# Self-Paced Fair Ranking with Loss as a Proxy for Bias


## Abstract

Neural rankers often mirror societal biases in training data, such as gender. Prior work typically requires access to protected-attribute labels or model changes, limiting applicability. We propose a simple, model-agnostic approach that uses the model's own loss values as a proxy for bias. Through self-paced learning, our method first prioritizes lower-loss (and less biased) examples, then gradually introduces harder (more biased) ones. This loss-aware curriculum reduces reliance on biased samples without demographic annotations. We prove the gender loss gap decreases monotonically during training and show on MS MARCO that our method reduces bias while maintaining or improving ranking effectiveness, outperforming strong baselines.


## 1 Introduction

Neural ranking models have become central to modern information retrieval systems due to their ability to learn deep semantic relationships between queries and documents [1–3, 10, 12, 13, 26, 27, 30]. These models, often based on pre-trained Transformers, have significantly improved ranking effectiveness across benchmark datasets [15, 17]. However, recent studies have demonstrated that such models are not immune to societal biases present in their training data, particularly with respect to protected attributes like gender, occupation, and race [7, 20]. For instance, ranking models can disproportionately favor content associated with one gender group, leading to uneven treatment of users and biased exposure of documents to different gender groups. This challenge has motivated a growing body of research that aims to build fairer ranking systems without compromising on ranking quality. To this end, researchers have proposed a variety of bias mitigation strategies [14, 19, 24, 29]. For example, representation learning methods have focused on explicitly separating protected attributes from semantic information during training, encouraging the model to rely on content rather than demographic signals when scoring relevance [24]. Others have explored adversarial training or regularization techniques to suppress the effect of gender and other attributes in the learned embeddings [14, 29]. More recently, curriculum learning has emerged as an effective strategy for reducing bias by controlling how the model is exposed to biased examples during training. By prioritizing low-bias samples in earlier epochs, models have shown to achieve both improved fairness and competitive effectiveness [5, 24].

Although recent methods for mitigating bias in neural ranking models have shown promise, they often depend on sensitive or costly resources such as protected-attributes, external bias measurements, or changes to model architecture [19, 23]. While effective, such interventions are difficult to scale in real-world ranking systems. This raises the foundational question of how bias can be reduced without relying on external demographic annotations. More specifically, can a model's own learning signals, such as loss values, serve as a proxy for bias mitigation?

Prior work has established connections between early-stage loss dynamics and factors such as dataset structure, semantic coherence, and lexical features [4], each of which can differentially affect population subgroups. From another perspective, bias in model predictions frequently emerges as a skewed distribution of attributes, where underrepresented groups are systematically harder for the model to learn. Assuming that harder samples correspond to higher loss values, loss can serve as an indicator along two dimensions: (1) a high-loss sample may be intrinsically challenging due to its lexical or semantic complexity, or (2) the sample may reflect bias, which makes it systematically harder for the model to learn and thus produces elevated loss values. Supporting this hypothesis, Seyedsalehi et al. [23] have demonstrated a misalignment between the similarity of male and female queries (gendered queries) to their relevant documents in large-scale datasets such as MS MARCO [16]. Specifically, male queries tend to enjoy higher similarity scores with their associated relevant documents compared to their female counterparts (we have replicated this finding and report it: https://anonymous.4open.science/r/SPL-bias-727D/plots/loss_gap.png). Given neural rankers optimize a loss function directly over query–document similarity, gender bias embedded in the training data becomes encoded in the model's loss. This implies that loss is not only a marker of learning difficulty but also a signal of genderedness and possibly societal bias in the data. This reframing positions loss as a dual-purpose signal: it measures difficulty while also serving as a practical proxy for bias, such as gender disparities, that influence learning.

Motivated by this observation, we propose an alternative strategy in which the model's own loss values, rather than external bias scores or demographic annotations, are used to guide the training schedule in a way that implicitly reduces bias. Our central idea is to treat the loss as a signal of sample reliability, prioritizing low-loss query–document pairs early in training, where the semantic match is clearer and less likely to be shaped by confounding bias. As training continues, higher-loss pairs are gradually introduced, allowing the model to generalize to more difficult cases after establishing a stable semantic foundation. To this end, we introduce a *lightweight self-paced curriculum learning* approach that dynamically adjusts the sampling weights of training instances based on their current loss values. Unlike prior debiasing methods, our approach requires no protected-attribute labels or precomputed bias scores. It is fully model-driven and applicable to any neural ranker architecture. We evaluate the method on the MS MARCO passage ranking dataset, following the same experimental protocol as [24]. Our results show that using loss as a sampling proxy leads to significant reductions in gender bias, measured by various bias metrics including ARaB [20], NFaiRR [19], and LIWC[18], while maintaining or improving ranking effectiveness.

