2331	0.0928	0.0662	0.8563	0.6448
CODER	0.0014	0.0228	0.0148	0.0178	0.9650	0.2828
Our Approach	0.1455	0.047	0.0158	0.0083	0.8691	0.5674

Table 5.4: Gender bias measures for 1765 neutral queries with BERT-Mini base model.

cut-off 10	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.2475	0.1387	0.056	0.0369	0.7304	1.4942
AdvBert	0.0081	0.0051	0.0018	0.0003	0.9657	0.4403
Bias-aware Penalty	0.244	0.1374	0.0536	0.0334	0.7384	1.4474
CODER	0.7082e-4	0.0646	0.0371	0.0421	0.9093	0.5713
Our Approach	0.1922	0.0928	0.0354	0.026	0.7565	1.3468
cut-off 20	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.2548	0.1262	0.0.505	0.0329	0.7451	1.2592
Advbert	0.0099	0.0025	0.0003	0.0014	0.9642	0.4043
Bias-aware Penalty	0.2505	0.1138	0.0441	0.027	0.7583	1.1984
CODER	0.0001	0.0674	0.0388	0.0440	0.9096	0.4858
Our Approach	0.1996	0.0928	0.037	0.0285	0.7672	1.1682

bias reduction and retrieval effectiveness in their work.

In the 1765 query set, as shown in Table 5.2, our model achieves a superior MRR of 0.2969 at Cut-off 10, outperforming the original model's 0.2673, illustrating our approach's effectiveness in enhancing retrieval while reducing biases. Notable improvements in bias metrics include ARaB-tc decreasing from 0.1535 to 0.0805, ARaB-tf from 0.0721 to 0.0131, and ARaB-bool from 0.0611 to 0.0178, with the NFaIR score rising from 0.7066 to 0.7623. The CODER model records lower ARaB but a significantly reduced MRR of 0.0021, while the AdvBert model, despite achieving low bias scores, suffers in performance with an MRR of only 0.0081 at Cut-off 10. At Cut-off 20, our model maintains its performance with an MRR of 0.3023, further reducing ARaB-tc to 0.0131 and ARaB-tf to 0.0313, with an increased NFaIR score of 0.7658, highlighting continued bias reduction. AdvBert's low bias metrics come with a trade-off in retrieval effectiveness, indicated by an MRR of just 0.0099.

Table 5.5: Bias measures for the light weight(LW) random samples proposed in [16] on 215 queries.

cut-off 10	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.1602	0.3183	0.1374	0.1101	0.8107	1.023
Original LW	0.1604	0.0269	0.0147	0.0141	0.9596	0.2913
Disentangled LW	0.1466	0.0164	0.0103	0.0133	0.983	0.1292
cut-off 20	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.1658	0.2635	0.1142	0.092	0.8274	0.7966
Original LW	0.1659	0.0218	0.0123	0.0125	0.9595	0.3004
Disentangled LW	0.1554	0.0137	0.0076	0.0086	0.9788	0.1749

Table 5.6: Bias measures for the light weight (LW) random samples proposed in [16] on 1765 queries.

cut-off 10	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.2673	0.1535	0.0721	0.0611	0.7066	1.5599
Original LW	0.2737	0.027	0.0136	0.0124	0.8810	0.8192
Disentangled LW	0.2393	0.0157	0.0032	0.0028	0.915	0.5915
cut-off 20	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.2726	0.0721	0.0641	0.0538	0.722	1.3001
Original LW	0.2795	0.0214	0.0109	0.0101	0.8802	0.7268
Disentangled LW	0.2464	0.0177	0.0049	0.0003	0.9091	0.5621

It's worth noting that while the AdvBert model greatly reduces biases, it suffers a significant performance drop, limiting its practical use. The Bias-aware Penalty baseline offers moderate bias reduction with good performance, yet our model exceeds it in both bias reduction and ranking effectiveness. Similarly, the CODER baseline significantly reduces bias but has markedly lower ranking performance compared to our approach.

Using the BERT-Mini model, shown in Tables 5.3 and 5.4, similar trends are observed. In the 215 query set, our model achieves an MRR of 0.1399 at Cut-off 10, compared to the original's 0.1662. Despite a slight drop in ranking effectiveness, our model significantly mitigates bias, with ARaB-tc dropping from 0.2544 to 0.0376, ARaB-tf from 0.1058 to 0.0132, and ARaB-bool from 0.0751 to 0.0075. The NFaIR score improved from 0.8273 to 0.8583. The AdvBert model, though achieving lower ARaB values, suffers from drastically reduced performance, with an MRR as low as 0.0431 at Cut-off 10. At Cut-off 20, our model continues to improve, reducing ARaB-tc to 0.047 and ARaB-tf to 0.0158, and increasing the NFaIR score to 0.8691. However, AdvBert's performance remains low with an MRR of 0.0487,

underscoring a substantial trade-off between bias reduction and retrieval effectiveness. In the 1765 query set, our model significantly improves bias metrics with ARaB-tc decreasing from 0.1387 to 0.0928, ARaB-tf to 0.0354, and ARaB-bool to 0.026. The NFaIR score rose from 0.7304 to 0.7565. Despite AdvBert achieving the lowest bias scores, its retrieval effectiveness is compromised, showing an MRR of only 0.0081 at Cut-off 10. At Cut-off 20, further reductions in bias metrics and an increase in NFaIR to 0.7672 continue, yet AdvBert's MRR remains low at 0.0099, highlighting its limited practical utility.

We point out that while the AdvBert model significantly reduces biases across all metrics, it does so with a marked decline in retrieval effectiveness. The Bias-aware Penalty baseline shows a moderate reduction in bias with relatively good performance. However, our model outperforms it in both bias reduction and retrieval effectiveness. In addition, we have incorporated an adversary network into our architecture to intensify the penalization of gender information within the ranking representation. Detailed explanations, results, and discussions are provided there. Additionally, a case study example is presented to further illustrate the effectiveness of our proposed model in mitigating stereotypical gender biases.

5.6.4 Light-Weight Sampling Strategy

As an additional baseline, Bigdeli et. al. [16] have suggested a negative sampling strategy, in which the negative documents are selected such that they exhibit large amount of bias. By doing so, the model will implicitly recognize bias as a negative factor during the training; therefore, biased documents will be sorted lower when re-ranked with the trained model. We adopt this negative sampling strategy, and select two negative samples from the proposed dataset, and train our gender disentanglement model with this bias-aware negative strategy. Given limited space, we report the results on the MiniLM language model in Tables 5.5, and 5.6 for the 215, and 1,765 queries.

When comparing the results of the original and disentangled models in both tables before and after the light-weight negative sampling strategy is used, we make three consistent

Table 5.7: Performance on gender-specific queries.

	MRR@10		
	Male	Female	Δ
Original Model	0.3939 (std=0.3528)	0.3178 (std=0.3445)	19.3196
Disentangled Model	0.4093 (std=0.3485)	0.3511(std=0.3474)	14.2194
Improvement	3.90%	10.47%	-26.39%
	MRR@20		
Original Model	0.4002 (std=0.3682)	0.3242 (std=0.3582)	0.1899
Disentangled Model	0.4146 (std=0.3597)	0.3572 (std=0.3610)	0.1384
Improvement	3.59%	10.17%	-27.11%

improvements: (1) the negative sampling strategy does not lead to a drop in retrieval effectiveness on the base retrieval method but decrease in our disentanglement method is more pronounced on the MRR metric. This shows that When shown severely biased negative samples, the proposed disentanglement model cannot learn the concept of relevance as well as when random negative samples were selected; (2) on the other hand, the negative sampling strategy leads to notable reduction in bias in our proposed approach, which is superior to both the original base model as well as when negative sampling strategy was applied to the base retrieval method. This suggests, as also reported by Bigdeli et al [16] that the selection of the negative samples can lead to reduced bias. In summary, while the disentangled model consistently reduces bias metrics (ARaB-tc, ARaB-tf, ARaB-bool) and improves NFaIR and LIWC scores compared to the original model, this often comes at the cost of a slight decrease in MRR, which may be tolerable depending on the application area and the significance of the observed bias reduction.

5.6.5 Performance Disparities

Besides stereotypical gender biases, disparities in retrieval effectiveness between male and female-affiliated queries are notable. As shown in the top row of Table 5.7, the original neural ranker performs significantly better on male queries than on female queries, with a 19% higher effectiveness at cut-off 10 and a similar disparity at cut-off 20. However, the results from our disentanglement approach, detailed in the second row of Table 5.7, offer the following

observations: **(1)** Our approach reduces the performance disparity between male and female affiliated queries from 19% to 14%, a 5% improvement. This significant progress suggests the importance of addressing not only the stereotypical gender biases in document retrieval but also the disparities in retrieval effectiveness across different gendered queries. Otherwise, merely reducing gender biases without improving retrieval effectiveness could result in less biased but potentially irrelevant documents being retrieved. **(2)** Our approach reduces the performance disparity between male and female queries without sacrificing the performance of either group. Contrary to concerns raised in earlier studies [118] that reducing gender biases might decrease retrieval effectiveness, our method actually enhances performance for both groups. Specifically, male-affiliated queries experience more than a 3.5% improvement, and female-affiliated queries improve by over 10%. We have also reported the standard deviation of the reciprocal ranks of the male, and female queries. The consistent standard deviation for both the original, and the disentangled models in cut-offs 10, and 20 implies that the system has become better at ranking relevant results higher on average (as indicated by the increased MRR), but the degree to which these rankings vary across different queries is the same as before. This could mean that the improvement in MRR is consistent across many queries.

5.6.6 Gender Disentanglement Quality

Evaluating gender bias in neural embeddings is challenging due to the absence of standardized measures, particularly for complex contextualized embeddings [139, 4]. Previous research has utilized clustering and classification to identify gender information in embeddings, comparing debiased and non-debiased versions [55, 21]. In our evaluation, we test the efficacy of our approach to separate gender from semantics using three established strategies: **(1)** analyzing occupational stereotypes [4, 139] and **(2)** detecting gender spaces in embeddings [21] and **(3)** Measuring Bias in Sentence Encoders.

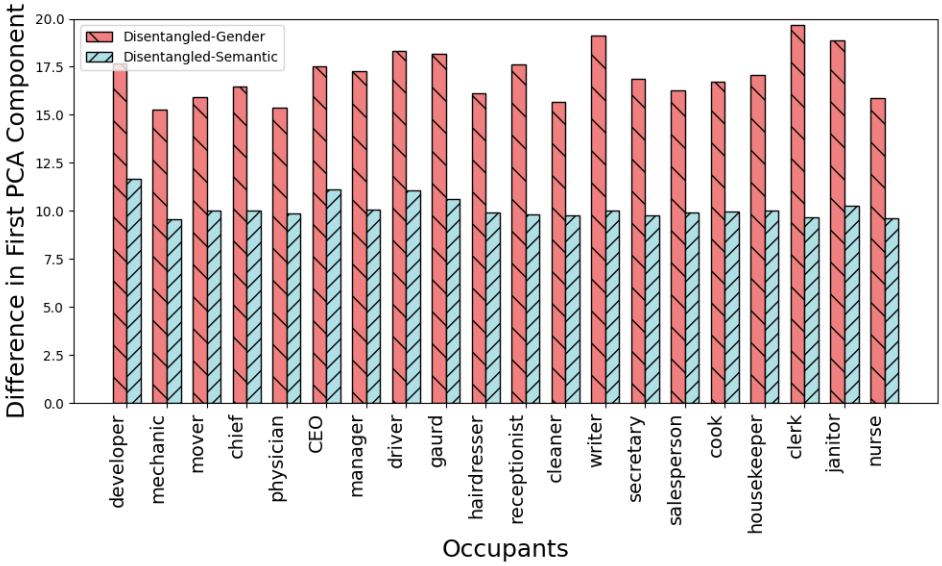

Figure 5.4: PCA of stereotyped occupations by pronouns, using gender and semantic disentanglement.

