Bias-Aware Curriculum Sampling For Fair Ranking

Abstract

Neural ranking models are widely used in information retrieval (IR) to retrieve and rank relevant documents. However, these models may inherit and amplify biases present in the training data, posing challenges for fairness and relevance in ranking outputs. In this paper, we propose a novel curriculum-based training approach that manages bias exposure throughout the training process. We design a bias-aware curriculum that stages the exposure of the model to biased samples during the training stages, allowing the model to establish a fair relevance baseline. We conduct extensive experiments across different LLMs and datasets to evaluate the effectiveness of our approach. Our results demonstrate that our proposed curriculum-based strategy outperforms state-of-the-art bias reduction methods in terms of both fairness and relevance, without sacrificing retrieval effectiveness.

1 Introduction

Neural rankers are central to recent IR systems, enabling efficient document retrieval across applications [2, 3, 11, 32, 36]. However, their effectiveness comes with a vulnerability: sensitivity to biases in training data [24]. Gold standard datasets often contain inherent biases, as human-generated content reflects demographic or ideological skews [9]. When trained on such data, neural rankers inherit and may amplify these biases, prioritizing frequent patterns as relevance indicators [24]. This risks embedding societal biases in ranking outputs, undermining fairness and trustworthiness.

Existing bias mitigation approaches for neural rankers often focus on data-level debiasing [6, 8] and modifying learning objectives [23, 26, 34]. Data-level debiasing adjusts biased samples before training, reducing explicit biases but risking loss of valuable information and failing to adapt to bias exposure during training. Model-level methods introduce regularization terms or bias-specific penalties [23, 26, 34], embedding bias mitigation in training but altering relevance learning. In either case, optimizing for both fairness and relevance may lead to degradation of ranking performance potentially entangling bias within the relevance criteria [15, 22].

While existing approaches focus on data-level debiasing or modifying training objectives, we hypothesize that an alternative strategy lies in controlling the exposure sequence of biased samples during training. Research has shown that the order in which samples are presented to neural models significantly impacts performance, a process known as curriculum learning [30]. Depending on the context, models can benefit from a structured progression of training data, moving from harder to easier samples or vice versa. In the case of bias mitigation, our key hypothesis is that controlling the sequence in which biased samples are introduced could influence the extent to which biases are embedded in the final model.

While this curriculum-based approach offers a possible alternative to traditional debiasing methods [7–9, 34], it presents key challenges. *First*, designing an effective curriculum is non-trivial, as it must prioritize relevance learning while gradually introducing biased samples. Poor sequencing risks premature bias internalization or an inadequate relevance foundation. *Second*, a dynamic sampling strategy is essential to adjust bias exposure in alignment with the model's progress, ensuring training stability and controlled bias integration. To address these challenges in this paper, inspired by the works on Curriculum Learning in IR [17, 20, 33], we provide the following contributions: *(i)* We formalize the use of *curriculum learning* for structured bias exposure in neural rankers, balancing bias mitigation and relevance learning. *(ii)* We introduce a bias-aware curriculum that sequences training samples based on bias scores, progressively incorporating staged exposure to biased samples. *(iii)* We propose a dynamic sampling strategy that adjusts sampling probabilities based on bias scores, ensuring gradual exposure without compromising convergence. *(iv)* We conduct extensive experiments demonstrating the effectiveness of our approach compared to existing bias mitigation techniques.

2 Problem Formulation

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

Training datasets often contain biases that subtly influence model learning [4, 10, 18, 35]. A document d_j may encode biases, such as inclination towards a certain gender, quantifiable via a bias scoring function, $\Psi(d_j)$ [1, 21, 23, 24]. These biases risk entangling with relevance signals, leading the model to misinterpret biased patterns as relevance indicators, compromising ranking fairness. Mitigating this requires (1) maintaining high ranking effectiveness, $\Lambda(Q)$, while minimizing bias in outputs, $\Pi(Q)$; and (2) sequencing training samples to positively shape learning dynamics.

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

$$\arg\max_{\theta} \Lambda(Q) \quad \text{subject to} \quad \Pi(Q) \to 0. \tag{1}$$

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

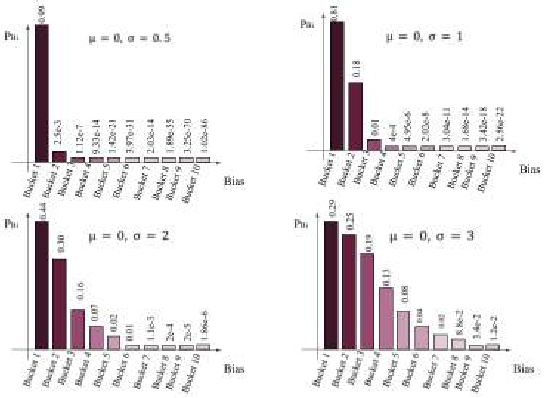

Figure 1: Sampling probs for 10 buckets, with $\sigma = \{0.5, 1, 2, 3\}$.