## 2 Problem Formulation

We first formalize the neural ranking setup that underlies our approach. Let $\mathcal{Q} = \{q_1, \ldots, q_N\}$ denote a set of queries and $\mathcal{D} = \{d_1, \ldots, d_M\}$ a collection of candidate documents. A ranker with parameters $\theta \in \mathbb{R}^d$ assigns a real-valued relevance score $s_\theta(q, d) : \mathcal{Q} \times \mathcal{D} \longrightarrow \mathbb{R}$. Training typically relies on triples $\mathcal{T} = \{(q, d^+, d^-)\}$, where $d^+$ is relevant and $d^-$ is non-relevant for query $q$. The standard pairwise loss is defined as

$$\ell_\theta(q, d^+, d^-) = \max\Big\{0,\ m - s_\theta(q, d^+) + s_\theta(q, d^-)\Big\}, \qquad (1)$$

where $m > 0$ is a fixed margin. This loss enforces that relevant documents are ranked above non-relevant ones by at least a margin.

**Self-Paced Learning.** Building on this formulation, we incorporate self-paced learning to mitigate bias during training. Self-paced learning assigns weights to training instances based on their difficulty. Given a per-example difficulty score $d(x)$, a weighting function $w(d(x), \lambda)$ determines each instance's contribution, where the parameter $\lambda \in (0, 1)$ gradually relaxes to admit harder samples.

In our approach, we adopt $d(x) = \ell_\theta(x)$, i.e., the model's own loss, as the difficulty signal. This creates a loss-aware curriculum where low-loss examples dominate early stages of training, while higher-loss ones are introduced progressively. Intuitively, loss reflects both sample difficulty and potential bias where examples that are systematically harder to learn, such as those from certain gender groups, tend to incur higher loss. By emphasizing low-loss examples at the beginning, the model implicitly focuses on less biased data, postponing exposure to biased, high-loss instances until later in training.

**Learning Objective.** With this curriculum in place, our objective is to learn parameters that maximize relevance while reducing gender disparities. The self-paced learning objective is defined as:

$$\mathcal{L}(\theta, \lambda) = \mathbb{E}_{x \sim \mathcal{T}}\Big[w\big(\ell_\theta(x), \lambda\big)\, \ell_\theta(x)\Big]. \tag{2}$$

Training proceeds with a schedule $\lambda_0 > \lambda_1 > \cdots \to 0$, such that the model gradually incorporates increasingly difficult (and possibly more biased) samples. The optimal parameters are given:

$$\theta^* = \arg\min_\theta \; \mathcal{L}(\theta, \lambda_T), \tag{3}$$

where $\lambda_T$ is the final curriculum stage. In this setup, samples most associated with bias (i.e., high-loss examples) are down-weighted in early training, reducing their undue influence. As $\lambda$ decreases, these harder cases are introduced gradually, ensuring the model learns to handle them without letting bias dominate. This weighted optimization therefore balances ranking effectiveness with fairness.

It is important to emphasize that we do not claim that all high-loss samples are biased. Some may simply be inherently difficult due to lexical or semantic complexity. Our hypothesis is more nuanced where we argue that biased samples are often harder for the model to learn and therefore more likely to yield high loss values. Prioritizing low-loss samples at the beginning reduces the likelihood of overrepresenting biased examples, allowing the model to first establish a balanced representation space.

Finally, it is worth noting that training neural rankers typically involves fine-tuning a pretrained language model, which already encodes semantic knowledge. As a result, early-stage loss values are not random but already carry information about both example difficulty and potential biases present in the data. Fine-tuning with self-paced learning therefore mitigates biases originating from both training data and also inherited from pretrained representations.

## 3 Methodology

We now describe our proposed training framework for reducing gender bias in neural rankers using a loss-aware self-paced learning scheme. The key idea is to modulate training by controlling the influence of each sample based on its current loss, encouraging the model to learn from lower-loss (potentially less biased) examples first, and gradually incorporating more difficult instances over time.

