A Regularization Framework for Gender Bias Mitigation in Dense Neural Rankers

Shirin Seyedsalehi[1*], Morteza Zihayat[2] and Ebrahim Bagheri[3]

[1*]Toronto Metropolitan University, Toronto, Canada.
[2]Toronto Metropolitan University, Toronto, Canada.
[3]University of Toronto, Toronto, Canada.

*Corresponding author(s). E-mail(s): shirin.seyedsalehi@torontomu.com;
Contributing authors: mzihayat@torontomu.ca;
ebrahim.bagheri@utoronto.ca;

Abstract

Dense neural retrievers have improved retrieval effectiveness but can also amplify social biases in ranked results. This paper investigates gender bias in retrieval systems and introduces a fairness-aware training approach that regularizes standard ranking losses with bias and fairness terms. The formulation applies penalty or reward signals at the document level within pairwise objectives, enabling a tunable trade-off between effectiveness and fairness. We evaluate the approach on MS MARCO-derived benchmarks using two encoders (BERT-mini and ELECTRA-small) and two query sets (gender-neutral and socially sensitive). Across ARaB, LIWC, and NFaiRR, our best configurations substantially reduce gender bias while preserving MRR@10 within small to moderate deltas, and in some cases improving effectiveness. We also compare against fairness-aware baselines such as adversarial and neutrality-regularized rankers and find competitive or superior bias reduction under comparable effectiveness. The findings are empirical and scoped to binary gender bias in English on the evaluated datasets and models, without claims of broader generality.

1 Introduction

The introduction of dense neural retrievers has significantly improved the effectiveness of information retrieval (IR) systems by encoding queries and documents into high-dimensional vector spaces that capture semantic similarity [1–3]. Unlike sparse methods based on exact term matching [4], neural approaches mitigate vocabulary mismatch [5, 6] and benefit from large-scale pretraining and domain-specific fine-tuning [7]. However, despite their empirical success, these models have been shown to exhibit and even amplify gender bias in ranked results [8–12]. Neutral queries such as "How

to become an engineer?" or "Best tips for parenting" often yield skewed rankings that disproportionately emphasize one gender [13], reinforcing stereotypical associations.

The biases exhibited by dense neural rankers can be attributed to several interrelated factors [14, 15]. 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. [16] 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 [17]. Third, the optimization objectives of dense retrievers prioritize relevance, often measured through *ranking metrics*, without explicitly incorporating fairness considerations [10]. This tendency can exacerbate biases when rankers disproportionately emphasize features linked to gendered patterns in the data.

Among the challenges previously identified, this work focuses on the optimization objective of dense retrievers, as it directly governs ranker behavior. We aim to develop a *fair ranker* that jointly optimizes for relevance and fairness, ensuring high retrieval effectiveness while mitigating biases, such as disproportionate gender representations, that reinforce harmful stereotypes. Dense retrievers are typically trained with objectives narrowly centered on maximizing *relevance*, guided by loss functions that overlook fairness considerations [18]. To address this, we propose regularizing the loss function with *fairness constraints*, enabling the model to learn rankings that are both effective and 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 *pairwise* [19, 20] ranking frameworks. 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*. The key contributions of our work can be enumerated as follows:

1. We propose a framework for fairness-aware ranking, incorporating fairness constraints directly into the loss function to address biases in the ranked results.
2. We design and implement *fairness-aware regularization terms*, enabling the optimization process to balance *relevance* and *fairness* effectively.
3. 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.

4. 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* against a host of strong *state-of-the-art baselines* to demonstrate its effectiveness.

While the idea of augmenting loss functions with fairness constraints is not new and can be conceptually grounded in established loss function regularization techniques [21], our contribution lies not in proposing a novel training paradigm but in systematically adapting and evaluating this approach for gender bias mitigation in dense neural retrievers. Specifically, we rigorously explore how fairness-aware loss regularization can be integrated into neural rankers, and empirically assess their trade-offs with retrieval effectiveness across varied pretrained language models and benchmark datasets.

2 Related Work

Neural approaches in natural language processing have significantly advanced language understanding tasks, but they also reflect and amplify societal biases present in the data they are trained on [22, 23]. Caliskan et al. [24] demonstrated that word embeddings reflect implicit biases, such as associating European-American names more strongly with positive sentiments compared to African-American names, revealing deeply ingrained societal prejudices in the data. Zhao et al. [25] extended this analysis to contextual embeddings, showing that models like BERT perpetuate similar biases even in nuanced contexts. Mehrabi et al. [26] conducted a broad survey of societal biases, highlighting their prevalence across tasks and categorizing them into representational and allocative harms. These findings emphasize the broader challenges posed by societal biases in natural language processing and showing the need for interventions to mitigate them, particularly in downstream applications such as that of the focus of this paper on neural ranking methods.

Bias in dense neural rankers often stems from relevance judgment datasets [27], which play a critical role in training these models. Bigdeli et al. [28] analyzed these datasets using Linguistic Inquiry and Word Count (LIWC) categories and identified systematic imbalances in the linguistic content of relevance judgments. Their findings showed that categories such as *achievement*-related terms were disproportionately associated with male-oriented contexts, while *family*-related terms were more prevalent in judgments tied to female-oriented queries. These biases in the relevance judgment datasets, which are used for training neural rankers, influence the ranker's training, leading to skewed ranking outcomes. To address these issues, a *bias-aware pseudo-relevance feedback framework* [9] has been proposed in the literature, which modifies the feedback process by incorporating *fairness-aware re-ranking* during relevance updates, reducing bias amplification while preserving retrieval effectiveness. Additionally, *bias-aware negative sampling* [29] techniques have been developed to balance the representation of sensitive attributes in training data by carefully selecting irrelevant documents that minimize overrepresentation of specific attributes. Other approaches involve *reweighting relevance judgment collections* [30] to address disparities by ensuring that query-document pairs across gendered contexts are equitably distributed, ultimately reducing performance gaps in retrieval outcomes.

Pre-trained language models (PLMs), such as BERT, form the foundation of dense neural rankers by providing contextualized embeddings that are fine-tuned for ranking

tasks. Despite their effectiveness, PLMs inherit societal biases from the large-scale corpora on which they are pre-trained. Several techniques have been proposed to mitigate these biases and make PLMs more equitable for downstream applications. *Adversarial learning frameworks*, such as those that incorporate fairness-specific adversarial objectives that minimize the predictability of sensitive attributes in latent representations during training [31]. *Counterfactual data augmentation* [32] introduces synthetic examples by altering sensitive attributes, such as gender, in the training data to balance its representation. Another notable approach is *Iterative Nullspace Projection (INLP)* by Ravfogel et al. [33], which removes bias-inducing dimensions from the embeddings while preserving task-relevant information. These debiasing techniques aim to address biases at the representation level, ensuring that PLMs provide a fairer starting point for dense rankers without compromising their contextual and semantic understanding.