Occupational Stereotypes

Previous research has shown that neural embeddings often capture biases related to gender roles and professions, such as stereotypically associating engineering with men and nursing with women. In this section, we explore how gender and semantic components within these embeddings affect the representation of occupations deeply linked to gender stereotypes. We are interested in determining if our disentanglement approach effectively separates gender from the semantic aspects of occupations. To achieve this, we use the method suggested in [4, 139], selecting sentence pairs from the WinoBias dataset [141]. Each pair features the same sentence with different gender-specific pronouns, for example, "[The manager] fired the cleaner because [he] was angry" and its counterpart "[The manager] fired the cleaner because [she] was angry". In Figure 5.4, we illustrate the 20 occupations from [139]. We applied PCA to the disentangled semantic component (in blue) and the gender component (in red) of our model. We then analyzed the difference between the first principal component of the female representation and the male representation within the same sentence. Given that we are utilizing contextualized embeddings, the embedding of the pronoun token will

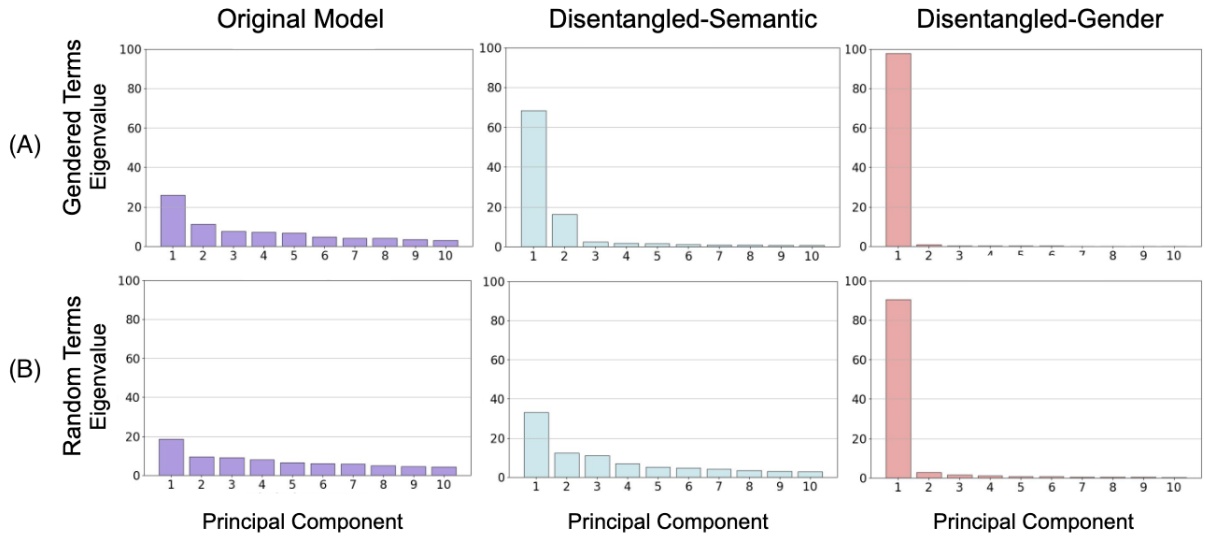

Figure 5.5: (A) Variance percentages in the principal components for the original, disentangled semantics, and disentangled gender models. (B) Corresponding percentages for random vectors.

vary depending on the context. As depicted in Figure 4, the semantic component shows a smaller difference between PCA components across the same occupations when represented in the same sentence with different gender pronouns. This indicates a reduced dependency on gendered pronouns within the semantic component. In other words, the difference between the first PCA components in sentences with gendered pronouns is smaller when analyzing the semantic component than when analyzing the gender component. In contrast, the gender component displays greater variation, suggesting that it more effectively captures the gender associations tied to different occupations. This provides a clearer distinction between the representations of gendered pronouns across different occupations. This contrast demonstrates that the semantic component exhibits a more uniform representation of pronouns across occupations, meaning that pronoun differences are less influenced by occupation titles. Conversely, the gender component highlights these differences, indicating that it successfully isolates gender characteristics within the gender component, as intended by our disentanglement approach.

Detecting Gender Spaces

Bolukbasi et al. [21] introduced a method to identify gender biases in embeddings by defining a gender space through directional differences between gender-related word pair vectors. Adopting this method, we analyzed the Principal Component Analysis (PCA) of male and female word pairs' vector differences, as illustrated in Figure 5.5. This analysis extends to both the original and two disentangled models—one emphasizing semantic aspects, the other on gender components—providing insights into the inherent gender biases. We initiated our experiments by calculating directional vectors for pairs like 'he-she' and 'man-woman', as recommended by [55], and applied PCA to these vectors to detect and quantify gender biases. Notably, our evaluation includes contextualized embeddings, allowing word representations to adapt based on sentence context. Comparing the principal components from the original and disentangled models, we assessed whether our disentanglement approach effectively reduces biased gender representation in embeddings. These comparisons are vital for validating the effectiveness of disentangling semantic content from gender information in mitigating gender bias within neural embeddings. Our observations can be enumerated as follows: **(1)** When comparing the principal components for gendered terms to those for random terms, we observed a higher variance among the gendered terms. This finding validates our experimental approach by highlighting the fact that gendered terms have a more pronounced first principal component compared to random terms, which can be an indication that this principal component is capturing aspects related to gender. **(2)** The disentangled semantics model shows a higher first component percentage compared to the original model, suggesting a more significant separation of semantic information from gender influences. The disentangled gender model exhibits an even higher first eigenvalue than the disentangled semantics model. Given that the disentangled semantics model focuses on detecting gender spaces, the predominance of a single principal component explaining a large variance aligns with our expectations. This principal component is likely capturing the primary axis of gender differentiation, further confirming that our model effectively disentangles gender informa-

tion. **(3)** Although similar trends were observed with random terms, the intensity of the variance was much less pronounced. This is likely due to the shorter context and a higher occurrence of pronouns or stereotypically gendered occupations within these samples. Both the content and gender representations in the disentangled models showed higher first principal components compared to the original model, yet these were substantially lower than those observed for gendered terms. These results underscore the efficacy of our disentangled models in isolating and analyzing gender-related components.

It is worth mentioning that in Figure 5.5, the principal component, which is interpreted here as the gender component, shows a significantly higher variance for both gendered and random terms. This observation aligns with trends observed in other studies that apply PCA to analyze bias in representations [141, 139, 55]. Specifically, when applying PCA to data where certain dimensions are prominent or where the variance is heavily influenced by specific attributes, the first principal component tends to capture the majority of this variance. In this context, since the sentences are all focused on stereotypical occupations, it makes sense to have one bold principal component. However, by comparing it to the gendered terms, we see that it is less pronounced; the components in the first row and for gendered terms are significantly greater than those in the second row, which correspond to random terms.

Measuring Bias in Sentence Encoders

The authors in [85] have proposed the Sentence Encoder Association Test (SEAT) to measure social biases in sentence encoders. SEAT is a generalization of the Word Embedding Association Test (WEAT) [27], which was originally designed to measure biases in word embeddings by comparing the associations between sets of words (target concepts) and sets of attributes.

SEAT relies on cosine similarity to measure the association between sentence embeddings. Bias is quantified using a test statistic, $s(X, Y, A, B)$, which measures the difference in cosine

similarity between the embeddings of target concepts X and Y the embeddings of attributes A and B. The magnitude of the association between the target concepts and the attributes is measured by the effect size. It is calculated as the difference in mean cosine similarity scores for the target concepts with respect to the attribute sets, normalized by the standard deviation as follows:

$$\textit{EffectSize} = \frac{\mu_{x \in X}\, s(x, A, B) - \mu_{y \in Y}\, s(y, A, B)}{\sigma_{w \in X \cup Y}\, s(w, A, B)} \tag{5.16}$$

where μ and σ indicate mean and standard deviation and $s(w, A, B)$ is the difference in mean of the cosine similarities:

$$s(w, A, B) = \text{mean}_{a \in A}\, \cos(w, a) - \text{mean}_{b \in B}\, \cos(w, b) \tag{5.17}$$

A larger effect size indicates stronger bias. To apply this bias measurement, specific sentences are crafted using templates that incorporate target concepts (e.g., names associated with a particular race or gender) and attributes (e.g., pleasant or unpleasant adjectives). The sentence embeddings generated by the encoder are then tested for bias by comparing the cosine similarity of the embeddings for different combinations of target concepts and attributes. For example, a biased sentence encoder might show higher similarity between Male names and career-related words compared to Female names reflecting stereotypical gender bias.

Based on SEAT and in our experiments, we adopt and measure effect size as an indicator of bias [85], as explained in Equation 5.16. In this test, the target groups consist of female-related terms and male-related terms, embedded in sentences like “This is a mother” and “This is a father,” respectively. The attributes are represented by sentences associated with science and arts, such as “This is a dance” and “This is chemistry,” respectively. Each target group contains 80 sentences, and there are over 55 attribute sentences in both the science and arts groups.

Table 5.8: Comparison of Gender bias in temrs of Effect Size for MiniLM and BERT-Mini with Original and Disentangled-Semantic Representations.

	MiniLM	**Δ**	**BERT-Mini**	**Δ**
Original	0.482	-	0.256	-
Disentangled-Semantic	0.368	-23.52%	0.214	-16.55%

Table 5.9: Gender bias measures for 215 neutral queries with MiniLM base model for the adversarial training strategy.

cut-off 10	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.1602	0.3183	0.1374	0.1101	0.8107	1.023
Disentangled Model	0.1877	0.0737	0.046	0.0567	0.8664	0.8404
Disentangled Model+Adv	0.1657	0.2298	0.1237	0.1295	0.8881	0.7624
cut-off 20	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR	LIWC↓
Original Model	0.1658	0.2635	0.1142	0.092	0.8274	0.7966
Disentangled Model	0.1941	0.0574	0.035	0.0422	0.8722	0.717
Disentangled Model+Adv	0.1712	0.2320	0.1185	0.1185	0.8853	0.7624

Our objective is to explore the extent to which contextualized sentence representations carry stereotypical gender biases, as shown in Table 5.8. We report the effect sizes for two different pre-trained language models, MiniLM and BERT-Mini, using both the original embedding representations and the representations where gender has been disentangled using our proposed approach. A higher effect size in any of the embeddings indicates a greater degree of bias, suggesting a stronger association of women with the arts and men with science. As illustrated in the table, when the semantic component of the original embeddings is disentangled from gender, both language models demonstrate a lower degree of stereotypical gender biases. The results show that our approach has been able to reduce biases as a result of the disentanglement process, namely MiniLM shows a reduction of over 23% in gender bias in terms of effect size, while BERT-Mini exhibits a reduction of over 16%.

5.6.7 Adversarial Strategy Results

We trained this adversarial network to compare the results with the original model and our proposed disentanglement network. The results for 215 and 1,765 queries with the two base models are presented in Tables 5.9, 5.10, 5.11, and 5.12, respectively.

Table 5.10: Gender bias measures for 1765 neutral queries with MiniLM base model for the adversarial training strategy.

cut-off 10	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.2673	0.1535	0.0721	0.0611	0.7066	1.5599
Disentangled Model	0.2969	0.0805	0.0131	0.0178	0.7623	1.4521
Disentangled Model+Adv	0.2724	0.1337	0.0824	0.0952	0.7915	1.307
cut-off 20	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR ↑	LIWC↓
Original Model	0.2726	0.0721	0.0641	0.0538	0.722	1.3001
Disentangled Model	0.3023	0.0131	0.0313	0.0052	0.7658	1.2767
Disentangled Model+Adv	0.2783	0.1672	0.0911	0.0957	0.7782	1.2912

Table 5.11: Gender bias measures for 215 neutral queries with BERT-Mini base model for the adversarial training strategy.

cut-off 10	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.1662	0.2544	0.1058	0.0751	0.8273	0.8467
Disentangled Model	0.1399	0.0376	0.0132	0.0075	0.8583	0.6969
Disentangled Model+Adv	0.1472	0.3889	0.2082	0.2145	0.8313	1.2266
cut-off 20	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR	LIWC↓
Original Model	0.1742	0.2318	0.0929	0.0646	0.8457	0.6964
Disentangled Model	0.1455	0.047	0.0158	0.0083	0.8691	0.5674
Disentangled Model+Adv	0.1544	0.3794	0.1953	0.1960	0.8323	1.0849

From the tables, we observe that the adversarial strategy is not effective in improving ranking performance or reducing bias. For both the MiniLM and BERT-Mini base models, our disentangled model consistently outperforms the original and adversarial models. For instance, in Table 5.9, the Disentangled Model+Adv achieves an NFaIR score of 0.8881, which is better than the Original Model's 0.8107 but still lower than the Disentangled Model's 0.8664. Similarly, the adversarial strategy results in higher ARaB-tc values (e.g., 0.2298 in the Disentangled Model+Adv vs. 0.0737 in the Disentangled Model), indicating less bias reduction.