**Adaptive curriculum.** Let $\mathcal{B}_t \subset \mathcal{T}$ be a mini-batch at step $t$, and let $\ell_{\theta_t}(x)$ denote the margin loss for triple $x \in \mathcal{B}_t$, as defined in Eq. (1). To enable self-paced weighting, we first normalize the loss values within each batch to the range $[0, 1]$ using min–max normalization:

$$\hat{\ell}_t(x) = \frac{\ell_{\theta_t}(x) - \min_{z \in \mathcal{B}_t} \ell_{\theta_t}(z)}{\max_{z \in \mathcal{B}_t} \ell_{\theta_t}(z) - \min_{z \in \mathcal{B}_t} \ell_{\theta_t}(z) + \varepsilon},$$

where $\varepsilon > 0$ is a small constant for numerical stability. This normalization ensures that the difficulty of examples is measured relative to their peers in the same batch, making the curriculum adaptive to training dynamics. Each sample is then assigned a weight based on its normalized loss:

$$w_t(x) = 1 - \lambda_t\, \hat{\ell}_t(x), \qquad 0 < \lambda_t \le 1, \tag{4}$$

where $\lambda_t$ is a curriculum parameter that determines the influence of the loss-based reweighting at training step $t$. Intuitively, lower-loss examples receive weights closer to 1 and are emphasized more during optimization. As training progresses, we linearly decay $\lambda_t$:

$$\lambda_t = \lambda_0 \left(1 - \frac{t}{T}\right),$$

where $\lambda_0$ is the initial curriculum strength and $T$ is the total number of training steps. This decay gradually flattens the curriculum, allowing harder (higher-loss) examples to play a more prominent role in later stages of training.

**Weighted ranking objective.** Given the sample weights $w_t(x)$, the self-paced training objective for the mini-batch $\mathcal{B}_t$ becomes a weighted sum of losses:

$$\mathcal{L}_t = \frac{1}{|\mathcal{B}_t|} \sum_{x \in \mathcal{B}_t} w_t(x)\, \ell_{\theta_t}(x). \tag{5}$$

This objective biases the gradient updates toward examples with lower loss, enabling the model to build its semantic understanding from more confident training signals.

**Theoretical Analysis.** The focus in our work is specifically on reducing stereotypical gender biases. Therefore, we provide a theoretical analysis in terms of gender groups. Let us assume each document is assigned a gender label $g(d) \in \{\text{female}, \text{male}\}$ obtained via the proposed metrics in [21]. We partition $\mathcal{T}$ into

$$\mathcal{T}_{\text{f}} = \{x \in \mathcal{T} \mid g(d^+) = \text{female}\}, \qquad \mathcal{T}_{\text{m}} = \{x \in \mathcal{T} \mid g(d^+) = \text{male}\}.$$

The **gender loss gap** is defined as:

$$G(\theta) = \mathbb{E}_{x \sim \mathcal{T}_{\text{f}}}[\ell_\theta(x)] \; - \; \mathbb{E}_{x \sim \mathcal{T}_{\text{m}}}[\ell_\theta(x)]. \tag{6}$$

We now present a theoretical analysis showing that our loss-aware self-paced learning approach monotonically reduces the gender loss gap. Specifically, we show that weighting examples by loss consistently decreases the disparity across male- and female-associated samples, as measured by $|G_t|$. Let $\theta_t$ be the model parameters at training step $t$, and let $G_t = G(\theta_t)$ denote the gender loss gap defined in Eq. (6). Our goal is to show that the sequence $\{|G_t|\}$ is non-increasing under the update rule defined by our training objective.