Several methods have been developed to explicitly train fair neural rankers by integrating fairness constraints directly into their design and training objectives. The Convolutional Debiasing for Retrieval (`CODER`) method [34, 35] introduces a specialized debiasing layer into the ranker's architecture, which adjusts the learned representations to mitigate overrepresentation of sensitive attributes in relevance scores. Specifically, `CODER` uses convolutional filters to identify and neutralize bias-inducing features in document-query embeddings before computing relevance scores, ensuring that the final rankings are less influenced by stereotypes encoded in the data. `AdvBERT` [36], on the other hand, incorporates adversarial training into the ranking framework. This method introduces a discriminator in the learning process that attempts to predict sensitive attributes, such as gender, from the intermediate representations. Simultaneously, the main ranking model is trained to minimize the discriminator's ability to do so while optimizing for relevance metrics, effectively removing bias-inducing information from the latent space. `CODER` achieves fairness through architectural modifications that directly target embeddings at the feature level, whereas `AdvBERT` relies on adversarial objectives to enforce fairness indirectly during training. Together, these methods exemplify the growing emphasis on principled solutions for mitigating bias and achieving fairness in neural ranking systems, addressing both representation and optimization challenges.

Our work aligns with this category of methods that modify the training process of neural rankers to enhance fairness. Similar to `CODER` and `AdvBERT`, we aim to reduce biases by introducing fairness constraints. However, our approach differs by focusing on regularizing the ranker's loss function rather than modifying its architecture or employing adversarial training. Specifically, we propose a flexible framework that integrates fairness-aware *penalty* and *reward mechanisms* directly into the optimization process, allowing it to balance relevance and fairness dynamically during training. Unlike methods that require architectural modifications, our approach is *model-agnostic* and compatible with existing ranking paradigms [37]. This design enables our framework to achieve a dual optimization of *retrieval effectiveness* and *fairness* by explicitly incorporating bias mitigation into the ranker's objective function. By dynamically adjusting relevance scores during training, our method offers a principled, computationally efficient, and scalable solution for *fairness-aware ranking*.

3 Methodology

3.1 Problem Definition

Let $Q_n = \{q_1, q_2, \ldots, q_n\}$ represent a set of *gender-neutral queries*, which are general, non-gendered information requests. Let $D = \{d_1, d_2, \ldots, d_m\}$ denote a collection of

documents, and let $R_q = \{d_{q_1}, d_{q_2}, \ldots, d_{q_k}\}$ be the ranked set of documents retrieved by an information retrieval system for a query $q \in Q_n$. The ranking in R_q is determined based on the system's assessment of the relevance of each document to the query. The primary objective of an information retrieval system is to ensure high *ranking effectiveness*, measured using $\lambda(R_q)$, which evaluates the relevance of the retrieved documents to the query.

While ranking effectiveness is crucial, it is equally important to ensure that the system does not propagate unintended biases. For instance, the query "How to be a data scientist?" is a *gender-neutral* query. However, *stereotypical gender biases* can emerge if the ranked list R_q for such a query exhibits an undue emphasis on representations associated with a specific gender. If the documents retrieved for this query consistently focus on one gender over another, the system may inadvertently reinforce harmful biases. To quantify such biases, let $\Psi(R_q)$ measure the degree of gender bias present in the ranked list R_q. A desirable situation is to strike the right balance between $\lambda(R_q)$ and $\Psi(R_q)$. High ranking effectiveness ensures that the system effectively meets the user's information needs, while low bias promotes fairness and prevents the perpetuation of harmful stereotypes. Balancing these two objectives is essential because prioritizing only relevance might lead to unfair outcomes, whereas focusing solely on bias reduction could degrade the quality of retrieved results. By achieving this balance, the system can deliver results that are both relevant and equitable, thereby enhancing user satisfaction and contributing to societal fairness.

Let us formalize this. Let Π denote a state-of-the-art ranker, and let $\hat{\Pi}$ represent a fair ranker designed to mitigate gender biases. To ensure both fairness and effectiveness, the fair ranker $\hat{\Pi}$ must satisfy the following criteria:

1. *Maintain Retrieval Performance:* The fair ranker $\hat{\Pi}$ must ensure that its ranking effectiveness is comparable to that of the state-of-the-art ranker Π. This condition ensures that improvements in fairness do not come at the expense of the system's ability to retrieve relevant documents. Formally:

$$\lambda(\hat{\Pi}, Q_n) \sim \lambda(\Pi, Q_n) \tag{1}$$

   where $\lambda(\Pi, Q_n)$ represents the effectiveness of Π over the set of gender-neutral queries Q_n, measured using standard retrieval metrics.
2. *Reduce Gender Bias:* The fair ranker $\hat{\Pi}$ must also demonstrate reduced levels of gender bias compared to the state-of-the-art ranker Π. The bias metric $\Psi(\Pi, Q_n)$ quantifies the degree of stereotypical gender representation in the ranked documents for the neutral query set. To ensure fairness, the following condition must be met:

$$\Psi(\hat{\Pi}, Q_n) < \Psi(\Pi, Q_n) \tag{2}$$

   where $\Psi(\Pi, Q_n)$ represents the biases exposed by Π when addressing the set of gender-neutral queries Q_n. This requirement ensures that $\hat{\Pi}$ provides rankings with reduced bias while maintaining its relevance for gender-neutral queries.

The goal of our work in this paper is to develop and propose a *fair neural ranker* that simultaneously satisfies the dual objectives of fairness and effectiveness. Specifically, the ranker aims to minimize $\Psi(\hat{\Pi}, Q_n)$, the measure of gender bias, while ensuring

that $\lambda(\hat{\Pi}, Q_n)$, its ranking effectiveness, remains comparable to $\lambda(\Pi, Q_n)$, the effectiveness of a state-of-the-art ranker. Achieving this balance requires addressing the inherent trade-offs between fairness and performance.

3.2 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:

$$\mathcal{L} = -\mathbb{E}_{\mathcal{T}}\big[\log P(\Phi \mid \mathcal{T}, \mathcal{Y})\big] \tag{3}$$

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

$$P(\Phi \mid \mathcal{T}, \mathcal{Y}) \propto P(\mathcal{Y} \mid \mathcal{T}, \Phi) \cdot P(\Phi) \tag{4}$$

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:

$$\mathcal{L} = -\mathbb{E}_{\mathcal{T}}\big[\log P(\mathcal{Y} \mid \mathcal{T}, \Phi)\big] \tag{5}$$

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 $\mathcal{L}_{\text{Penalty}}$:

$$\mathcal{L}_{\text{Penalty}} = -\mathbb{E}_{\mathcal{T}}\big[\log P(\mathcal{Y} \mid \mathcal{T}, \Phi, \Psi)\big] \tag{6}$$

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 $\mathcal{L}_{\text{Reward}}$:

$$\mathcal{L}_{\text{Reward}} = -\mathbb{E}_{\mathcal{T}}\big[\log P(\mathcal{Y} \mid \mathcal{T}, \Phi, \zeta)\big] \tag{7}$$

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. While penalizing biased documents and rewarding fair ones are conceptually equivalent in that both encourage the model to prioritize fairness, their practical implementation may diverge depending on the context. In real-world scenarios, the choice between these strategies often depends on the structure of the dataset and the availability of well-defined metrics. In some cases, robust bias metrics may exist that enable effective penalization, while in others, fairness metrics may be more accessible or interpretable, making reward-based strategies more practical.