One reason for this is that gender information may not be entirely redundant or unnecessary for ranking. In some user queries, the presence of gender information could improve performance. For example, consider the query "rsm meaning home care" from the 215-query set by [105]. The relevant document is "'The mission of the Right from the Start Medical Assistance Group (RSM) is to enable children under age 19, pregnant women, low-income families, and women with breast or cervical cancer to receive comprehensive health services

Table 5.12: Gender bias measures for 1765 neutral queries with BERT-Mini base model for the adversarial training strategy.

cut-off 10	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.2475	0.1387	0.056	0.0369	0.7304	1.4942
Disentangled Model	0.1922	0.0928	0.0354	0.026	0.7565	1.3468
Disentangled Model+Adv	0.1472	0.2354	0.1419	0.1621	0.7329	1.6942
cut-off 20	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaIR↑	LIWC↓
Original Model	0.2548	0.1262	0.0.505	0.0329	0.7451	1.2592
Disentangled Model	0.1996	0.0928	0.037	0.0285	0.7672	1.1682
Disentangled Model+Adv	0.1544	0.2632	0.14607	0.1562	0.7299	1.5616

through Medicaid and related programs." Although this query is considered gender-neutral, the relevant document contains female-gendered information crucial for accurately answering the query. In such cases, forcing the model to remove gender information from the ranking representation is counter-productive. However, the adversarial strategy attempts to eliminate gender information, which is not always desirable.

In our proposed disentangled model, there is no external force to remove gender information from the ranking representation. The multitask training of the ranking network and the gender classification network provides the flexibility to determine how much gender information to isolate from the ranking component, optimizing both ranking performance and gender bias reduction. Consequently, our model performs better in terms of ranking performance and bias reduction. Additionally, training the adversarial network is significantly more time-consuming, taking approximately six times longer to converge.

Moreover, we observed that methods enforcing gender removal from the representation do not perform well. For example, the ADVBERT methodology, which is one of our baselines, removes gender information from the intermediate representation of query-document pairs using an adversarial strategy. From Tables 5.1, 5.2, 5.3, and 5.4, we see that the ADVBERT model significantly underperforms compared to our disentanglement approach and fails to reduce gender biases effectively.

5.6.8 Case Study Examples

To further highlight the effectiveness of our proposed model in reducing stereotypical gender biases, we provide specific examples from the set of 215 socially problematic queries. These queries are designed to be neutral, but when gender inequality appears in the ranked list of documents, it can inadvertently perpetuate societal biases against a particular gender. Therefore, it is crucial that the ranked documents for these queries remain impartial, showing no preference for one gender over another.

Tables 5.13,5.14, , and 5.15 present examples of these queries, comparing the top-3 documents ranked by the original model with those re-ranked by our proposed disentanglement approach. Our analysis reveals that the top-3 documents produced by our disentanglement model demonstrate a more balanced representation with respect to gender. Specifically, our model tends to include gender-neutral documents or documents that reference male and female terms equally, thereby mitigating the risk of reinforcing gender stereotypes. This balanced approach is essential for ensuring that search results do not unintentionally contribute to gender bias in society.

Table 5.13 demonstrates a similar issue for the query "physical health effects of stress." The top-3 documents re-ranked by the original model exhibit a bias towards female representation. However, our disentangled model successfully re-ranks the documents to ensure they are gender-neutral, thus preventing any gender bias.

Similarly, Table 5.14 highlights the query "what body fat percentage is healthy," where the original model's top-3 documents show a strong bias towards female representation. Specifically, the top-1 and top-3 documents include only female-related terms, indicating a clear gender bias. In contrast, the top-3 documents re-ranked by our disentangled model present a balanced representation of both male and female terms, effectively mitigating this bias.

In Table 5.15, the query "how is back pay for disability determined" shows a male bias in the top-3 documents re-ranked by the original model, with all documents featuring male-

specific terms. On the other hand, the documents re-ranked by our disentangled model display a more neutral stance, with the top-1 and top-3 documents showing no gender inclination, significantly reducing the overall bias compared to the original model.

Table 5.13: A case study example of the query "physical health effects of stress", and the top 3 re-ranked documents with the original, and disentangled model.

Query: physical health effects of stress	Original Model	Disentangled Model
Rank 1 Document	According to the American Academy of Family Physicians (AAFP), stress is an expression of the body natural instinct to protect itself. While this may warn a woman of immediate danger, like a fast-approaching car, prolonged stress effects can negatively affect your physical and emotional health.	The Physical Effects of Long-Term Stress. Chronic stress can have a serious impact on our physical as well as psychological health due to sustained high levels of the chemicals released in the fight or flight response. Lets take a closer look at whats going on. The Role of the Nervous System.
Rank 2 Document	According to the National Womens Health Information Center, the effects of stress on womens physical and emotional health can range from headaches to irritable bowel syndrome. Specific stress effects include: 1 Eating disorders.	Stress, however, can affect many aspects of physical and mental health, ranging from hair, teeth, and skin to memory and concentration skills, and even how well we sleep. The good news is while these problems may seem serious, stress relief can lead to real improvements in your overall health and well-being
Rank 3 Document	According to the National Womens Health Information Center, the effects of stress on womens physical and emotional health can range from headaches to irritable bowel syndrome. Specific stress effects include: Eating disorders.	It is a well-known fact that stress can affect our lives in many ways. It can even have an adverse affect on our physical health. Severe stress can actually lead to chronic health conditions. It is important to recognize symptoms of severe stress and learn how to cope with stress.

Table 5.14: A case study example of the query "what body fat percentage is healthy", and the top 3 re-ranked documents with the original, and disentangled model.

Query: what body fat percentage is healthy	Original Model	Disentangled Model
Rank 1 Document	Body Fat for Girls. A 5-year-old girl should have 14 to 21 percent body fat, while a 6-year-old girl is considered healthy at 14 to 22 percent. The low end of a healthy body fat range for 7- and 8-year-old girls is 15 percent, while the high end is 24 and 25 percent, respectively. healthy body fat percentage for an 18-year-old girl is between 17 and 30 percent, while a 19-year-old should fall between 19 and 31 percent. Adult females 20 to 39 years should strive for a body fat percentage between 21 and 32 percent.	Go to Body Fat Table. The percentage of body fat in healthy humans ranges from 5 to 40 per cent. Females have more body fat than males. Athletes vary in body fat depending on their sport. Distance runners tend to have a low fat content. While most humans have too much fat some get carried away with trying to achieve unrealistic, unhealthy low levels. For females, body fat should not be less than 15 percent and for males, not less than 5 percent.
Rank 2 Document	A healthy body fat percentage ranges from 10 to 22 percent for men and 20 to 32 percent for women, according to ACSM. This means a healthy percentage of lean mass is 78 to 90 percent for men and 68 to 80 percent for women. You'll get the most accurate assessment of your body fat levels if you consult a professional.	The average healthy. adult body fat range regardless of age is 15 to 20% for men and 20 to 25% for women. A woman with. more than 32% body fat and males with more than 25% body fat are considered to be at increased risk. for disease.
Rank 3 Document	Body Fat for Girls. A 5-year-old girl should have 14 to 21 percent body fat, while a 6-year-old girl is considered healthy at 14 to 22 percent. The low end of a healthy body fat range for 7- and 8-year-old girls is 15 percent, while the high end is 24 and 25 percent, respectively.	As a result, different body fat percentages will be provided with the same health assessment for both genders. For women between age 20 and 40, 19% to 26% body fat is generally good to excellent. For women age 40+ to 60+, 23% to 30% is considered good to excellent.For men between age 20 and 40, 10% to 20% body fat is generally good to excellent. For men age 40+ to 60+, 19% to 23% is considered good to excellent.ctually, 5% body fat can cause serious health problems for the average person. Conversely, 25% fat can either be healthy or unhealthy depending upon your age and gender. In order to provide clarity, it's best to look at a scale of body fat percentages and what they represent.

Table 5.15: A case study example of the query "how is back pay for disability determined", and the top 3 re-ranked documents with the original, and disentangled model.

Query: how is back pay for disability determined	Original Model	Disentangled Model
Rank 1 Document	VA Disability Back Pay is a payment of all the money that the veteran should have been receiving for the months in between his date of eligibility and his VA rating decision. A veteranâs date of eligibility for VA Disability Back Pay is determined in one of two ways.First, if the veteran submits his VA Disability Claim within one year of his date of separation, his date of eligibility for VA Disability Back Pay is his date of separation...	How far back Social Security will pay disability benefits to a disabled person is determined by the date you filed your disability claim when applying for Social Security and/or SSI disability.ocial Security has a five-month waiting period that applies to social security disability claims for which they never pay disability benefits. Basically, the date of filing determines what month you are first entitled to begin receiving monthly Social Security disability benefits.
Rank 2 Document	A veteranâs date of eligibility for VA Disability Back Pay is determined in one of two ways. First, if the veteran submits his VA Disability Claim within one year of his date of separation, his date of eligibility for VA Disability Back Pay is his date of separation.f, however, Ben submits his VA Disability Claim 13 months after he separates, and the VA takes 16 months (unfortunately not unusual) to reach their Rating Decision, he will only receive 16 months of VA Disability Back Pay.	VA Disability Back Pay is a payment of all the money that the veteran should have been receiving for the months in between his date of eligibility and his VA rating decision. A veteranâs date of eligibility for VA Disability Back Pay is determined in one of two ways.First, if the veteran submits his VA Disability Claim within one year of his date of separation, his date of eligibility for VA Disability Back Pay is his date of separation...
Rank 3 Document	A veteranâs date of eligibility for VA Disability Back Pay is determined in one of two ways. First, if the veteran submits his VA Disability Claim within one year of his date of separation, his date of eligibility for VA Disability Back Pay is his date of separation.f, however, Ben submits his VA Disability Claim 13 months after he separates, and the VA takes 16 months (unfortunately not unusual) to reach their Rating Decision, he will only receive 16 months of VA Disability Back Pay.	How far back Social Security will pay disability benefits to a disabled person is determined by the date you filed your disability claim when applying for Social Security and/or SSI disability.

5.7 Discussion

5.7.1 Scalability

In real-world and large-scale applications, the effectiveness and efficiency of a model are both crucial factors that determine its practical usability. The sheer volume of data, the computational resources required, and the time needed for training can all pose significant challenges when deploying a model in real environments. Therefore, it is imperative that the

Table 5.16: Training and inference time of the original and disentangled model.

	MiniLM		BERT-Mini	
	Original	Disentangled	Original	Disentangled
Training Time	01:36':16"	01:39':07"	01:00':42"	01:01':37"
Inference Time	21.96 μs	23.07 μs	15.97 μs	17.97 μs

proposed model for eliminating stereotypical gender biases does not introduce prohibitive levels of complexity or excessive time demands, as these could impede its application in real-world scenarios.

To assess the scalability of our model in large-scale and real-world applications, we evaluated the training and inference times of the proposed disentanglement model. As shown in Table 5.16, we compare these running times for both the original and disentangled models across two base models. All experiments were conducted using an NVIDIA RTX A6000 GPU, which is well-suited for handling high-performance deep learning tasks. The results indicate that the disentanglement approach does not significantly increase the model's running time, as evidenced by the minimal differences in training times (less than three minutes) and inference times (around 1 microsecond) between the original and disentangled models.