Lemma 3.1. *Let $x = (q, d^+, d^-)$ be a training triple and define the sign indicator:*

$$s(x) = \mathbb{I}\{x \in \mathcal{T}_f\} - \mathbb{I}\{x \in \mathcal{T}_m\},$$

*which takes value $+1$ for examples with female-associated documents and $-1$ for those with male-associated ones. Then the inner product between the gradients of the gender loss gap $G_t$ and the weighted training loss $\mathcal{L}_t$ satisfies:*

$$\left\langle \nabla_\theta G_t,\, \nabla_\theta \mathcal{L}_t \right\rangle = \lambda_t \operatorname*{Cov}_{x\sim\mathcal{B}_t} \left( s(x),\, \hat{\ell}_t(x)\, \|\nabla_\theta \ell_{\theta_t}(x)\|^2 \right), \tag{7}$$

*where $\hat{\ell}_t(x)$ is the normalized loss of $x$ in batch $\mathcal{B}_t$.*

Proof. From definitions of $G_t$ and $\mathcal{L}_t$, and the chain rule:

$$\nabla_\theta G_t = \mathbb{E}_{x\sim\mathcal{T}_{\text{f}}}[\nabla_\theta \ell_\theta(x)] - \mathbb{E}_{x\sim\mathcal{T}_{\text{m}}}[\nabla_\theta \ell_\theta(x)] = \mathbb{E}_{x\sim\mathcal{T}}[s(x)\nabla_\theta \ell_\theta(x)].$$

Similarly,

$$\nabla_\theta \mathcal{L}_t = \mathbb{E}_{x\sim\mathcal{B}_t}[\nabla_\theta (w_t(x)\ell_\theta(x))].$$

Since $w_t(x) = 1 - \lambda_t \hat{\ell}_t(x)$ and $\hat{\ell}_t(x)$ is differentiable with respect to $\ell_\theta(x)$, we apply the product rule and simplify to isolate the covariance term in Eq. (7). □

Theorem 3.2 (Monotonic Gender Gap Reduction). *Let $\theta_t$ be the model parameters at training step $t$, and assume $\theta_{t+1}$ is obtained via one step of stochastic gradient descent:*

$$\theta_{t+1} = \theta_t - \eta_t \nabla_\theta \mathcal{L}_t.$$

*Then the absolute gender loss gap satisfies:*

$$|G_{t+1}| \le |G_t|. \tag{8}$$

Proof. A first-order Taylor expansion of $G(\theta)$ around $\theta_t$ gives

$$G_{t+1} \approx G_t - \eta_t \left\langle \nabla_\theta G_t, \nabla_\theta \mathcal{L}_t \right\rangle.$$

Substituting Lemma 3.1,

$$G_{t+1} \approx G_t - \eta_t \lambda_t \cdot \operatorname*{Cov}_{x\sim\mathcal{B}_t} \left( s(x),\, \hat{\ell}_t(x)\, \|\nabla_\theta \ell_{\theta_t}(x)\|^2 \right).$$

Consider the pairwise hinge loss in Eq. (1). If $\ell_\theta(x) = 0$ then $\nabla_\theta \ell_\theta(x) = 0$, and larger margin violations yield larger $\ell_\theta(x)$ together with larger $\|\nabla_\theta \ell_\theta(x)\|$. Hence $\hat{\ell}_t(x)\, \|\nabla_\theta \ell_{\theta_t}(x)\|^2$ is a non-decreasing function of $\ell_\theta(x)$, and since $\hat{\ell}_t$ is a batchwise monotone transform of $\ell_\theta$, the covariance

$$\operatorname{Cov}\left( s(x),\, \hat{\ell}_t(x)\, \|\nabla_\theta \ell_{\theta_t}(x)\|^2 \right)$$

has the same sign as $\operatorname{Cov}(s(x), \hat{\ell}_t(x))$. The latter has the same sign as the groupwise difference in mean losses in the batch, which agrees with the sign of $G_t$. Therefore $\langle \nabla_\theta G_t, \nabla_\theta \mathcal{L}_t \rangle$ has the same sign as $G_t$. If $G_t > 0$ the update decreases $G_t$, and if $G_t < 0$ the update increases $G_t$. In both cases the absolute gap contracts, so $|G_{t+1}| \le |G_t|$, with strict inequality whenever the batch exhibits a nonzero disparity. □

Eq. (8) shows that the absolute gender loss gap is *non-increasing* during training, with each update reducing or at worst leaving it unchanged. This follows directly from the weight construction, without extra assumptions. As $\lambda_t$ decays to 0, the curriculum becomes uniform, ensuring the final model preserves relevance while achieving a strictly smaller gap ($|G_T| \le |G_0|$). Since most group-fair IR metrics (e.g., ARaB, NFaiRR, LIWC bias) are monotone transforms of group losses, a non-increasing $|G_t|$ yields the same trend, as confirmed empirically.