3.3 Fair Neural Rankers

Neural rankers[1] often 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. Pairwise loss function 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 in isolation, this approach models the relative preference between a relevant document d^+ and an irrelevant document d^- for a given query q. The conditional probability for such a ranking approach is expressed as:

$$P(\mathcal{Y} \mid \mathcal{T}, \Phi) = \prod_{(i,j)} P\big((y_i, y_j) \mid (\Phi(q, d_i), \Phi(q, d_j))\big), \tag{8}$$

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

$$P\big((y_i, y_j) \mid (\Phi(q, d_i), \Phi(q, d_j))\big) = \sigma\big(\Phi(q, d_i) - \Phi(q, d_j)\big), \quad \sigma(z) = \frac{1}{1 + \exp(-z)}. \tag{9}$$

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

$$\mathcal{L}_{\text{pairwise}} = -\sum_{(i,j)} \log P\big((y_i, y_j) \mid (\Phi(q, d_i), \Phi(q, d_j))\big). \tag{10}$$

[1] In this section, we present our proposed method based on the pairwise loss function. The methodology and experiments related to the pointwise loss are provided in the appendix.

Expanding this expression gives the following loss function:

$$\mathcal{L}_{\text{pairwise}} = \sum_{(i,j)} \log\big(1 + \exp(-(\Phi(q, d_i) - \Phi(q, d_j)))\big). \tag{11}$$

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:

$$\mathcal{L}_{\text{margin}} = \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - \Phi(q, d^+) + \Phi(q, d^-)). \tag{12}$$

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. In the following, we are going to explain how the penalizing, and rewarding scenarios can be applied on the pairwise loss function.

3.3.1 Penalizing Biased Documents

Penalizing *Biased Irrelevant* Documents. In the training of pairwise neural rankers, the distance between the vector representation of a query, and its irrelevant documents is maximized in the embedding space. In other words, irrelevant documents are expected to be positioned farther from the query in the vector space, resulting in low relevance scores between the query, and the irrelevant document ($\Phi(q, d)$). On the other hand, when irrelevant documents exhibit high levels of bias, their proximity to the query can inadvertently influence the ranking, 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 to signal the model to move these documents farther from the query in the vector space. In other words, **biased irrelevant** documents would need to receive an extra penalty relative to their degree of bias compared to a **fair irrelevant document** because they not only are irrelevant, but also biased. By explicitly discouraging the model from associating biased irrelevant documents with the query, we aim to de-prioritize 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:

$$\Phi_B(q, d^-) = \alpha(\Phi(q, d^-)) + \lambda\Psi(d^-) \tag{13}$$

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. Following the document gender magnitude formulation proposed by Rekabsaz et al.[38], we compute the gender magnitude of a document based on the presence of gender-definitional words. Let $\text{mag}_f(d)$ and $\text{mag}_m(d)$ denote the female and male magnitudes of a document d, respectively, defined as:

$$\text{mag}_g(d) = \begin{cases} 1, & \text{if } \sum_{w \in G_g} \#\langle w, d\rangle > 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{for } g \in \{f, m\} \tag{14}$$

where G_f and G_m are the sets of female and male definitional words, and $\#\langle w, d\rangle$ is the count of word w in document d. The bias score of a document is then defined as the difference between its male and female magnitudes:

$$\Psi(d) = \text{mag}_f(d) - \text{mag}_m(d) \tag{15}$$

A positive value indicates a female bias, a negative value indicates a male bias, and zero indicates no observable gender bias in boolean terms.

While this work focuses on gender bias and uses a keyword-based metric for bias quantification, we emphasize that the proposed framework is not limited to this specific setting. Our approach is designed to be modular with respect to the bias scoring function Ψ. As long as a suitable metric exists to quantify a particular form of bias, be it gender, racial, socioeconomic, or others, it can be integrated into our framework without architectural changes. This flexibility allows our method to be adapted to various types of social biases, including those that may require more nuanced or learned representations.

By increasing the relevance score of the biased irrelevant documents in the loss function, the model is signaled that this document is not far enough from the query in the vector space, so it has to push it farther from the query, 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 *Biased 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. 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:

$$\Phi_B(q, d^+) = \alpha(\Phi(q, d^+)) + \lambda\Psi(d^+) \tag{16}$$

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. Increasing the relevance score of the biased relevant documents, misleads the model that this document is close enough to the query in the vector space, and there is no need to bring it

Algorithm 1 Training of the Ranking Network with the Bias-aware Pair-wise Loss.

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

closer. 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 All Biased 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. As described in Equations 13 and 16. 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.

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

$$\mathcal{L}_{\text{Penalty}} = \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - (\tanh(\Phi(q, d^+)) + \lambda\Psi(d^+)) + (\tanh(\Phi(q, d^-)) + \lambda\Psi(d^-))), \tag{17}$$

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.

The derivative of the penalized loss function with respect to the relevance score $\Phi(q, d)$ is defined as:

$$\frac{\partial \mathcal{L}^{\text{neg}}_{\text{Penalty}}}{\partial \Phi(q, d)} = \begin{cases} 0 & \text{if } l_i < 0 \\ \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} (\tanh^2(\Phi(q, d^+)) - \tanh^2(\Phi(q, d^-))) & \text{if } l_i \geq 0 \end{cases} \tag{18}$$

where the decision boundary l_i is computed as:

$$l_i = \underbrace{(m - \lambda\Psi(q, d^+) + \lambda\Psi(q, d^-))}_{m'} - (\tanh(\Phi(q, d^+)) - \tanh(\Phi(q, d^-))) \tag{19}$$

In this formulation, the bias terms $\Psi(q, d^+)$ and $\Psi(q, d^-)$ act as bias-aware adjustments to the margin in the loss function. The adjusted margin, denoted by m', replaces the original margin m and explicitly accounts for the presence of bias in both the relevant and irrelevant documents. If the relevant document is biased, the bias term $\Psi(q, d^+)$ is large, which reduces the adjusted margin m'. This reduction lowers the likelihood of the loss being positive and contributing to the model's gradient updates. As a result, the model is discouraged from pulling biased relevant documents closer to the query in the embedding space. If the irrelevant document is biased, the term $\Psi(q, d^-)$ has a positive value, which enlarges the margin m'. This makes it more likely for the loss to be positive and of higher magnitude. Consequently, the model is encouraged to push the biased irrelevant document further away from the query representation, resulting in a lower rank in the final output. In summary, this penalized loss selectively suppresses updates for biased relevant documents while amplifying updates that move biased irrelevant documents farther from the query, thereby promoting fairness in the ranking without sacrificing relevance.

Algorithm 1 outlines the training procedure for a ranking network using the bias-aware pairwise loss. The process begins with calculating the gender bias of the documents in the training set ($\Psi(d)$)(Line 2), and initializing the model parameters θ and the bias term b randomly (Line 3). Training proceeds over T iterations, during which each query-document pair (q, d^+, d^-) in the batch is processed iteratively (Line 4-5). 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 6-7). 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 8-9). Using these scores, the bias-aware pairwise loss $\mathcal{L}_{\text{Penalty}}$ 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 11-12), 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 13). 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. It is worth noting that the bias scores used as the penalty term for the relevance scores ($\Psi(d)$) can be precomputed offline (Line 2), thus incurring no additional computational cost during training. Moreover, because the metric is computed using straightforward word-count operations, it imposes negligible computational overhead on the overall framework.