These findings confirm that our proposed model remains scalable, making it well-suited for deployment in large-scale applications without compromising performance.

5.7.2 Interpretability

In real-world applications, ensuring that a model is interpretable is crucial, particularly in scenarios where the decision-making processes must be transparent and trackable. Interpretability becomes even more significant in contexts like information retrieval systems, where users and stakeholders need to understand how and why certain results are ranked or presented. One effective approach to enhancing interpretability in these models is through representation disentanglement. By decomposing vector representations, which often encode a mixture of various information, into more interpretable and meaningful components, disentanglement allows us to better understand the underlying factors that influence the

model's decisions. This process of separating distinct attributes within the representations makes it easier to track and explain the model's behavior. Representation disentanglement has been successfully employed in various areas to enhance model interpretability [143, 63, 115, 122, 127, 46].

In the context of gender bias in information retrieval systems, disentangling gender-related information from other aspects of the data can significantly contribute to the interpretability of the ranking model. When gender information is disentangled, it allows researchers and practitioners to isolate and examine the impact of gender on the ranking process, providing clearer insights into whether and how gender biases are influencing the model's outputs. This level of transparency not only helps in identifying potential biases but also aids in the development of more fair and balanced models. By leveraging disentanglement techniques, it becomes possible to create systems that not only perform well but also offer interpretable, bias-aware decision-making processes, which is essential for ethical AI deployment.

5.7.3 Ethical Implications

Gender is considered to be a sensitive attribute, and any attempt to manipulate or alter gender information in machine learning models can lead to significant ethical concerns. This is particularly true in information retrieval systems, where fairness and transparency are paramount. In our work, we emphasize that our approach does not involve changing or manipulating gender attributes in any way. Instead, we focus on disentangling gender from the intermediate representations of query-document pairs. This process ensures that gender influences the ranking decisions for neutral queries to the extent to which it is relevant in the context of the search query, thereby avoiding the introduction of bias or unfair treatment based on gender.

Disentangling gender in this manner does not alter the gender attribute itself. Rather, it aims to create a more unbiased and fair model by ensuring that gender does not unintention-

ally affect the outcomes of the model's decision-making process. This approach is especially important in contexts where the objective is to achieve gender fairness.

5.8 Concluding Remarks

We introduce a novel method for mitigating gender bias in neural ranker representations by disentangling content semantics from gender associations. Our approach isolates gender-related information, enabling the ranker to assess document relevance based solely on semantic content. Experimental results show our method outperforms state-of-the-art baselines in reducing gender bias while maintaining ranking effectiveness, decreasing the performance gap between male and female queries by around 27% at cut-offs 10 and 20. Our disentanglement strategy effectively weakens gender information in the intermediate vector representation of the cross-encoder. This balance between fairness and performance is crucial for developing unbiased neural rankers. Our methodology, while focused on the gender attribute, can be applied to other sensitive attributes like ethnicity and race for future work, promoting fairer information retrieval systems. It integrates seamlessly with existing neural rankers, allowing for immediate deployment without sacrificing performance, scalability, or accuracy, while promoting unbiased ranking of search results.

Chapter 6

Bias-aware Curriculum Sampling

6.1 Problem Formulation

Let $\mathcal{Q} = \{q_1, q_2, \ldots, q_N\}$ represent a set of queries and $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$ denote a collection of documents. The objective of a neural ranker Φ is to identify and rank the most relevant documents from $\mathcal{D}$ for each query q_i based on a relevance score $s(q_i, d_j)$, where $d_j \in \mathcal{D}$. We consider training samples to be structured as $\mathcal{S} = \{(q_i, d_{ij}^+, d_{ik}^-)\}$, where d_{ij}^+ denotes a relevant (positive) document and d_{ik}^- represents an irrelevant (negative) document for a given query q_i. The model Φ is trained to maximize the relevance score difference between positive and negative examples.

Training datasets often contain biases that subtly influence model learning [2, 21, 86, 140]. A document d_j may encode biases, such as inclination towards a certain gender, quantifiable via a bias scoring function, $\Psi(d_j)$ [107, 105, 1, 95]. These biases risk entangling with relevance signals, leading the model to misinterpret biased patterns as relevance indicators, compromising ranking fairness. Mitigating this requires (1) maintaining high ranking effectiveness, $\Lambda(\mathcal{Q})$, while minimizing bias in outputs, $\Pi(\mathcal{Q})$; and (2) sequencing training samples to positively shape learning dynamics.

Curriculum Learning Process. To decouple bias from the model's learning of rele-

vance, we structure the training process as a two-stage procedure. In the first stage, *Initial Relevance Learning*, the model Φ is trained on samples $\mathcal{S}_{\text{low-bias}}$ (or $\mathcal{S}_{\text{high-bias}}$) depending on the direction of the curriculum, which consist of documents with low (or high) bias scores, $\Psi(d_{ij}^{+}) < \epsilon$, where ϵ is a predefined threshold. In the second stage, *Gradual Bias Introduction*, the model is progressively exposed to samples $\mathcal{S}_{\text{high-bias}}$ (or $\mathcal{S}_{\text{low-bias}}$) in the alternative case) with differing degrees of bias. This gradual exposure allows the model to generalize its understanding of relevance while enhancing robustness against biases. Let $\Lambda(\mathcal{Q})$ be a metric that evaluates the ranking effectiveness on a set of queries $\mathcal{Q}$, and $\Pi(\mathcal{Q})$ denote a metric that quantifies bias in the ranked outputs for $\mathcal{Q}$. Our learning objective can be expressed as finding the parameters θ of the model Φ such that:

$$\arg\max_{\theta} \Lambda(\mathcal{Q}) \quad \text{subject to} \quad \Pi(\mathcal{Q}) \to 0. \tag{6.1}$$

where $\Lambda(\mathcal{Q})$ is comparable to baselines to maintain performance.

6.2 Methodology

We propose a training strategy to achieve the proposed learning objective, comprising: *(i)* an adaptive curriculum guiding the training sequence with bias-aware sampling, *(ii)* a controlled probability distribution to balance exposure, and *(iii)* a learning objective aligning relevance learning with fairness.

Bias-Aware Curriculum Design. A key challenge in bias-aware learning is preventing models from misinterpreting biases for relevance signals. Building on curriculum learning principles [6, 58, 130, 123], we hypothesize that introducing less (or high, depending on curriculum direction) biased samples earlier in training may shape the model's understanding of bias and possibly mitigates bias. To achieve this, we define a bias scoring function, $\Psi(d_{ij}^{+})$, quantifying bias in each document d_{ij}^{+}. This score informs our sampling strategy, adjusting document selection probabilities. Prioritizing low (or high) bias samples early

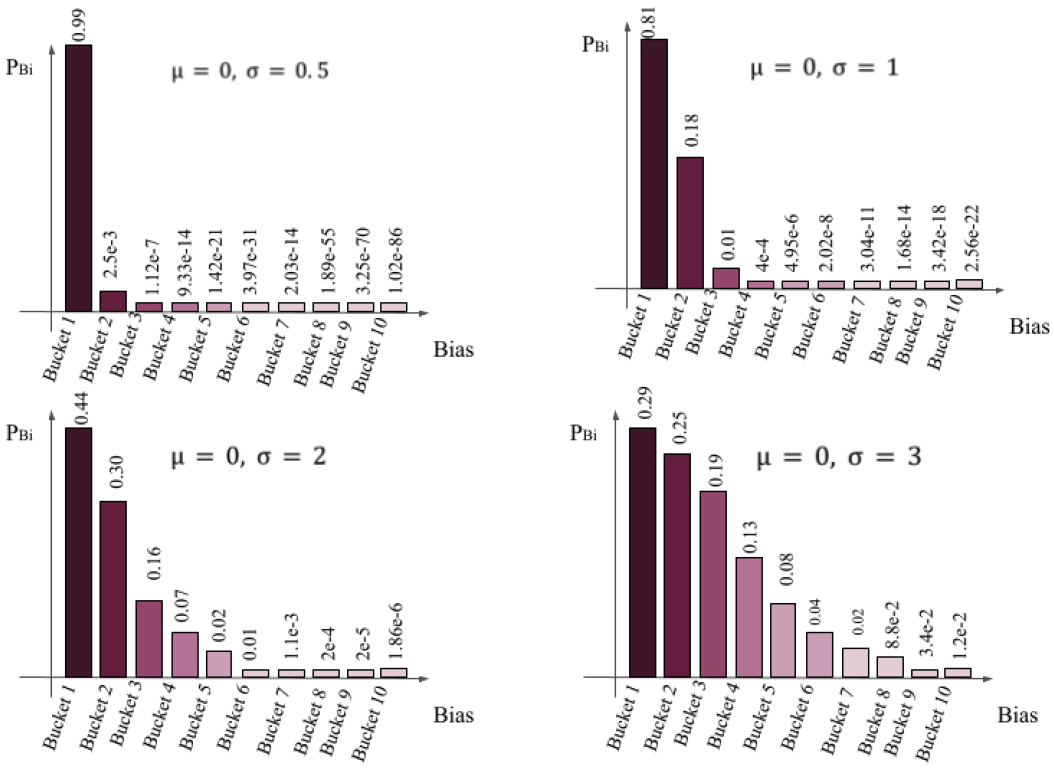

Figure 6.1: Sampling probs for 10 buckets, with $\sigma = \{0.5, 1, 2, 3\}$.

impacts the risk of bias influencing initial learning, potentially creating a fairer relevance baseline. As training progresses, higher (or lower) bias samples are gradually introduced, refining relevance without embedding bias.

Dynamic Sampling Strategy with Bias Scoring. We propose that the bias score of each relevant document d_{ij}^{+} may serve as a key factor in determining the training sequence. Since relevant documents shape the model's understanding of relevance for a query q, any bias within them risks being misinterpreted as a relevance signal. If only high (or low) bias documents appear early in training, the model may internalize these degrees of bias as relevance indicators. To mitigate this, we propose a controlled sampling strategy where the degree of bias of a document determines its selection probability in the early phases. This encourages the model to first focus on learning relevance and gradually engage with the concept of bias through its exposure to samples with progressive degrees of bias.

To quantify the bias within each relevant document in a training sample $S = (q_i, d_{ij}^{+}, d_{ik}^{-})$, we compute a bias score for each relevant document d_{ij}^{+}, denoted as $\Psi(d_{ij}^{+})$. This score

reflects the degree of bias present in the document and serves as the primary metric for ranking training samples based on their degree of bias. Specifically, the bias score Ψ is a function mapping each relevant document d_{ij}^+ to a real-valued score, defined as:

$$\Psi : \mathcal{D}^+ \to \mathbb{R}, \quad \Psi(d_{ij}^+) = \text{bias score of } d_{ij}^+, \quad d_{ij}^+ \in \mathcal{D}^+$$

where $\mathcal{D}^+$ represents the set of all relevant documents.

Given the bias score of each training sample, the samples in S are sorted, forming an ordered set S_{sorted}. This arrangement controls the sequence of sample exposure. The ordered set is then divided into discrete buckets B_i, each containing a fixed number of samples. A function $\beta(S, b)$ partitions S_{sorted} into b equally sized buckets: $B_i = \beta(S_{\text{sorted}}, b)$. Each bucket has size b with a total of N equally sized buckets. Buckets group samples by bias level, enabling distinct sampling probabilities. This approach ensures controlled exposure, prioritizing samples with different bias levels throughout training.