## 4 Experiments

**Research Questions.** Our experiments are designed to address 3 Research Questions (RQs): **RQ1.** Can self-paced learning effectively mitigate gender bias in neural ranking models? We specifically examine whether weighting training samples based on their individual loss values helps reduce bias in the ranking outputs while preserving ranking effectiveness. **RQ2.** How does our method compare to existing state-of-the-art bias mitigation approaches? **RQ3.** Is the proposed method robust across different language model architectures? To evaluate the generalizability of our approach, we conduct experiments using two pretrained language models: BERT-L2 [6, 25], and ELECTRA-small-discriminator [11].

**Datasets and Experimental Setup.** We conduct our experiments using the MS MARCO passage ranking dataset [16], which contains ~200,000 queries and 8.8M passages. For training, we randomly sample 3,000,000 query–positive–negative triples and train the model for one epoch using the Adam optimizer with a sigmoid activation function. Our approach follows the architecture, implementation, and hyperparameter settings of the OpenMatch framework [22]. Complete implementation details, and source code are publicly available on GitHub: https://anonymous.4open.science/r/SPL-bias-727D.

**Bias Measurement and Evaluation Datasets.** To assess both ranking performance and bias mitigation, we focus on measuring *gender bias*. We evaluate the models using two bias-sensitive query sets: (a) *Gender-neutral queries*: Designed to test whether the model introduces gender stereotypes in otherwise neutral contexts. We use the dataset introduced by [19], which consists of 1,765 gender-neutral queries selected from the MS MARCO dataset. (b) *Socially sensitive queries*: This set comprises 215 queries that, if biased, have the potential to reinforce or exacerbate societal inequalities.

**Evaluation Metrics.** To assess ranking performance, we use Mean Reciprocal Rank (MRR) [16]. To evaluate bias, we employ three complementary metrics: (1) *Average Rank Bias (ARaB)* [20], which quantifies the prominence of gendered terms in retrieved documents based on Term Count (TC), Term Frequency (TF), and Boolean presence; (2) *NFaiRR* [19], a fairness metric that captures document-level gender balance, with higher values indicating less biased rankings; and (3) *Linguistic Inquiry and Word Count (LIWC)* [18], used to analyze the frequency of gender-related words in retrieved texts, following the methodology of [9].

**Baseline Methods.** To benchmark our approach, we compare it against several established baselines representing diverse bias mitigation strategies: (1) `AdvBert` [19] uses adversarial debiasing in the ranker's intermediate layers; (2) `Bias-aware Loss` [23] integrates a bias penalty in the loss function for targeted bias reduction during training; (3) `CODER` [28] applies a neutrality regularization term in a transformer model; (4) `Light-Weight Sampling Strategy (LWS)` [8] selects biased documents as negative samples, training the model to recognize and mitigate bias. (5) `Gender Disentanglement` [24] Disentangles gender from semantics in ranker representations by jointly training a relevance predictor and gender classifier, ensuring retrieval relies only on meaning.

**Findings.** To address **RQ1**, we examine whether our proposed self-paced learning approach can effectively mitigate gender bias in neural rankers. We conduct experiments comparing the original model (the model without the self-paced strategy) to our self-paced variant across both bias and effectiveness metrics. Figure 1 summarizes the results, showing scores for bias-related metrics on the left side of the dotted line, where lower values indicate reduced bias, and for performance and fairness metrics on the right side, where higher values are preferred. Our results demonstrate a consistent reduction in bias across both evaluation datasets, as well as across ranking cutoffs (i.e., top-10 and top-20). Notably, our approach achieves up to a 95% reduction in bias on the 215 queries and a 42%

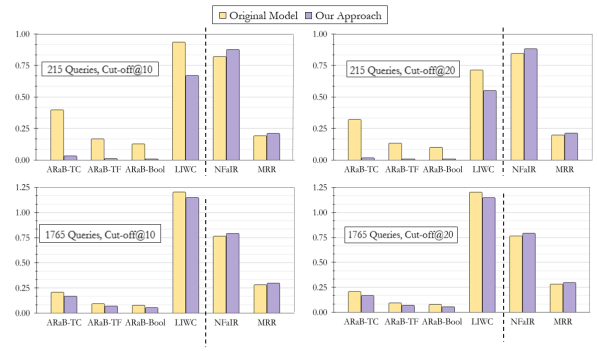


**Figure 1: Effectiveness, and bias metrics for BERT. Metrics to the left of the dotted line lower is better and to the right higher is better.**