We note that our approach differentiates between irrelevant documents in general and those that are specifically biased. While the model naturally pushes irrelevant documents away from the query, it does not inherently distinguish between biased

and unbiased irrelevants. By applying an additional penalty to biased irrelevant documents, we explicitly ensure they are pushed further down the ranked list relative to their unbiased counterparts. This step helps prevent biased irrelevant content from disproportionately influencing the final output and thereby reduces overall unfairness in the system's rankings.

3.3.2 Rewarding Fair Documents

In contrast to penalizing biased content, rewarding fair documents serves as the dual concept by encouraging the model to preferentially treat documents that exhibit desirable fairness properties. While the underlying goal remains similar, namely, integrating fairness into the learning process, the reward-based regularization introduces distinct subtleties in how relevance scores are adjusted. To avoid redundancy with the previous section, we present this formulation more concisely, while noting that the operational nuances of reward-driven loss regularization warrant careful tuning. For irrelevant documents d^-, the model is expected to assign low relevance scores. However, overly penalizing irrelevant documents that exhibit high fairness may introduce unintended bias against such content. To address this, we adjust the relevance score downward by an amount proportional to the fairness measure $\zeta(d^-)$, thereby reducing the extent to which fair irrelevant documents are pushed away in the embedding space. This adjustment is formalized as:

$$\Phi_R(q, d^-) = \alpha(\Phi(q, d^-)) - \lambda\zeta(d^-) \tag{20}$$

where α is an activation function and λ controls the influence of fairness in the scoring adjustment. We employ the concept of document neutrality proposed by Rekabsaz et al. [36] to assess the extent to which a document provides a balanced representation of a protected attribute, i.e. firness of the document. Specifically, the magnitude of representation for each group $a \in A$ is computed as:

$$\text{mag}_a(d) = \sum_{w \in V_a} \#\langle w, d\rangle \tag{21}$$

where V_a is the set of representative words for group a, and $\#\langle w, d\rangle$ is the count of word w in document d. Based on this, document neutrality $\omega(d)$ is defined as:

$$\zeta(d) = \begin{cases} 1 & \text{if } \sum_{a \in A} \text{mag}_a(d) \leq \tau \\ 1 - \sum_{a \in A} \left| \frac{\text{mag}_a(d)}{\sum_{x \in A} \text{mag}_x(d)} - J_a \right| & \text{otherwise} \end{cases} \tag{22}$$

Here, J_a denotes the expected fair proportion for group a, and τ is a threshold to filter out documents with minimal gendered content.

Similarly, for relevant documents d^+, the objective is to promote those that are not only relevant but also fair. By reducing their relevance scores according to their fairness score $\zeta(d^+)$, we signal the model that these fair documents deserve additional emphasis in the final ranking. This leads to the following fairness-aware adjustment:

$$\Phi_R(q, d^+) = \alpha(\Phi(q, d^+)) - \lambda\zeta(d^+) \tag{23}$$

To operationalize these adjustments within the ranking objective, we define a reward-based loss function that incorporates fairness into the relative comparison between

relevant and irrelevant documents:

$$\mathcal{L}_{\text{Reward}} = \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max\left(0, m - \left(\tanh(\Phi(q, d^+)) - \lambda\zeta(d^+)\right) + \left(\tanh(\Phi(q, d^-)) - \lambda\zeta(d^-)\right)\right) \tag{24}$$

This formulation encourages the model to differentiate documents not just by relevance, but also by their degree of fairness. Training follows the same optimization pipeline as the penalty-based approach, replacing bias $\Psi(\cdot)$ with fairness $\zeta(\cdot)$.

While the preceding description formalizes the reward adjustment, it is equally important to explain how this mechanism shapes training dynamics in practice. The key distinction from the penalty formulation lies in how the gradients are modified. Penalty-based updates suppress the contribution of biased documents by shrinking their gradient magnitude, effectively slowing down the learning of representations that would otherwise amplify bias. In contrast, reward-based regularization acts as a reweighting mechanism that amplifies the gradient contributions of documents exhibiting desirable fairness properties. This does not simply reduce the impact of bias but actively encourages the optimization process to favor fairer content.

The consequence is a shift in the effective decision boundary learned by the model. Margin improvements are no longer distributed uniformly across documents but are preferentially allocated to those whose fairness scores are higher. This means that, during training, the optimization procedure is repeatedly nudged toward ranking outcomes where fair documents accumulate greater representational weight. Mathematically, the additive reward term modifies the loss landscape by tilting gradient trajectories toward regions of parameter space that emphasize fairness without discarding relevance. Whereas the penalty approach introduces steeper "slopes" away from biased regions, the reward approach builds "attractive areas" around fair regions, altering the convergence pathways the optimizer may follow.

From an optimization theory perspective, this implies that the model is guided toward convergence points that reflect a trade-off between maximizing relevance margins and maximizing representational parity. Importantly, this trade-off is not simply the inverse of the penalty case. Penalization acts as a deterrent against certain undesirable updates, while rewarding fairness creates positive reinforcement for desirable updates. These are asymmetric interventions that lead to different local minima, which explains why our empirical results (presented later in the paper) show that reward-based strategies sometimes improve relevance metrics at the cost of weaker bias reduction, while penalty-based strategies achieve the opposite balance.

4 Experiments

4.1 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, i.e., penalizing or rewarding relevant, irrelevant, or both document types, on 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:** *How does the choice of the regularization coefficient (λ) impact the performance of the fairness-aware framework?* We test the best-performing models 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.
- **RQ4:** *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 [36], 2) `CODER`: A transformer-based model that incorporates neutrality regularization [34], 3) `Light-Weight Sampling Strategy (LWS)`: A bias-aware negative sampling approach that trains models to mitigate bias [29].

4.2 Experimental Setup

Datasets and Setup. We conduct our experiments on the MSMARCO passage ranking dataset [39], 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 [40], leveraging its architecture, implementation, and hyperparameter settings to ensure consistency with prior work. OpenMatch utilizes the cross-encoder architecture as its neural ranker; however, it is important to note that our proposed approach is model-agnostic. Alternative architectures—such as ColBERT [41] can equally benefit from the bias-aware loss functions introduced in this work. Each experiment was conducted on a server with two NVIDIA RTX A6000 GPUs (48 GB each), running for approximately two hours. Based on an estimated total power draw of 700 watts, each run consumed around 1.4 kWh of electricity. This corresponds to approximately 0.094 kg CO_2 e in Canada [42]. To ensure the robustness and stability of our method, we report the average performance over five independent runs using different random seeds. For a fair comparison, we follow the official implementation, hyperparameter settings, and training commands provided in the respective papers of each baseline. In line with standard practice, we compare against the best reported configuration for each baseline method, as published by the original authors. Full implementation details and the source code for our work are publicly available on GitHub[2].