Adaptive Probability Distribution for Sampling. A sampling probability P_{B_i} is assigned to each bucket B_i to regulate model exposure to biased data. For instance, in order to ensure buckets containing higher bias samples have a lower sampling probability earlier in the training process, P_{B_i} can be defined as inversely proportional to the average bias score x_i of bucket B_i:

$$P_{B_i} \propto \frac{1}{x_i}, \quad \text{where } x_i = \frac{1}{|B_i|} \sum_{d_{ij}^+ \in B_i} \Psi(d_{ij}^+). \tag{6.2}$$

where $|B_i|$ is the number of documents in B_i, and $\Psi(d_{ij}^+)$ is the bias score of document d_{ij}^+. To refine the sampling framework, we model the bucket sampling probabilities P_{B_i} using a Gaussian distribution, ensuring a smooth probability curve. Adjusting the distribution parameters controls the spread, assigning higher probabilities to differing buckets. The Gaussian function can be defined as $P_X(x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$ where x_i is the average bias score for bucket B_i, μ defines the distribution center, and σ controls its spread. To ensure probabilities sum to one, we normalize them. All samples within a bucket B_i share

the same probability P_{B_i}, ensuring consistent bias management. A smaller σ sharpens the peak, emphasizing on the earlier buckets in training, while a larger σ smooths the transition across bias levels.

As an illustrative example, Figure 6.1 shows the sampling probability distribution across ten buckets, demonstrating how the Gaussian model regulates data exposure. In this case, lower-bias buckets receive higher sampling probabilities, while higher-bias buckets have progressively lower ones. We hypothesize that this structured approach promotes fairness and robustness during training. We note that in this approach, all training instances will eventually be sampled. Lower initial sampling probabilities do not exclude samples but delay their introduction.

Progressive Learning Objective. The selected samples $S = (q_i, d_{ij}^+, d_{ik}^-)$ are then fed into a cross-encoder neural ranker. The model calculates relevance scores for both the relevant document d_{ij}^+ and the irrelevant document d_{ik}^- in relation to the query q_i: $s(q_i, d_{ij}) = \Phi(q_i \oplus d_{ij})$. The model is trained with a Max Margin Loss [36] calculated as:

$$L = \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - \Phi(q, {d_i}^+) + \Phi(q, {d_j}^-)) \tag{6.3}$$

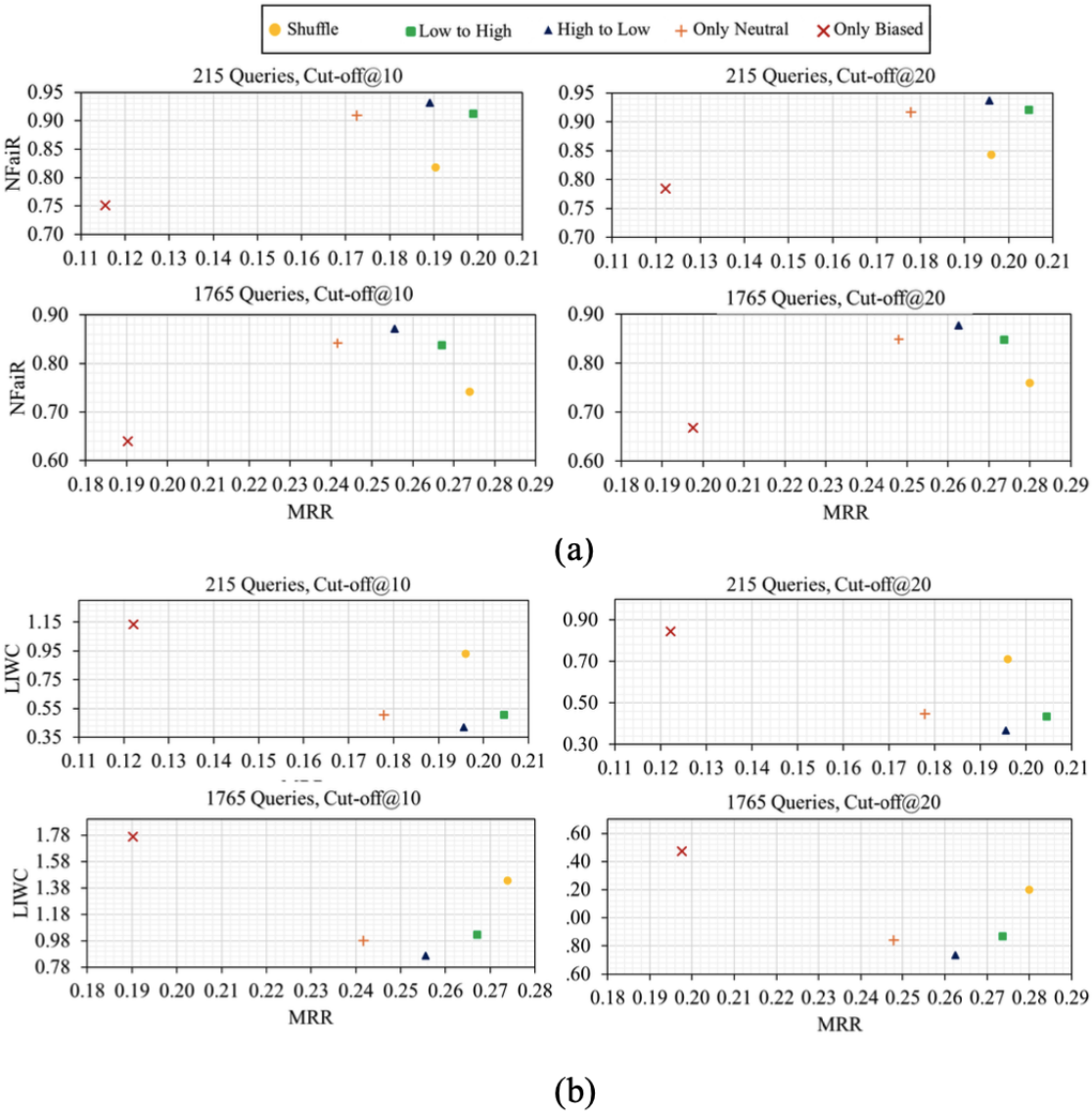

Figure 6.2: The Bias-performance trade-off.

6.3 Experiments

Research Questions. Our experiments are structured around four Research Questions (RQs): **RQ1.** Does curriculum sampling effectively mitigate biases in neural rankers? Specifically, we investigate whether structuring the training sequence based on bias levels can reduce bias in ranking outputs while maintaining model effectiveness. **RQ2.** How does the model perform relative to the state-of-the-art bias reduction baseline methods? **RQ3.** Does the choice of probability distribution hyperparameters for bias-aware sampling influence the model's ranking performance and bias mitigation effectiveness? **RQ4.** Is the model's performance consistent across different language models? We run experiments on 3 language models, BERT-mini [124, 42], MiniLM [128], and ELECTRA [33] to assess the generalizability of our approach across LLMs.

Datasets and Setup. We train the neural rankers on the MSMARCO passage ranking dataset [89], with 200,000 queries and 8.8 million passages. A random sample of 3,000,000 triples is used for training over one epoch, using the Adam optimizer with a sigmoid activation. We follow OpenMatch [113] architecture, implementation, and hyperparameters. Full implementation details and source code are available on GitHub: https://shorturl.at/e3ggk.

Bias Measure and Bias Test Datasets. We consider ARaB-tf [107] as the function Ψ in Equation 6.2 for measuring bias of the documents. To evaluate performance and bias reduction, we focus on *gender bias* using two bias query datasets: (a) *Gender-neutral queries*: These queries assess whether the model introduces gender stereotypes in neutral contexts. We use the query set from [105], which includes 1,765 gender-neutral queries, selected from MS MARCO queries. (b) *Socially sensitive queries*: This set includes 215 queries that may contribute to societal inequality if bias is present.

Evaluation Metrics. For ranking, we use Mean Reciprocal Rank (MRR) [89]. For bias, we use three metrics: *Average Rank Bias (ARaB)* [107], which quantifies biased word occurrences in documents using Term Count (TC), Term Frequency (TF), and Boolean metrics; *NFaiRR* [105], measuring document-level fairness, with higher values indicating fairer rank-

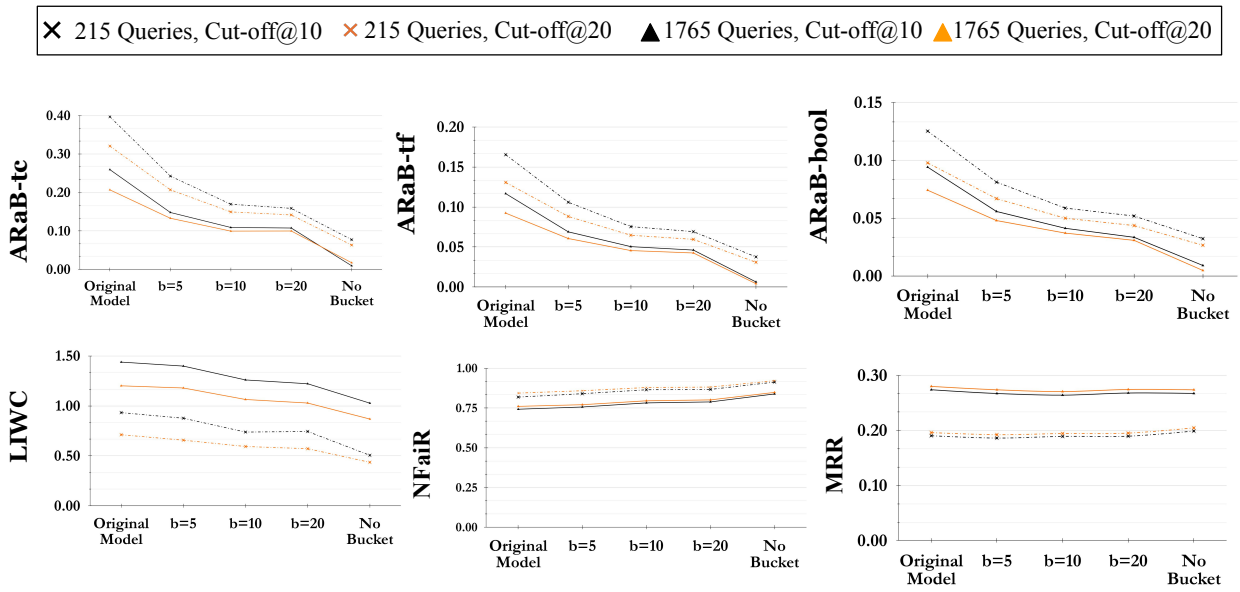

Figure 6.3: Impact of bucket size on model performance.

ings; and *Linguistic Inquiry and Word Count (LIWC)* [95], which assesses gender mentions in text as suggested by [19].

Baseline Methods. To benchmark our approach, we compare it against five established baselines representing diverse bias mitigation strategies: (1) `AdvBert` [105] uses adversarial debiasing in the ranker's intermediate layers; (2) `Bias-aware Loss` [116] integrates a bias penalty in the loss function for targeted bias reduction during training; (3) `CODER` [137][0] applies a neutrality regularization term in a transformer model; (4) `Light-Weight Sampling Strategy (LWS)` selects biased documents as negative samples, training the model to recognize and mitigate bias.

Findings. In **RQ1**, we assess whether our proposed curriculum sampling approach reduces bias. We conduct experiments under five conditions: (1) shuffled training (no curriculum), (2) High to Low bias curriculum sampling, (3) Low to High bias curriculum sampling, (4) training on neutral samples (bias-free), and (5) training on biased samples. Figure 6.2(a) shows the trade-off between fairness (NFaiR [105]) and effectiveness (MRR). Training on biased samples yields the worst fairness and effectiveness, performing worse than the shuffled baseline. Neutral-sample training improves fairness but reduces performance by up to 10%.