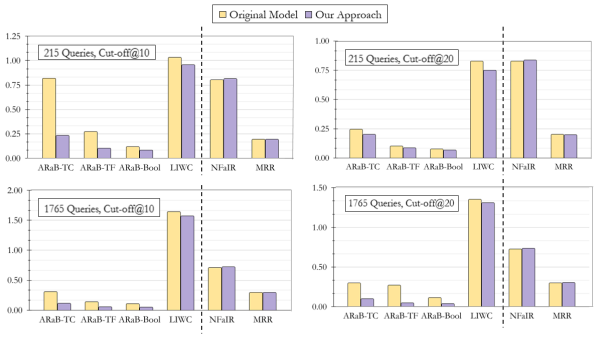


**Figure 2: Effectiveness, and bias metrics for ELECTRA (Metrics interpretation similar to Figure 1).**

reduction on the 1,765 queries, as measured by ARaB and LIWC-based metrics. At the same time, the model preserves, and in some cases slightly improves, ranking effectiveness as measured by MRR. These findings support the conclusion that our self-paced learning strategy successfully prioritizes less biased examples during early training, thereby steering the model toward fairer behavior without sacrificing ranking quality. We note that our approach introduces no additional computational overhead compared to the standard training setup.

**RQ2.** To evaluate the effectiveness of our method relative to state-of-the-art bias mitigation techniques, we compare performance across several baselines, including Bias-Aware Loss, Light-Weight Sampling, ADVBERT, CODER, and Disentanglement. Tables 1 and 2 present results for the two bias-sensitive query sets. Our findings show that the proposed self-paced learning approach consistently outperforms Bias-Aware Loss, Light-Weight Sampling, ADVBERT, and Disentanglement in terms of bias reduction, while also achieving higher MRR scores, indicating better ranking effectiveness. Although CODER demonstrates a slightly greater reduction in bias than our method, this comes at a substantial cost to effectiveness. Specifically, CODER's MRR drops to 0.0014 at cutoff-10 and to 0.0001 at cutoff-20 for the 215 socially sensitive queries, rendering its ranking performance practically unusable. These results highlight that our method strikes a more favorable balance between fairness and effectiveness.

**Table 1: Bias & ranking effectiveness on the 215 query set.**

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | cutoff @10 | | | |
| Models | MRR | ARaB-tc↓ | ARaB-tf↓ | ARaB-bool↓ | NFaiR ↑ | liwc ↓ |
| Bias-Aware Loss [23] | 0.1820 | 0.3419 | 0.1492 | 0.1176 | 0.8209 | 0.9202 |
| Light-Weight-Sampling [8] | 0.1823 | 0.2017 | 0.0938 | 0.0782 | 0.9087 | 0.5636 |
| CODER [28] | 0.0014 | 0.0260 | 0.0171 | 0.0205 | 0.9649 | 0.2998 |
| ADVBERT [19] | 0.1753 | 0.1975 | 0.1054 | 0.1113 | 0.8747 | 0.7850 |
| Disentanglement [24] | 0.1859 | 0.6806 | 0.3067 | 0.2630 | 0.8124 | 1.1716 |
| **Our Approach** | **0.2082** | **0.0317** | **0.0079** | **0.0074** | **0.8797** | **0.6693** |
| | | | cutoff @20 | | | |
| Models | MRR | ARaB-tc↓ | ARaB-tf↓ | ARaB-bool↓ | NFaiR ↑ | liwc ↓ |
| Bias-Aware Loss [23] | 0.1873 | 0.2783 | 0.1169 | 0.0899 | 0.8519 | 0.6650 |
| Light-Weight-Sampling [8] | 0.1876 | 0.1618 | 0.0746 | 0.0616 | 0.9168 | 0.4681 |
| CODER [28] | 0.0001 | 0.0227 | 0.0148 | 0.0178 | 0.9650 | 0.2828 |
| ADVBERT [19] | 0.1799 | 0.1144 | 0.0653 | 0.0710 | 0.8795 | 0.6432 |
| Disentanglement [24] | 0.1939 | 0.6019 | 0.2688 | 0.2309 | 0.8324 | 0.9517 |
| **Our Approach** | **0.2126** | **0.0167** | **0.0049** | **0.0052** | **0.8870** | **0.5493** |