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

- **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. [38], consisting of 1,765 queries annotated by three Amazon Mechanical Turk workers. These queries were derived from an initial pool of MSMARCO development set 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

[2] https://github.com/shirinssalehi/LossRegularizationJournal

Table 1 Percentage of change in performances of the model across the six proposed scenarios using the pairwise loss function with the "BERT-mini" base model on the 215-query dataset [36]. We performed statistical significance tests on all the values reported in the table. Values marked with "*" indicate changes that are statistically significant at 95% confidence.

		Cut-off@10					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Base Model		0.18	0.38	0.17	0.13	0.99	0.81
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
Base Model		0.18	0.30	0.13	0.10	0.74	0.84
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%*

the rankings. These queries are designed to evaluate the ranker's ability to mitigate biases in contexts with inherent societal sensitivity [36].

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 [39]. To assess gender bias, we employ three complementary metrics:

- **Average Rank Bias (ARaB)** [38]: 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. Because this metric quantifies gender bias in the ranking results, lower values signify better performance.
- **NFaiRR** [36]: A document-level fairness metric designed to evaluate ranking fairness, where higher values indicate more equitable rankings with respect to gender-neutral queries. Therefore, higher values for this metric is desirable.
- **Linguistic Inquiry and Word Count (LIWC)** [43]: This metric examines the frequency of gendered terms in retrieved text by counting references to male and female pronouns, quantifying the linguistic attributes of retrieved content. Since the metric captures the degree of gender bias in ranking outcomes, smaller scores indicate superior performance.

4.3 Findings

Findings for RQ1. This research question is designed as an ablation study to isolate the effects of fairness constraints under different configurations. Specifically, we examine how applying fairness-aware regularization only to relevant documents, only to irrelevant documents, or to both affects retrieval effectiveness and bias mitigation. We also compare penalty-based versus reward-based mechanisms to assess their independent contributions. Tables 1 and 2 present the percentage of change in performance (MRR; positive numbers are desirable as indicate the performance is increased), bias (ARaB-TC, ARaB-TF, ARaB-Bool, and LIWC; negative numbers are desirable as indicate the bias is reduced), and fairness (NFaiRR; positive numbers are desirable

Table 2 Percentage of change in 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 [38].We performed statistical significance tests on all the values reported in the table. Values marked with "*" indicate changes that are statistically significant at 95% confidence.

		Cut-off@10					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Base Model		0.28	0.24	0.11	0.10	1.60	0.72
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
Base Model		0.28	0.19	0.09	0.08	1.31	0.74
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%*

Table 3 Percentage of change in 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 [36]. We performed statistical significance tests on all the values reported in the table. Values marked with "*" indicate changes that are statistically significant at 95% confidence.

		Cut-off@10					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Base Model		0.18	0.29	0/13	0.10	0.98	0.82
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
Base Model		0.19	0.25	0.11	0.08	0.77	0.84
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%*

as indicate the fairness is increased) after applying 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 methods often results 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*.

Table 4 Percentage of change in 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 [38]. We performed statistical significance tests on all the values reported in the table. Values marked with "*" indicate changes that are statistically significant at 95% confidence.

		Cut-off@10					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Base Model		0.30	0.13	0.07	0.06	1.60	0.72
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
Base Model		0.30	0.12	0.06	0.05	1.33	0.74
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%*

The results in **RQ1** suggest that penalty-based strategies tend to achieve more consistent bias mitigation, albeit with slight compromises in ranking effectiveness.

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 3-4. 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 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 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 *Electra-small*, where the ranking performance sometimes degrades more significantly than with *BERT-mini*.

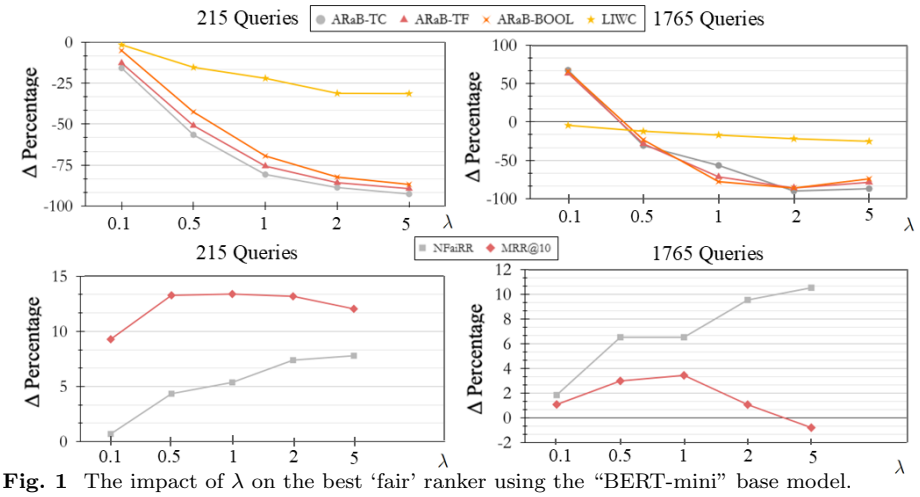

Fig. 1 The impact of λ on the best 'fair' ranker using the "BERT-mini" base model.

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.

> The results in **RQ2** 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.

Findings for RQ3. 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 penalizing the biased relevant documents scenario in this research question, in that this scenario shows the best bias-performance trade-off across all of the six scenarios. It improves the performance, and fairness, and reduces bias for both datasets as shown in Tables 1, and 2. 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). 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, 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.

> The findings in **RQ3** show the regularization coefficient λ can be tuned to effectively control the tension between bias mitigation and retrieval effectiveness.

Findings for RQ4. 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

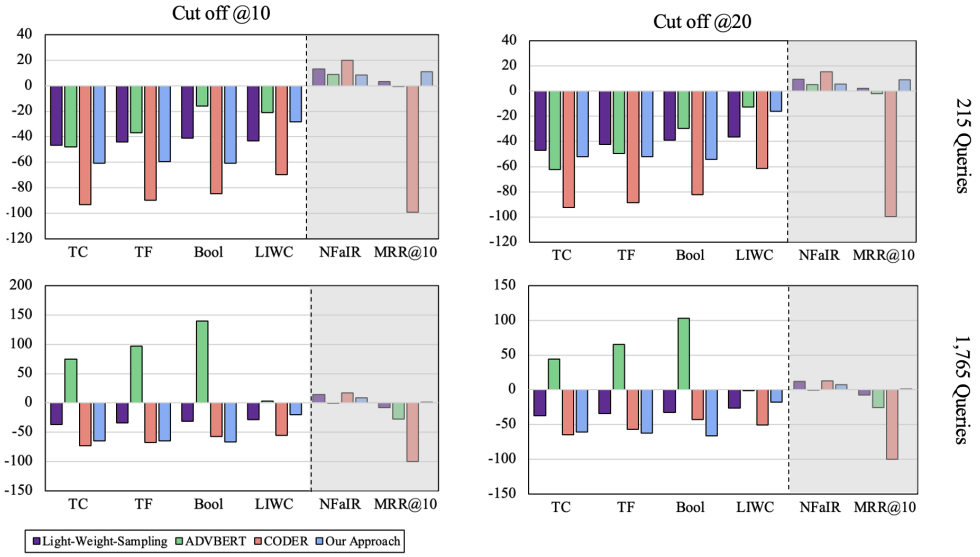

Fig. 2 Comparison of our proposed approach with the state-of-the-art methods on the 215-query and 1,765-query datasets based on the best 'fair' pairwise ranker using the "BERT-mini" base model. We note negative values on the left side of each figure and positive values on the right side are desirable. We note that all the changes reported in the figure are statistically significant at the 95% confidence level, except for the MRR on the 1765-query dataset.

chosen based on Tables 1 and 2 (penalizing biased relevant documents) as representative of our proposed approach. As illustrated in Figure 2, 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 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.