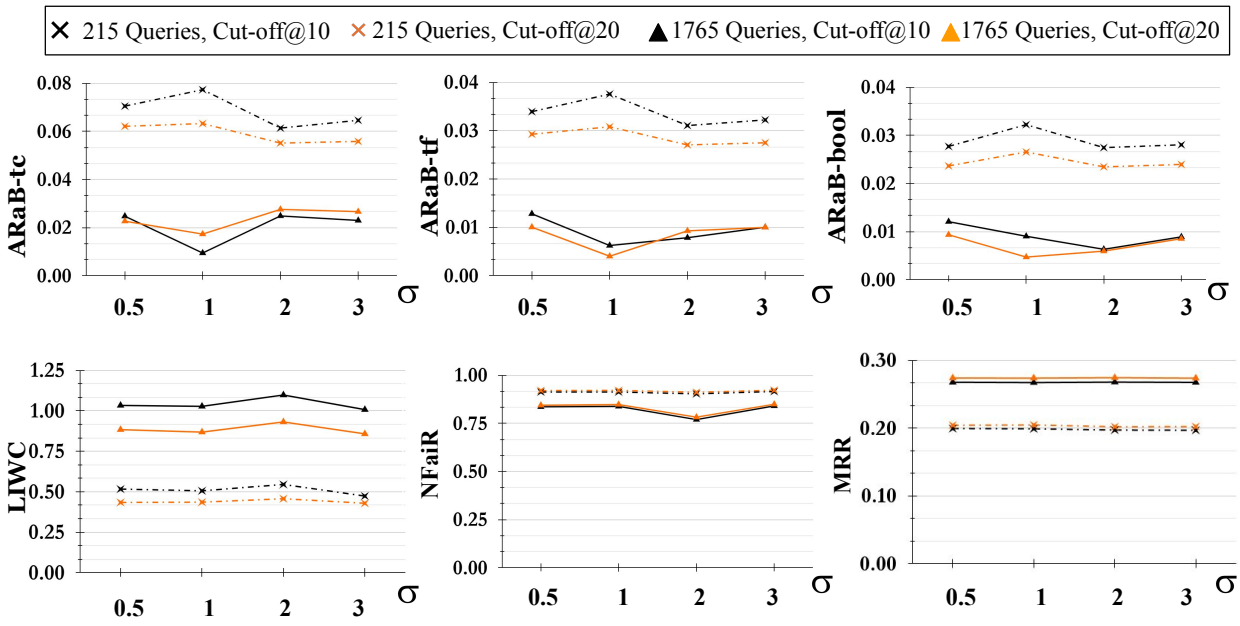

Figure 6.4: Impact of σ on model performance.

The High to Low strategy improves fairness but still lowers effectiveness (up to 7%). In contrast, our proposed *Low to High curriculum strategy* achieves the best fairness-effectiveness trade-off, as further confirmed in Figure 6.2(b), where lower LIWC values indicate reduced bias. Empirical results show that *Low to High curriculum sampling* best balances bias reduction and retrieval effectiveness. By prioritizing low-bias samples early, the model establishes a robust relevance foundation before gradually incorporating higher-bias samples. This controlled exposure prevents early bias internalization, allowing the model to learn relevance signals more effectively. Without loss of generality and to save space, we adopt this strategy for reporting subsequent RQs. In **RQ2**, we compare the performance of our proposed approach with the state-of-the-art baseline methods. Tables 6.1 and 6.2 show this comparison. We observe that our proposed method outperforms the bias-aware loss, Light-Weight-Sampling, and ADVBERT methods in terms of bias reduction, while having higher MRR. The other baseline method, CODER, although is able to reduce the bias more than our proposed approach, but it significantly reduces retrieval effectiveness to 0.0014 for cut-off 10, and 0.0001 for cut-off 20 on the 215 query set, effectively making unhelpful retrieval.

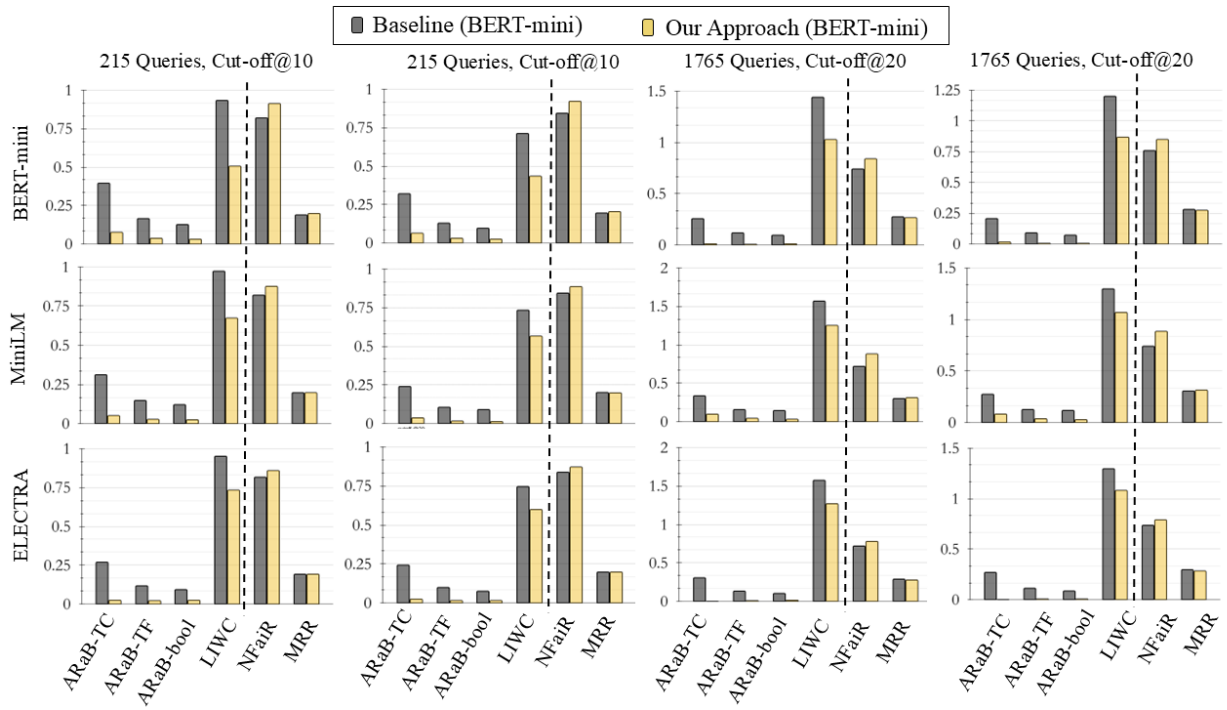

Figure 6.5: Generalizability of our proposed approach on different LLMs.

In **RQ3,** we examine the impact of probability distribution parameters on model performance, focusing on bucket size b and standard deviation σ in sampling probabilities. Figure 6.3 shows the effect of varying bucket sizes using a Normal Distribution ($\mu = 0, \sigma = 1$) and BERT-mini as the baseline. The analysis covers two query sets (215 and 1,765 queries) at cut-offs 10 and 20. The results reveal: *(1) Bias Reduction:* Increasing b from 5 to 20 reduces bias across all metrics (ARaB-TC, ARaB-TF, ARaB-Bool, LIWC, and NFaiR). *(2) Effect of Larger Buckets:* Larger bucket sizes further enhance bias reduction, with $b = 20$ outperforming $b = 5$. *(3) No-Bucket Strategy:* Treating each sample as an individual bucket maximizes performance and minimizes bias, aligning with the trend that larger bucket sizes reduce bias more effectively. Figure 6.4 examines the effect of varying σ on model stability and bias metrics. Our results show *(i) MRR Stability:* MRR remains stable across sigma values (0.27 on 1,765 queries, 0.22 on 215 queries), showing minimal impact on retrieval effectiveness. *(ii) Consistency in Bias Metrics:* Bias measures (ARaB-TC, ARaB-TF, ARaB-BOOL, LIWC, NFaiR) exhibit less than 5% variation across sigma values.

Table 6.1: Bias & retrieval effectiveness on the 215 query set.

			cutoff @10			
Models	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaiR ↑	liwc ↓
Bias-Aware Loss [116]	0.1820	0.3419	0.1492	0.1176	0.8209	0.9202
Light-Weight-Sampling [14]	0.1823	0.2017	0.0938	0.0782	0.9087	0.5636
CODER [137]	0.0014	0.0260	0.0171	0.0205	0.9649	0.2998
ADVBERT [105]	0.1753	0.1975	0.1054	0.1113	0.8747	0.7850
Our Approach	**0.1989**	**0.0773**	**0.0376**	**0.0322**	**0.9126**	**0.5057**

			cutoff @20			
Models	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaiR ↑	liwc ↓
Bias-Aware Loss [116]	0.1873	0.2783	0.1169	0.0899	0.8519	0.6650
Light-Weight-Sampling [14]	0.1876	0.1618	0.0746	0.0616	0.9168	0.4681
CODER [137]	0.0001	0.0227	0.0148	0.0178	0.9650	0.2828
ADVBERT [105]	0.1799	0.1144	0.0653	0.0710	0.8795	0.6432
Our Approach	**0.2046**	**0.0632**	**0.0308**	**0.0265**	**0.9212**	**0.4355**

Table 6.2: Bias & retrieval effectiveness on the 1,765 query set.

			cutoff @10			
Models	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaiR ↑	liwc ↓
Bias-Aware Loss [116]	0.2591	0.2109	0.0949	0.0755	0.7289	1.5142
Light-Weight-Sampling [14]	0.2558	0.1540	0.0764	0.0680	0.8204	1.1500
CODER [137]	0.0001	0.0646	0.0371	0.0421	0.8404	0.7199
ADVBERT [105]	0.2019	0.4222	0.2260	0.2363	0.7132	1.6427
Our Approach	**0.2671**	**0.0095**	**0.0062**	**0.0090**	**0.8382**	**1.0275**

			cutoff @20			
Models	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaiR ↑	liwc ↓
Bias-Aware Loss [116]	0.2653	0.1644	0.0730	0.0574	0.7578	1.2169
Light-Weight-Sampling [14]	0.2622	0.1192	0.0587	0.0516	0.8313	0.9614
CODER [137]	0.0014	0.0674	0.0388	0.0440	0.8407	0.6467
ADVBERT [105]	0.2106	0.2731	0.1475	0.1554	0.7424	1.2933
Our Approach	**0.2737**	**0.0173**	**0.0040**	**0.0047**	**0.8478**	**0.8684**

In **RQ4**[1]**,** we assess whether our approach generalizes across different LLMs while reducing bias and maintaining retrieval effectiveness. We repeat experiments with MiniLM and ELECTRA alongside BERT-mini. Since Figure 6.3 indicates the best results occur in the no-bucket scenario, we train models with no bucket, and $\sigma = 1$. Figure 6.5 presents bias reduction and ranking performance across the LLMs. The first, second, and third rows show results for BERT-mini, MiniLM, and ELECTRA, respectively. Metrics to the left of the dotted line measure bias (lower is better), while those on the right assess fairness and effectiveness (higher is better). Our approach significantly reduces bias compared to the baseline (no curriculum sampling) while increasing the NFaiR fairness metric. Additionally,

[1]All results are statistically significant based on a paired t-test with a p-value < 0.05.

MRR remains comparable to the original model, confirming that bias reduction does not compromise ranking effectiveness in our proposed curriculum sampling approach.

6.4 Concluding Remarks

In this chapter, we introduced a curriculum learning approach for addressing bias in neural rankers. By structuring the training process through a staged exposure to biased samples, we enabled neural rankers to learn relevance while minimizing the risk of embedding biases into their trained model. Our proposed bias-aware curriculum and dynamic sampling strategy achieved gradual bias exposure in a controlled, systematic manner, supporting both model stability and performance. Our experimental results demonstrated that this approach not only improved bias mitigation but also enhanced ranking effectiveness, underscoring its potential for advancing fairness in information retrieval systems.

Chapter 7

Conclusions

In this thesis, we addressed the challenge of gender bias in neural information retrieval (IR) systems by proposing a set of principled interventions across key stages of the learning pipeline: training strategies, embedding representations, and data sampling. Our goal was to develop bias-mitigation approaches that preserve retrieval effectiveness while promoting fairness and social responsibility in ranked outputs.

First, we introduced bias-aware training strategies that incorporate fairness directly into the model's optimization objectives. By modifying standard loss functions to penalize biased documents and reward unbiased ones, we designed a training paradigm that guides neural rankers toward equitable decision-making. A document-level gender bias penalty dynamically adjusted the learning process, leading to consistent improvements in the fairness-performance trade-off compared to baseline methods.