**Table 2: Bias & ranking effectiveness on the 1,765 query set.**

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | cutoff @10 | | | |
| Models | MRR | ARaB-tc↓ | ARaB-tf↓ | ARaB-bool↓ | NFaiR ↑ | liwc ↓ |
| Bias-Aware Loss [23] | 0.2591 | 0.2109 | 0.0949 | 0.0755 | 0.7289 | 1.5142 |
| Light-Weight-Sampling [8] | 0.2558 | 0.1540 | 0.0764 | 0.0680 | 0.8204 | 1.1500 |
| CODER [28] | 0.0001 | 0.0646 | 0.0371 | 0.0421 | 0.8404 | 0.7199 |
| ADVBERT [19] | 0.2019 | 0.4222 | 0.2260 | 0.2363 | 0.7132 | 1.6427 |
| Disentanglement [24] | 0.2571 | 0.8824 | 0.4150 | 0.3746 | 0.7317 | 1.6988 |
| **Our Approach** | **0.2881** | **0.1778** | **0.0715** | **0.0539** | **0.7792** | **1.3268** |
| | | | cutoff @20 | | | |
| Models | MRR | ARaB-tc↓ | ARaB-tf↓ | ARaB-bool↓ | NFaiR ↑ | liwc ↓ |
| Bias-Aware Loss [23] | 0.2653 | 0.1644 | 0.0730 | 0.0574 | 0.7578 | 1.2169 |
| Light-Weight-Sampling [8] | 0.2622 | 0.1192 | 0.0587 | 0.0516 | 0.8313 | 0.9614 |
| CODER [28] | 0.0014 | 0.0674 | 0.0388 | 0.0440 | 0.8407 | 0.6467 |
| ADVBERT [19] | 0.2106 | 0.2731 | 0.1475 | 0.1554 | 0.7424 | 1.2933 |
| Disentanglement [24] | 0.2641 | 0.8190 | 0.3823 | 0.3433 | 0.7389 | 1.5031 |
| **Our Approach** | **0.2941** | **0.1662** | **0.0692** | **0.0541** | **0.7886** | **1.1466** |

**RQ3.** To evaluate the generalizability of our approach across different language models, we replicate our experiments using the ELECTRA architecture. Figure 2 presents both bias and effectiveness metrics for the baseline model and our self-paced variant. The results show that our self-paced learning strategy consistently reduces bias compared to the standard ELECTRA model without self-paced training. In particular, we observe a notable increase in the NFaiRR fairness score, indicating more equitable exposure of documents. Importantly, the Mean Reciprocal Rank (MRR) remains comparable to that of the original model, demonstrating that the improvements in fairness and bias mitigation do not come at the cost of ranking effectiveness.

## 5 Concluding Remarks

We propose a model-agnostic debiasing technique for neural rankers that does not require protected-attribute labels or architectural changes. By using each sample's loss as an implicit bias indicator and applying a self-paced learning curriculum, prioritizing low-loss (less biased) examples, we achieve a steady reduction in the gender loss gap. Our theoretical analysis confirms this monotonic bias decrease, and experiments on MS MARCO demonstrate significant bias reduction without sacrificing, and in some cases improving, ranking effectiveness compared to strong baselines.

# References


[1] Qingyao Ai, Keping Bi, Jiafeng Guo, and W Bruce Croft. 2018. Learning a deep listwise context model for ranking refinement. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 135–144.

[2] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2021. Context-aware target apps selection and recommendation for enhancing personal mobile assistants. *ACM Transactions on Information Systems (TOIS)* 39, 3 (2021), 1–30.

[3] Negar Arabzadeh, Amin Bigdeli, and Charles LA Clarke. 2024. Adapting standard retrieval benchmarks to evaluate generated answers. In *European Conference on Information Retrieval*. Springer, 399–414.

[4] Negar Arabzadeh, Radin Hamidi Rad, Maryam Khodabakhsh, and Ebrahim Bagheri. 2023. Noisy perturbations for estimating query difficulty in dense retrievers. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 3722–3727.

[5] Reza Barzegar, Marco Nikola Kurepa, and Hossein Fani. 2025. Adaptive Loss-based Curricula for Neural Team Recommendation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*. 914–923.