The results for **RQ4** indicate our approach not only surpasses the state-of-the-art methods in bias reduction but also ensures that fairness enhancements do not come at the cost of retrieval effectiveness by retaining comparable rates of retrieval effectiveness.

5 Limitations and Future Work

This study is bounded by four interrelated limitations. First, we focus exclusively on gender-based disparities. While this focus allows for a clear empirical analysis, it leaves other important attributes, such as race, ethnicity, age, or disability status, unexamined. Second, all benchmark corpora used in this study annotate gender only as male or female, and existing bias-quantification metrics follow the same binary structure. Consequently, our findings do not account for non-binary, gender-fluid, or otherwise gender-diverse populations. Third, our analysis relies solely on English-language collections. Language-specific vocabulary, morphology, and cultural context can shape both the formation and detection of bias; therefore, our conclusions may not fully generalize to other languages. Fourth, we used the ARaB-BOOL metric to assess gender bias, which outputs only a binary label indicating whether a document is biased or not. This boolean measure limits our ability to capture the degree or magnitude of bias present in documents.

Future work can address these limitations in several ways. The proposed debiasing framework itself is attribute-agnostic, and could be extended beyond gender to other sensitive attributes, provided that reliable labels and appropriate fairness metrics are available for those dimensions. Applying the framework to other attributes and languages will require careful attention to the social context, data quality, and metric validity specific to each new setting. Furthermore, future studies could incorporate more nuanced bias metrics that produce graded or continuous values, allowing for a richer assessment of bias intensity. We have examined our methodology using pairwise and listwise loss functions. While the current framework is not directly applicable to listwise loss functions, an important future direction would be to adapt and extend the proposed method for use with listwise loss functions. Expanding along these directions would improve the generalizability, inclusivity, and interpretability of the framework.

6 Concluding Remarks

In this paper, 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 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.

References

[1] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. In: International Conference on Learning Representations (ICLR) (2020)

[2] Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
[3] Yates, A., Nogueira, R., Lin, J.: Pretrained transformers for text ranking: Bert and beyond. In: WSDM 2021, pp. 1154–1156 (2021)
[4] Robertson, S., Zaragoza, H., *et al.*: The probabilistic relevance framework: Bm25 and beyond. FnTIR **3**(4), 333–389 (2009)
[5] Wei, C.-P., Hu, P.J.-H., Tai, C.-H., Huang, C.-N., Yang, C.-S.: Managing word mismatch problems in information retrieval: A topic-based query expansion approach. Journal of Management Information Systems **24**(3), 269–295 (2007)
[6] Wang, X., MacAvaney, S., Macdonald, C., Ounis, I.: Generative query reformulation for effective adhoc search. arXiv preprint arXiv:2308.00415 (2023)
[7] Zhao, W.X., Liu, J., Ren, R., Wen, J.-R.: Dense text retrieval based on pretrained language models: A survey. ACM TOIS **42**(4), 1–60 (2024)
[8] Bigdeli, A.: Exploration and mitigation of stereotypical gender biases in information retrieval systems. Master's thesis, TorontoMet University (2021)
[9] Bigdeli, A., Arabzadeh, N., Seyersalehi, S., Zihayat, M., Bagheri, E.: On the orthogonality of bias and utility in ad hoc retrieval. In: SIGIR 2021 (2021)
[10] Seyedsalehi, S., Bigdeli, A., Arabzadeh, N., Mitra, B., Zihayat, M., Bagheri, E.: Bias-aware fair neural ranking for addressing stereotypical gender biases. In: EDBT, pp. 2–435 (2022)
[11] Krieg, K., Parada-Cabaleiro, E., Medicus, G., Lesota, O., Schedl, M., Rekabsaz, N.: Grep-biasir: A dataset for investigating gender representation bias in information retrieval results. In: CHIIR 2023, pp. 444–448 (2023)
[12] Kopeinik, S., Mara, M., Ratz, L., Krieg, K., Schedl, M., Rekabsaz, N.: Show me a" male nurse"! how gender bias is reflected in the query formulation of search engine users. In: CHI 2023, pp. 1–15 (2023)
[13] Krieg, K., Parada-Cabaleiro, E., Schedl, M., Rekabsaz, N.: Do perceived gender biases in retrieval results affect relevance judgements? In: International Workshop on Algorithmic Bias in Search and Recommendation, pp. 104–116 (2022)
[14] Dai, S., Xu, C., Xu, S., Pang, L., Dong, Z., Xu, J.: Unifying bias and unfairness in information retrieval: A survey of challenges and opportunities with large language models. arXiv preprint arXiv:2404.11457 (2024)
[15] Dai, S., Xu, C., Xu, S., Pang, L., Dong, Z., Xu, J.: Bias and unfairness in information retrieval systems: New challenges in the llm era. In: KDD 2024, pp. 6437–6447 (2024)
[16] Bolukbasi, T., Chang, K., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Neural Information Processing Systems (2016)
[17] Basta, C., Costa-jussà, M.R., Casas, N.: Evaluating the underlying gender bias in contextualized word embeddings. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing (2019). https://doi.org/10.18653/v1/w19-3805 . https://doi.org/10.18653/v1/w19-3805
[18] Bigdeli, A., Arabzadeh, N., Bagheri, E.: Learning to jointly transform and rank difficult queries. In: ECIR 2024, pp. 40–48 (2024). Springer
[19] Huang, S., Zhou, J., Feng, H., Zhou, D.-X.: Generalization analysis of pairwise learning for ranking with deep neural networks. Neural Computation **35**(6), 1135–1158 (2023)
[20] Cerrato, M., Köppel, M., Segner, A., Esposito, R., Kramer, S.: Fair pairwise learning to rank. In: DSAA 2020, pp. 729–738 (2020)