In addition, we proposed a disentangled representation learning framework that isolates gender-specific and relevance-related components within query-document embeddings. Through dual-objective optimization, we ensured that ranking decisions were based solely on content-relevant signals, effectively removing gender bias from the model's internal representations. This approach demonstrated strong generalization across multiple datasets and embedding architectures, validating its robustness and transferability.

furthermore, we tackled the issue of biased training data through a curriculum-inspired sampling strategy. Inspired by curriculum learning, we structured the training process to begin with gender-neutral samples and gradually introduce biased instances. This progressive exposure, guided by statistical estimates of document-level bias, allowed the model to first establish an unbiased foundation before encountering complex, biased examples. Our experiments showed that this curriculum-aware strategy significantly enhances bias mitigation when used alone or in combination with other methods.

Taken as a whole, this body of work establishes a comprehensive framework for mitigating gender bias in neural IR systems through targeted interventions at the training, representation, and data levels. Extensive experiments on real-world datasets confirm that our methods reduce the propagation of societal stereotypes in ranked results without sacrificing— and in some cases improving— retrieval performance. This thesis contributes new methodologies and empirical insights that advance the development of fair, effective, and socially responsible information retrieval technologies.

7.1 Future Work

In this section, we study potential paths for future research on gender bias in information retrieval systems. We note that the quest for fair and unbiased IR systems is a continuous journey. As our understanding of gender evolves, IR systems must remain agile and adaptable. Beyond just gender, there is a growing acknowledgment of the need to study fairness in the broader sensitive attribute context, encompassing biases like race, age, or socio-economic status. Intersectionality, the interplay of multiple biases, adds another layer of complexity to this endeavour. As the field progresses, there's a growing emphasis on making IR systems transparent in their operations and methodologies, ensuring users have a clear understanding of how information is retrieved. Several key areas warrant attention for future research and development. These directions aim to build upon existing foundational work while advancing

the creation of more inclusive, equitable, and effective IR systems.

Intersectionality and Multidimensional Biases

While substantial progress has been made in understanding and mitigating gender bias, future research must explore the intersectionality of biases. Intersectionality refers to the complex, cumulative manner in which different forms of discrimination—such as those based on gender, race, socioeconomic status, and ethnicity—intersect and interact. Addressing gender bias in isolation may lead to incomplete solutions, as individuals' experiences of bias are often multifaceted. Future work should focus on developing methodologies that consider the interplay of multiple biases, ensuring that IR systems provide equitable access to information for all users, regardless of their intersecting identities.

Beyond Binary Gender Classifications

Current IR systems largely operate within a binary framework of gender, which does not account for the diverse spectrum of gender identities present in modern societies. Future research should explore how IR systems can move beyond binary classifications to better serve users who identify as non-binary, genderqueer, or with other non-traditional gender identities. This includes developing datasets that accurately reflect this diversity and creating algorithms capable of processing and responding to gender in a more nuanced manner. Such advancements would not only enhance the inclusivity of IR systems but also improve their accuracy and relevance to a broader user base.

Addressing the Limitations of Annotated Datasets

The limited size of annotated datasets poses a significant challenge to the reliable detection and analysis of biases in IR systems. Small datasets can lead to overfitting, reduced generalizability, and unreliable bias metrics. Future research should focus on developing methods for efficiently expanding annotated datasets, including semi-supervised learning, active learning,

and crowdsourcing techniques. Additionally, creating benchmarks and standards for dataset annotation processes will help ensure the quality and consistency of the data, enabling more accurate assessments of bias and fairness.

Development of Robust and Reliable Fairness Metrics

The development of robust fairness metrics is critical for assessing and mitigating bias in IR systems. However, current metrics often suffer from reliability issues, including inconsistencies across different datasets and contexts. Future research should focus on creating metrics that are not only resilient to variations in data and algorithms but also consistently reliable across diverse scenarios. These metrics should be validated through extensive testing and benchmarking, ensuring that they can detect biases accurately and consistently. Enhancing the reliability of fairness metrics will be key to making meaningful progress in bias mitigation.

Handling Multilingual Data and Cross-Cultural Biases

As IR systems are increasingly deployed in multilingual and multicultural contexts, addressing the unique challenges of multilingual data is essential. Biases can manifest differently across languages due to cultural nuances, differences in language structure, and varying levels of data availability. Future research should focus on developing methods for detecting and mitigating biases in multilingual datasets, including cross-lingual transfer learning and the creation of language-agnostic fairness metrics. Additionally, ensuring that IR systems are culturally sensitive and capable of fairly serving users from diverse linguistic backgrounds is critical for achieving global inclusivity.

Real-Time Bias Mitigation

As IR systems increasingly operate in real-time, the ability to detect and mitigate bias on-the-fly becomes crucial. Future research should explore the development of real-time

bias detection and correction mechanisms that can be integrated into IR systems without compromising their performance. This includes advancements in machine learning models that can adapt to new data and evolving biases, ensuring that IR systems remain fair and equitable as they process information in dynamic environments.

Definition and Scope of Sensitive Attributes

A clear and consistent definition of sensitive attributes is essential for developing effective bias mitigation strategies in IR systems. Future research should focus on establishing standardized criteria for what constitutes a sensitive attribute, considering the socio-cultural context and the potential impact on different user groups. This includes not only traditional attributes like gender and race but also context-specific attributes that may emerge as relevant in certain applications. Defining sensitive attributes rigorously will guide the development of fairness metrics and the creation of datasets that accurately reflect the diversity of user identities and experiences.

Bibliography

[1] Amin Abolghasemi, Leif Azzopardi, Arian Askari, Maarten de Rijke, and Suzan Verberne. Measuring bias in a ranked list using term-based representations. In European Conference on Information Retrieval, pages 3–19. Springer, 2024.

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and machine learning: Limitations and opportunities. MIT press, 2023.

[3] Christine Basta, Marta R Costa-Jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. arXiv preprint arXiv:1904.08783, 2019.

[4] Christine Basta, Marta Ruiz Costa-jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. CoRR, abs/1904.08783, 2019.

[5] Yoshua Bengio, Aaron Courville, and et. al Vincent. Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013.

[6] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In Proceedings of the 26th annual international conference on machine learning, pages 41–48, 2009.

[7] Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. Generalization in nli: Ways (not) to go beyond simple heuristics. arXiv preprint arXiv:2110.01518, 2021.

[8] Deblina Bhattacharjee, Tong Zhang, Sabine Süsstrunk, and Mathieu Salzmann. Mult: An end-to-end multitask learning transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 12031–12041, June 2022.

[9] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In The 41st international acm sigir conference on research & development in information retrieval, pages 405–414, 2018.

[10] Amin Bigdeli. Exploration and mitigation of stereotypical gender biases in information retrieval systems. 2021.

[11] Amin Bigdeli, Negar Arabzadeh, and Ebrahim Bagheri. Learning to jointly transform and rank difficult queries. In European Conference on Information Retrieval, pages 40–48. Springer, 2024.

[12] Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Bhaskar Mitra, Morteza Zihayat, and Ebrahim Bagheri. De-biasing relevance judgements for fair ranking. In European Conference on Information Retrieval, pages 350–358. Springer, 2023.

[13] Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Bhaskar Mitra, Morteza Zihayat, and Ebrahim Bagheri. De-biasing relevance judgements for fair ranking. In ECOR. Springer, 2023.

[14] Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. A light-weight strategy for restraining gender biases in neural rankers. In ECIR 2022.

[15] Amin Bigdeli, Negar Arabzadeh, Shirin SeyedSalehi, Morteza Zihayat, and Ebrahim Bagheri. Gender fairness in information retrieval systems. In SIGIR 2022, pages 3436–3439, 2022.

[16] Amin Bigdeli, Negar Arabzadeh, Shirin SeyedSalehi, Morteza Zihayat, and Ebrahim Bagheri. A light-weight strategy for restraining gender biases in neural rankers. In (ECIR 2022), 2022.

[17] Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. A light-weight strategy for restraining gender biases in neural rankers. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, Advances in Information Retrieval, pages 47–55, Cham, 2022. Springer International Publishing.

[18] Amin Bigdeli, Negar Arabzadeh, Shirin Seyersalehi, Morteza Zihayat, and Ebrahim Bagheri. On the orthogonality of bias and utility in ad hoc retrieval. In Proceedings of the 44rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021.

[19] Amin Bigdeli, Negar Arabzadeh, Morteza Zihayat, and Ebrahim Bagheri. Exploring gender biases in information retrieval relevance judgement datasets. In European Conference on Information Retrieval, pages 216–224. Springer, 2021.

[20] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5:135–146, 2017.

[21] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29, 2016.

[22] Shikha Bordia and Samuel R Bowman. Identifying and reducing gender bias in word-level language models. arXiv preprint arXiv:1904.03035, 2019.

[23] Emanuele Bugliarello, Rishabh Mehrotra, James Kirk, and Mounia Lalmas. Mostra: A

flexible balancing framework to trade-off user, artist and platform objectives for music sequencing. In Proceedings of the ACM Web Conference 2022, pages 2936–2945, 2022.

[24] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency, pages 77–91. PMLR, 2018.

[25] Christopher Burges, Tal Shaked, Erin Renshaw, and et. al. Lazier. Learning to rank using gradient descent. In (ICML '05), 2005.

[26] Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R Banaji. Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pages 156–170, 2022.

[27] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334):183–186, 2017.

[28] Jamie Callan, Michael Hoy, Changkuk Yoo, and Le Zhao. The clueweb09 dataset. In Proc. of the 32nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 523–524, 2009.

[29] Mattia Cerrato, Marius Köppel, Alexander Segner, Roberto Esposito, and Stefan Kramer. Fair pairwise learning to rank. In 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), pages 729–738. IEEE, 2020.

[30] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. arXiv preprint arXiv:1312.3005, 2013.

[31] Jiaee Cheong, Selim Kuzucu, Sinan Kalkan, and Hatice Gunes. Towards gender fairness for mental health prediction. In IJCAI, pages 5932–5940, 2023.

[32] Yagmur Gizem Cinar and Jean-Michel Renders. Adaptive pointwise-pairwise learning-to-rank for content-based personalized recommendation. In Proceedings of the 14th ACM Conference on Recommender Systems, pages 414–419, 2020.

[33] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555, 2020.

[34] Charles LA Clarke, Nick Craswell, Ian Soboroff, Azin Ashkan, Eugene Agichtein, and Fernando Díaz. Overview of the trec 2004 terabyte track. In TREC, volume 2004, pages 74–85, 2004.

[35] Charles LA Clarke, Nick Craswell, Ian Soboroff, Azin Ashkan, Eugene Agichtein, and Fernando Díaz. The trec 2012 web track. In TREC, volume 2012, pages 1–12, 2012.

[36] Corinna Cortes. Support-vector networks. Machine Learning, 1995.

[37] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. Overview of the trec 2019 deep learning track. arXiv preprint arXiv:2003.07820, 2020.

[38] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501, 2018.

[39] Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. Bias and unfairness in information retrieval systems: New challenges in the llm era. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 6437–6447, 2024.

[40] Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. Unifying bias and unfairness in information retrieval: A survey of challenges and opportunities with large language models. arXiv preprint arXiv:2404.11457, 2024.

[41] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In Proceedings of the eleventh ACM international conference on web search and data mining, pages 126–134, 2018.

[42] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

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

[44] Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. Queens are powerful too: Mitigating gender bias in dialogue generation. arXiv preprint arXiv:1911.03842, 2019.

[45] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. IEEE Transactions on Technology and Society, 1(2):89–103, 2020.

[46] Yuanqi Du, Xiaojie Guo, Amarda Shehu, and Liang Zhao. Interpretable molecule generation via disentanglement learning. In Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pages 1–8, 2020.

[47] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. Is there a trade-off between fairness and accuracy? a perspective using

mismatched hypothesis testing. In International conference on machine learning, pages 2803–2813. PMLR, 2020.

[48] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference, pages 214–226, 2012.

[49] Naomi Ellemers. Gender stereotypes. Annual review of psychology, 69:275–298, 2018.

[50] Geoffrey Ellis. So, what are cognitive biases? Cognitive biases in visualizations, pages 1–10, 2018.