[6] Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics. arXiv:2110.01518 [cs.CL]

[7] Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. 2021. On the orthogonality of bias and utility in ad hoc retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1748–1752.

[8] Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. 2022. A light-weight strategy for restraining gender biases in neural rankers. In *European Conference on Information Retrieval*. Springer, 47–55.

[9] Amin Bigdeli, Negar Arabzadeh, Morteza Zihayat, and Ebrahim Bagheri. 2021. Exploring gender biases in information retrieval relevance judgement datasets. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*. Springer, 216–224.

[10] Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with recursive neural networks and its application to multi-document summarization. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 29.

[11] K Clark. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555* (2020).

[12] Seyed Mohammad Hosseini, Negar Arabzadeh, Morteza Zihayat, and Ebrahim Bagheri. 2024. Enhanced Retrieval Effectiveness through Selective Query Generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 3792–3796.

[13] Maryam Khodabakhsh and Ebrahim Bagheri. 2023. Learning to rank and predict: multi-task learning for ad hoc retrieval and query performance prediction. *Information Sciences* 639 (2023), 119015.

[14] Yingji Li, Mengnan Du, Rui Song, Xin Wang, Mingchen Sun, and Ying Wang. 2024. Mitigating social biases of pre-trained language models via contrastive self-debiasing with double data augmentation. *Artificial Intelligence* 332 (2024), 104143.

[15] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. Pretrained Transformers for Text Ranking: BERT and Beyond. *Synthesis Lectures on Human Language Technologies* 14, 4 (2021), 1–325. doi:10.2200/S01123ED1V01Y202108HLT052

[16] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *CoRR* abs/1611.09268 (2016). arXiv:1611.09268 http://arxiv.org/abs/1611.09268

[17] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Mike Lewis. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. *Findings of the Association for Computational Linguistics: EMNLP 2020* (2020), 1017–1029. doi:10.18653/v1/2020.findings-emnlp.92

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

[19] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal biases in retrieved contents: Measurement framework and adversarial mitigation of bert rankers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 306–316.

[20] Navid Rekabsaz and Markus Schedl. 2020. Do neural ranking models intensify gender bias?. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2065–2068.

[21] Navid Rekabsaz and Markus Schedl. 2020. Do Neural Ranking Models Intensify Gender Bias? *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020), 1145–1154. doi:10.1145/3397271.3401142

[22] Kuniaki Saito, Donghyun Kim, and Kate Saenko. 2021. OpenMatch: Open-set Consistency Regularization for Semi-supervised Learning with Outliers. *CoRR* abs/2105.14148 (2021). arXiv:2105.14148 https://arxiv.org/abs/2105.14148

[23] Shirin Seyedsalehi, Amin Bigdeli, Negar Arabzadeh, Bhaskar Mitra, Morteza Zihayat, and Ebrahim Bagheri. 2022. Bias-aware Fair Neural Ranking for Addressing Stereotypical Gender Biases.. In *EDBT*. 2–435.

[24] Shirin Seyedsalehi, Sara Salamat, Negar Arabzadeh, Sajad Ebrahimi, Morteza Zihayat, and Ebrahim Bagheri. 2025. Gender disentangled representation learning in neural rankers. *Machine Learning* 114, 5 (2025), 1–33.

[25] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-Read Students Learn Better: The Impact of Student Initialization on Knowledge Distillation. *CoRR* abs/1908.08962 (2019). arXiv:1908.08962 http://arxiv.org/abs/1908.08962

[26] Duc-Thuan Vo, Fattane Zarrinkalam, Ba Pham, Negar Arabzadeh, Sara Salamat, and Ebrahim Bagheri. 2023. Neural Ad-Hoc Retrieval Meets Open Information Extraction. In *European Conference on Information Retrieval*. Springer, 655–663.

[27] Pengtao Xie and Eric Xing. 2018. A neural architecture for automated ICD coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1066–1076.

[28] George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. 2021. CODER: An efficient framework for improving retrieval through Contextual Document Embedding Reranking. *arXiv preprint arXiv:2112.08766* (2021).

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

[30] Yucheng Zhou, Tao Shen, Xiubo Geng, Chongyang Tao, Can Xu, Guodong Long, Binxing Jiao, and Daxin Jiang. 2022. Towards robust ranker for text retrieval. *arXiv preprint arXiv:2206.08063* (2022).