[21] Moradi, R., Berangi, R., Minaei, B.: A survey of regularization strategies for deep models. Artificial Intelligence Review **53**(6), 3947–3986 (2020)
[22] Navigli, R., Conia, S., Ross, B.: Biases in large language models: origins, inventory, and discussion. ACM Journal of Data and Information Quality **15**(2), 1–21 (2023)
[23] Raza, S., Garg, M., Reji, D.J., Bashir, S.R., Ding, C.: Nbias: A natural language processing framework for bias identification in text. Expert Systems with Applications **237**, 121542 (2024)
[24] Caliskan, A., Ajay, P.P., Charlesworth, T., Wolfe, R., Banaji, M.R.: Gender bias in word embeddings. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (2022). https://doi.org/10.1145/3514094.3534162 . https://doi.org/10.1145/3514094.3534162
[25] Zhao, J., Zhou, Y., Li, Z., Wang, W., Chang, K.-W.: Learning gender-neutral word embeddings. arXiv preprint arXiv:1809.01496 (2018)
[26] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM CSUR **54**(6), 1–35 (2021)
[27] Rahmani, H.A., Craswell, N., Yilmaz, E., Mitra, B., Campos, D.: Synthetic test collections for retrieval evaluation. In: SIGIR 2024, pp. 2647–2651 (2024)
[28] Bigdeli, A., Arabzadeh, N., Zihayat, M., Bagheri, E.: Exploring gender biases in information retrieval relevance judgement datasets. In: ECIR 2021, pp. 216–224 (2021). Springer
[29] Bigdeli, A., Arabzadeh, N., Seyedsalehi, S., Zihayat, M., Bagheri, E.: A lightweight strategy for restraining gender biases in neural rankers. In: ECIR 2022, pp. 47–55 (2022)
[30] Seyedsalehi, S., Bigdeli, A., Arabzadeh, N., Zihayat, M., Bagheri, E.: Addressing gender-related performance disparities in neural rankers. In: SIGIR 2022, pp. 2484–2488 (2022)
[31] Arduini, M., Noci, L., Pirovano, F., Zhang, C., Shrestha, Y.R., Paudel, B.: Adversarial learning for debiasing knowledge graph embeddings. arXiv preprint arXiv:2006.16309 (2020)
[32] Qian, C., Feng, F., Wen, L., Ma, C., Xie, P.: Counterfactual inference for text classification debiasing. In: IJCNLP, pp. 5434–5445 (2021)
[33] Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., Goldberg, Y.: Null it out: Guarding protected attributes by iterative nullspace projection. arXiv preprint arXiv:2004.07667 (2020)
[34] Zerveas, G., Rekabsaz, N., Cohen, D., Eickhoff, C.: Mitigating bias in search results through contextual document reranking and neutrality regularization. In: SIGIR 2022, pp. 2532–2538 (2022)
[35] Zerveas, G., Rekabsaz, N., Cohen, D., Eickhoff, C.: Coder: An efficient framework for improving retrieval through contextual document embedding reranking. arXiv preprint arXiv:2112.08766 (2021)
[36] Rekabsaz, N., Kopeinik, S., Schedl, M.: Societal biases in retrieved contents: Measurement framework and adversarial mitigation for bert rankers. arXiv preprint arXiv:2104.13640 (2021)
[37] Liu, T.-Y., *et al.*: Learning to rank for information retrieval. Foundations and Trends® in Information Retrieval **3**(3), 225–331 (2009)
[38] Rekabsaz, N., Schedl, M.: Do neural ranking models intensify gender bias? In: SIGIR 2020, pp. 2065–2068 (2020)

[39] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: Ms marco: A human generated machine reading comprehension dataset. In: CoCo@ NIPS (2016)
[40] Liu, Z., Zhang, K., Xiong, C., Liu, Z., Sun, M.: Openmatch: An open source library for neu-ir research. In: SIGIR 2021, pp. 2531–2535 (2021)
[41] Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., Zaharia, M.: Colbertv2: Effective and efficient retrieval via lightweight late interaction. arXiv preprint arXiv:2112.01488 (2021)
[42] The Atmospheric Fund: Ontario Electricity Emissions Factors and Guidelines 2024. https://taf.ca/custom/uploads/2024/06/TAF-Ontario-Emissions-Factors-2024.pdf. Accessed September 17, 2025 (2024)
[43] Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic Inquiry and Word Count. Lawerence Erlbaum Associates, Mahwah, NJ (2001)
[44] Déjean, H., Clinchant, S., Formal, T.: A thorough comparison of cross-encoders and llms for reranking splade. arXiv preprint arXiv:2403.10407 (2024)
[45] Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., Grave, E.: Unsupervised dense information retrieval with contrastive learning. arXiv preprint arXiv:2112.09118 (2021)
[46] Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over bert. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 39–48 (2020)

A Fair pointwise ranking

Thus far, our empirical analysis has focused on pairwise ranking losses. In this appendix we clarify the extent to which the proposed bias-mitigation strategy can be transferred to other loss families, namely listwise and pointwise formulations.

Listwise objectives [37] compare an entire slate of documents returned for a query. Because our method injects a bias-penalty (or fairness reward) directly into each document's relevance score, the signals from biased and fair documents within the same slate are aggregated before the loss is computed. Positive and negative adjustments can therefore cancel out, producing a weak or ambiguous gradient that makes it difficult for the model to learn which individual documents are biased. Unlike the pairwise setting, where each relevant document is repeatedly contrasted with many irrelevant ones, the listwise setting typically exposes each document only once per query, leaving the model with too little information to disentangle fair from biased items. The cancellation issue is far less pronounced for pairwise objectives, because every relevant document is compared against multiple irrelevant counterparts. Across these comparisons the model observes enough variation to infer whether bias arises from the relevant or the irrelevant side, allowing the penalty term to guide learning effectively.

Pointwise approaches [37] score one document at a time, so the bias-penalty or reward directly modifies the single relevance score being optimised. This one-to-one correspondence creates an unambiguous training signal, ensuring that the model is explicitly informed whenever a document carries undesirable bias.

In summary, while our method extends naturally to pointwise objectives, additional adjustments, such as slate-level debiasing terms or document-specific weighting are required before it can be reliably applied to listwise losses. We leave this adaptation

for future work. In the next subsection we report experimental results demonstrating that our framework transfers cleanly to the pointwise setting.

A.1 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:

$$P(\mathcal{Y} \mid \mathcal{T}, \Phi) = \prod_i P(y_i \mid \Phi(q, d_i)) \tag{25}$$

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:

$$P(y_i \mid \Phi(q, d_i)) = \frac{1}{1 + e^{-\Phi(q,d_i)}} \tag{26}$$

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:

$$\mathcal{L}_{\text{pointwise}} = -\sum_i \log P(y_i \mid \Phi(q, d_i)) = \sum_i \log \left(1 + e^{-y_i \cdot \Phi(q,d_i)}\right) \tag{27}$$

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:

$$\mathcal{L}_{\text{pointwise}} = \sum_i \left(\frac{1}{1 + e^{-\Phi(q,d_i)}} - y_i\right)^2 \tag{28}$$

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.

A.1.1 Penalizing Documents

To incorporate fairness into the pointwise loss function, we introduce a *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:

$$\mathcal{L}_{\text{Penalty}} = \sum_i \left[\left(\frac{1}{1 + e^{-(\Phi(q,d_i)+(1-y_i)\cdot\Psi(d_i)+y_i\cdot\Psi(d_i))}} \right) - y_i \right]^2 \tag{29}$$

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 formulation supports three distinct scenarios:

1. For *penalizing irrelevant documents* (Section 3.3.1), 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 3.3.1), 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 3.3.1), 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:

$$\frac{\partial \mathcal{L}_{\text{Penalty}}}{\partial \Phi(q,d_i)} = 2\big[\sigma(\Phi(q,d_i)+\Psi(d_i)) - y_i\big]\cdot\sigma(\Phi(q,d_i)+\Psi(d_i))\cdot\big(1-\sigma(\Phi(q,d_i)+\Psi(d_i))\big), \tag{30}$$

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 2 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 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

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

(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.