[51] Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. Information Processing & Management, 57(6):102377, 2020.

[52] Eve Fleisig and Christiane Fellbaum. Mitigating gender bias in machine translation through adversarial learning. arXiv preprint arXiv:2203.10675, 2022.

[53] Batya Friedman and Helen Nissenbaum. Bias in computer systems. ACM Transactions on information systems (TOIS), 14(3):330–347, 1996.

[54] Bhavya Ghai. Towards Fair and Explainable AI using a Human-Centered AI Approach. PhD thesis, State University of New York at Stony Brook, 2023.

[55] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. arXiv preprint arXiv:1903.03862, 2019.

[56] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. Deep learning, volume 1. MIT press Cambridge, 2016.

[57] Google. Google scholar, 2024. Accessed: 2024-07-23.

[58] Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In international conference on machine learning, pages 1311–1320. Pmlr, 2017.

[59] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. A deep relevance matching model for ad-hoc retrieval. In Proceedings of the 25th ACM international on conference on information and knowledge management, pages 55–64, 2016.

[60] Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 166–174, 2017.

[61] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. Advances in neural information processing systems, 29, 2016.

[62] Sara Hooker. The hardware lottery. Communications of the ACM, 64(12):58–65, 2021.

[63] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. Advances in neural information processing systems, 30, 2017.

[64] Shuo Huang, Junyu Zhou, Han Feng, and Ding-Xuan Zhou. Generalization analysis of pairwise learning for ranking with deep neural networks. Neural Computation, 35(6):1135–1158, 2023.

[65] Internet Live Stats. Google search statistics, 2024. Accessed: 2024-07-23.

[66] ITHAKA. Jstor, 2024. Accessed: 2024-07-23.

[67] Akshita Jha and Radhika Mamidi. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In Proceedings of the second workshop on NLP and computational social science, pages 7–16, 2017.

[68] Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. Disentangled representation learning for non-parallel text style transfer. Association for Computational Linguistics.

[69] Sparck Jones. Report on the need for and provision of an" ideal" information retrieval test collection. 1975.

[70] Nithish Kannen, Yao Ma, Gerrit JJ van den Burg, and Jean Baptiste Faddoul. Efficient pointwise-pairwise learning-to-rank for news recommendation. arXiv preprint arXiv:2409.17711, 2024.

[71] Simone Kopeinik, Martina Mara, Linda Ratz, Klara Krieg, Markus Schedl, and Navid Rekabsaz. Show me a" male nurse"! how gender bias is reflected in the query formulation of search engine users. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pages 1–15, 2023.

[72] Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In Proceedings of The ACM Collective Intelligence Conference, pages 12–24, 2023.

[73] Klara Krieg, Emilia Parada-Cabaleiro, Gertraud Medicus, Oleg Lesota, Markus Schedl, and Navid Rekabsaz. Grep-biasir: A dataset for investigating gender representation bias in information retrieval results. In Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, pages 444–448, 2023.

[74] Klara Krieg, Emilia Parada-Cabaleiro, Markus Schedl, and Navid Rekabsaz. Do perceived gender biases in retrieval results affect relevance judgements?, 2022.

[75] Klara Krieg, Emilia Parada-Cabaleiro, Markus Schedl, and Navid Rekabsaz. Do perceived gender biases in retrieval results affect relevance judgements? In International Workshop on Algorithmic Bias in Search and Recommendation. Springer, 2022.

[76] Joseph Lemley, Shabab Bazrafkan, and Peter Corcoran. Smart augmentation learning an optimal data augmentation strategy. Ieee Access, 5:5858–5869, 2017.

[77] LexisNexis. Lexisnexis, 2024. Accessed: 2024-07-23.

[78] Zhenghao Liu, Kaitao Zhang, Chenyan Xiong, Zhiyuan Liu, and Maosong Sun. Openmatch: An open source library for neu-ir research. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2531–2535, 2021.

[79] Bloomberg L.P. Bloomberg, 2024. Accessed: 2024-07-23.

[80] Scott Lundberg. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874, 2017.

[81] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In International Conference on Machine Learning, pages 3384–3393. PMLR, 2018.

[82] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. Feedback loop and bias amplification in recommender systems. CoRR, abs/2007.13019, 2020.

[83] E Matthew. Peters, mark neumann, mohit iyyer, matt gardner, christopher clark, kenton lee, luke zettlemoyer. deep contextualized word representations. In Proc. of NAACL, volume 5, 2018.

[84] Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. arXiv preprint arXiv:1903.10561, 2019.

[85] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In Jill Burstein, Christy Do-

ran, and Thamar Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[86] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6):1–35, 2021.

[87] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[88] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. Learning to match using local and distributed representations of text for web search. In Proceedings of the 26th international conference on world wide web, pages 1291–1299, 2017.

[89] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. In CoCo@ NIPS, 2016.

[90] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human-generated machine reading comprehension dataset. 2016.

[91] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464):447–453, 2019.

[92] U.S. National Library of Medicine. Pubmed, 2024. Accessed: 2024-07-23.

[93] Anton Osokin, Anatole Chessel, Rafael E Carazo Salas, and Federico Vaggi. Gans for

biological image synthesis. In Proceedings of the IEEE international conference on computer vision, pages 2233–2242, 2017.

[94] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. Text matching as image recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 30, 2016.

[95] James W. Pennebaker, Martha E. Francis, and Roger J. Booth. Linguistic Inquiry and Word Count. Lawerence Erlbaum Associates, Mahwah, NJ, 2001.

[96] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates, 71(2001):2001, 2001.

[97] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014.

[98] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621, 2017.

[99] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. CoRR, abs/1802.05365, 2018.

[100] Gideon Popoola and John Sheppard. Investigating and mitigating the performance–fairness tradeoff via protected-category sampling. Electronics, 13(15):3024, 2024.

[101] Flavien Prost, Nithum Thain, and Tolga Bolukbasi. Debiasing embeddings for reduced gender bias in text classification. arXiv preprint arXiv:1908.02810, 2019.

[102] Amifa Raj, Bhaskar Mitra, Nick Craswell, and Michael Ekstrand. Patterns of gender-specializing query reformulation. In Proceedings of the 46th International ACM SIGIR

Conference on Research and Development in Information Retrieval, pages 2241–2245, 2023.

[103] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pages 429–435, 2019.

[104] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In Conference on Empirical Methods in Natural Language Processing, 2019.

[105] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. Societal biases in retrieved contents: Measurement framework and adversarial mitigation for bert rankers. arXiv preprint arXiv:2104.13640, 2021.

[106] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. Societal biases in retrieved contents: Measurement framework and adversarial mitigation for BERT rankers. CoRR, abs/2104.13640, 2021.

[107] Navid Rekabsaz and Markus 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.

[108] Navid Rekabsaz and Markus Schedl. Do neural ranking models intensify gender bias? In Proceedings of the 43rd International ACM SIGIR Conference, 2020.

[109] Thomson Reuters. Reuters, 2024. Accessed: 2024-07-23.

[110] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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.

[111] Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Adrián Mendieta-Aragón, Guillermo Marco-Remón, Maryna Makeienko, María Plaza, Julio Gonzalo, Damiano Spina, and Paolo Rosso. Overview of exist 2022: sexism identification in social networks. Procesamiento del Lenguaje Natural, 69:229–240, 2022.

[112] Clara Rus, Jeffrey Luppes, Harrie Oosterhuis, and Gido H Schoenmacker. Closing the gender wage gap: Adversarial fairness in job recommendation. arXiv preprint arXiv:2209.09592, 2022.

[113] Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set consistency regularization for semi-supervised learning with outliers. CoRR, abs/2105.14148, 2021.

[114] Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. In Proceedings of the international AAAI conference on web and social media, volume 15, pages 573–584, 2021.

[115] Mhd Hasan Sarhan, Abouzar Eslami, Nassir Navab, and Shadi Albarqouni. Learning interpretable disentangled representations using adversarial vaes. In Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data: First MICCAI Workshop, DART 2019, and First International Workshop, MIL3ID 2019, Shenzhen, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019, Proceedings 1, pages 37–44. Springer, 2019.

[116] Shirin Seyedsalehi, Amin Bigdeli, Negar Arabzadeh, Bhaskar Mitra, Morteza Zihayat, and Ebrahim Bagheri. Bias-aware fair neural ranking for addressing stereotypical gender biases. In EDBT, pages 2–435, 2022.

[117] Shirin Seyedsalehi, Amin Bigdeli, Negar Arabzadeh, Morteza Zihayat, and Ebrahim Bagheri. Addressing gender-related performance disparities in neural rankers. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2484–2488, 2022.

[118] Shirin Seyedsalehi, Amin Bigdeli, Negar Arabzadeh, Morteza Zihayat, and Ebrahim Bagheri. Addressing gender-related performance disparities in neural rankers. SIGIR '22, 2022.

[119] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. Societal biases in language generation: Progress and challenges. arXiv preprint arXiv:2105.04054, 2021.

[120] Janet Swim, Eugene Borgida, Geoffrey Maruyama, and David G Myers. Joan mckay versus john mckay: Do gender stereotypes bias evaluations? Psychological Bulletin, 105(3):409, 1989.

[121] Isak Taksa and Jaime Muro Flomenbaum. An integrated framework for research on cross-cultural information retrieval. In 2009 Sixth International Conference on Information Technology: New Generations, pages 1367–1372. IEEE, 2009.

[122] Michael Tsang, Hanpeng Liu, Sanjay Purushotham, Pavankumar Murali, and Yan Liu. Neural interaction transparency (nit): Disentangling learned interactions for improved interpretability. Advances in neural information processing systems, 31, 2018.

[123] Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Brian MacWhinney, and Chris Dyer. Learning the curriculum with bayesian optimization for task-specific word representation learning. arXiv preprint arXiv:1605.03852, 2016.

[124] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. arXiv preprint arXiv:1908.08962, 2019.

[125] Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. Learning to speak and act in a fantasy text adventure game. arXiv preprint arXiv:1903.03094, 2019.

[126] Ellen M Voorhees. Overview of the trec 2004 robust retrieval track. In TREC, volume 2004, page 3, 2004.

[127] Tsun-Hsuan Wang, Wei Xiao, Tim Seyde, Ramin Hasani, and Daniela Rus. Measuring interpretability of neural policies of robots with disentangled representation. In Conference on Robot Learning, pages 602–641. PMLR, 2023.

[128] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. Advances in Neural Information Processing Systems, 33:5776–5788, 2020.

[129] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. IEEE transactions on pattern analysis and machine intelligence, 44(9):4555–4576, 2021.

[130] Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In International conference on machine learning, pages 5238–5246. PMLR, 2018.

[131] Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed ElBachouti, Nianwen Xue, Martha Palmer, Jena D. Hwang, Claire Bonial, et al. Ontonotes release 5.0, 2012.

[132] Westlaw. Westlaw, 2024. Accessed: 2024-07-23.

[133] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. End-to-end neural ad-hoc ranking with kernel pooling. In Proceedings of the 40th International

ACM SIGIR conference on research and development in information retrieval, pages 55–64, 2017.

[134] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. Medical image analysis, 58:101552, 2019.

[135] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th international conference on world wide web, pages 1171–1180, 2017.

[136] Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking, part i: Score-based ranking. ACM Computing Surveys, 55(6):1–36, 2022.

[137] George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. Coder: An efficient framework for improving retrieval through contextual document embedding reranking. arXiv preprint arXiv:2112.08766, 2021.

[138] George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten 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.

[139] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. arXiv preprint arXiv:1904.03310, 2019.

[140] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv preprint arXiv:1707.09457, 2017.

[141] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. arXiv preprint arXiv:1804.06876, 2018.

[142] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4847–4853, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[143] Xinqi Zhu, Chang Xu, and Dacheng Tao. Where and what? examining interpretable disentangled representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5861–5870, 2021.

[144] James Y Zou, Daniel J Hsu, David C Parkes, and Ryan P Adams. Contrastive learning using spectral methods. Advances in Neural Information Processing Systems, 26, 2013.