A.1.2 Rewarding Documents

To incorporate fairness rewards into the pointwise loss function, we introduce a *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:

$$\mathcal{L}_{\text{Reward}} = \sum_i \left[\left(\frac{1}{1+e^{-(\Phi(q,d_i)-(1-y_i)\cdot\zeta(d_i)-y_i\cdot\zeta(d_i))}}\right) - y_i\right]^2 \tag{31}$$

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 framework integrates the following cases:

1. When *rewarding irrelevant documents* (Section 3.3.2), 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 3.3.2), 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 3.3.1), 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.

A.1.3 Experimental Results

Tables 5 and 6 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 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.

Table 5 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 [36]. We performed statistical significance tests on all the values reported in the table. Values marked with "*" indicate changes that are statistically significant at 95% confidence.

		Cut-off@10					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Base Model		0.19	0.13	0.05	0.03	0.61	0.89
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
Base Model		0.19	0.11	0.04	0.03	0.52	0.90
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%*

In Tables 7, and 8, we 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. 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 effectiveness. For instance, *applying penalty*

Table 6 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 [38]. We performed statistical significance tests on all the values reported in the table. Values marked with "*" indicate changes that are statistically significant at 95% confidence.

		Cut-off@10					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Base Model		0.28	0.14	0.05	0.04	1.25	0.79
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
Base Model		0.29	0.12	0.05	0.04	1.09	0.80
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%*

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 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 *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.

In tables 9, and 10, we 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 highlight notable differences in the performance of the pairwise and pointwise loss functions under the proposed bias mitigation scenarios. These observations

Table 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 [36]. We performed statistical significance tests on all the values reported in the table. Values marked with "*" indicate changes that are statistically significant at 95% confidence.

		Cut-off@10					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Base Model		0.22	0.15	0.06	0.04	0.61	0.89
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
Base Model		0.2	0.11	0.04	0.03	0.53	0.90
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 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 [38]. We performed statistical significance tests on all the values reported in the table. Values marked with "*" indicate changes that are statistically significant at 95% confidence.

		Cut-off@10					
		MRR	ARaB-TC	ARaB-TF	ARaB-Bool	LIWC	NFaIRR
Base Model		0.33	0.31	0.13	0.11	1.25	0.80
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
Base Model		0.33	0.29	0.13	0.11	1.09	0.80
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%*	1.29%*
	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%*

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

Table 9 Comparison of the best-performing pairwise and pointwise approaches using the "BERT-mini" base model on the 215-query dataset [36].

			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 10 Comparison of the best-performing pairwise and pointwise approaches using the "BERT-mini" base model on the 1765-query dataset [38].

			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

Figure 3 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 pointwise models reported in Tables 9 and 10. As shown in the figure, 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). 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, 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.

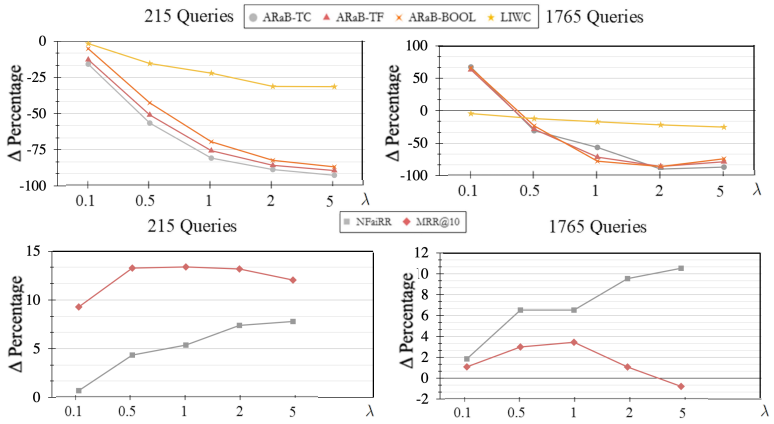

Fig. 3 The impact of varying the value of λ on the performance of the best 'fair' pointwise and pairwise ranker using the "BERT-mini" base model. These rankers are the same as those reported in Tables 9 and 10.

B Applicability of Our Method Across Architectures

In this work, we have employed the cross-encoder architecture as our primary model for re-ranking. In this appendix, we demonstrate how our proposed method can also be extended to alternative neural ranking architectures, namely bi-encoders and late-interaction models such as ColBERT with only minimal modifications.

In cross-encoders, the query q and document d are concatenated and jointly encoded by a large language model (LLM) [44]:

$$h = \text{Encoder}([q; d]), \tag{32}$$

where $[q; d]$ denotes the concatenated sequence. The final relevance score is then computed using the [CLS] token representation:

$$s(q, d) = f(h_{[\text{CLS}]}). \tag{33}$$

Since token-level interactions are computed across both segments within the transformer layers, this approach is highly expressive but computationally expensive. Cross-encoders are therefore most suitable for re-ranking after an initial candidate set is retrieved with a lightweight ranker such as BM25 [4].

Bi-encoders process the query and document independently [45]:

$$h_q = \text{Encoder}_q(q), \quad h_d = \text{Encoder}_d(d). \tag{34}$$

The relevance score is obtained via similarity between the two vector representations, often cosine similarity:

$$s(q, d) = \cos(h_q, h_d). \tag{35}$$

This formulation is computationally efficient, as query and document embeddings can be precomputed and reused. However, it captures fewer fine-grained token-level interactions than a cross-encoder.

ColBERT represents a hybrid approach. Instead of reducing queries and documents to single vectors, it encodes them into token-level embeddings [46]:

$$H_q = \{h_{q_1}, \ldots, h_{q_m}\}, \quad H_d = \{h_{d_1}, \ldots, h_{d_n}\}. \tag{36}$$

At scoring time, token-level similarities are computed, and for each query token, the maximum similarity with any document token is taken. The final score is then:

$$s(q,d) = \sum_{i=1}^{m} \max_j \cos(h_{q_i}, h_{d_j}). \tag{37}$$

This design achieves a favorable balance between computational efficiency and retrieval effectiveness.

Our proposed fairness-aware adjustment operates at the relevance scoring layer, regardless of architecture. Given a base score $s(q,d)$, we adjust it with either a penalty based on bias $\Psi(d)$ or a reward based on fairness $\zeta(d)$:

$$s'(q,d) = \begin{cases} s(q,d) \mp \lambda\, \Psi(d), & \text{(penalty formulation)} \\ s(q,d) \pm \lambda\, \zeta(d), & \text{(reward formulation)} \end{cases} \tag{38}$$

where λ controls the strength of the adjustment.

Since all three model families (cross-encoder, bi-encoder, ColBERT) ultimately reduce to computing a scalar relevance score, our fairness-aware adjustment may be seamlessly integrated into any of them. The difference lies only in how the base score $s(q,d)$ is computed; the post-hoc adjustment mechanism remains identical across architectures.