3 Methodology

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

Bias-Aware Curriculum Design. A key challenge in bias-aware learning is preventing models from misinterpreting biases for relevance signals. Building on curriculum learning principles [5, 16, 27, 31], we hypothesize that introducing less (or high, depending on curriculum direction) biased samples earlier in training may shape the model's understanding of bias and possibly mitigates bias. To achieve this, we define a bias scoring function, $\Psi(d^+_{ij})$, quantifying bias in each document d^+_{ij}. This score informs our sampling strategy, adjusting document selection probabilities. Prioritizing low (or high) bias samples early impacts the risk of bias influencing initial learning, potentially creating a fairer relevance baseline. As training progresses, higher (or lower) bias samples are gradually introduced, refining relevance without embedding bias.

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

To quantify the bias within each relevant document in a training sample $S = (q_i, d^+_{ij}, d^-_{ik})$, we compute a bias score for each relevant document d^+_{ij}, denoted as $\Psi(d^+_{ij})$. This score reflects the degree of bias present in the document and serves as the primary metric for ranking training samples based on their degree of bias. Specifically, the bias score Ψ is a function mapping each relevant document d^+_{ij} to a real-valued score, defined as:

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

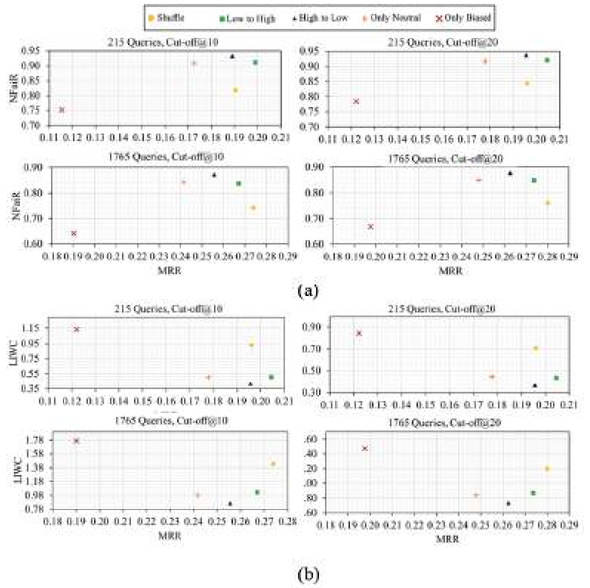

Figure 2: The Bias-performance trade-off.

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

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

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

$$P_{B_i} \propto \frac{1}{x_i}, \quad \text{where } x_i = \frac{1}{|B_i|} \sum_{d^+_{ij} \in B_i} \Psi(d^+_{ij}). \tag{2}$$

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

As an illustrative example, Figure 1 shows the sampling probability distribution across ten buckets, demonstrating how the Gaussian

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

cutoff @10

Models	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaiR ↑	liwc ↓
Bias-Aware Loss [26]	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 [34]	0.0014	0.0260	0.0171	0.0205	0.9649	0.2998
ADVBERT [23]	0.1753	0.1975	0.1054	0.1113	0.8747	0.7850
Our Approach	**0.1989**	**0.0773**	**0.0376**	**0.0322**	**0.9126**	**0.5057**

cutoff @20

Models	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaiR ↑	liwc ↓
Bias-Aware Loss [26]	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 [34]	0.0001	0.0227	0.0148	0.0178	0.9650	0.2828
ADVBERT [23]	0.1799	0.1144	0.0653	0.0710	0.8795	0.6432
Our Approach	**0.2046**	**0.0632**	**0.0308**	**0.0265**	**0.9212**	**0.4355**

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

cutoff @10

Models	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaiR ↑	liwc ↓
Bias-Aware Loss [26]	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 [34]	0.0001	0.0646	0.0371	0.0421	0.8404	0.7199
ADVBERT [23]	0.2019	0.4222	0.2260	0.2363	0.7132	1.6427
Our Approach	**0.2671**	**0.0095**	**0.0062**	**0.0090**	**0.8382**	**1.0275**

cutoff @20

Models	MRR	ARaB-tc↓	ARaB-tf↓	ARaB-bool↓	NFaiR ↑	liwc ↓
Bias-Aware Loss [26]	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 [34]	0.0014	0.0674	0.0388	0.0440	0.8407	0.6467
ADVBERT [23]	0.2106	0.2731	0.1475	0.1554	0.7424	1.2933
Our Approach	**0.2737**	**0.0173**	**0.0040**	**0.0047**	**0.8478**	**0.8684**

model regulates data exposure. In this case, lower-bias buckets receive higher sampling probabilities, while higher-bias buckets have progressively lower ones. We hypothesize that this structured approach promotes fairness and robustness during training. We note that in this approach, all training instances will eventually be sampled. Lower initial sampling probabilities do not exclude samples but delay their introduction.

Progressive Learning Objective. The selected samples $S = (q_i, d_{ij}^{+}, d_{ik}^{-})$ are then fed into a cross-encoder neural ranker. The model calculates relevance scores for both the relevant document d_{ij}^{+} and the irrelevant document d_{ik}^{-} in relation to the query q_i: $s(q_i, d_{ij}) = \Phi(q_i \oplus d_{ij})$. The model is trained with a Max Margin Loss [13] calculated as:

$$L = \frac{1}{n} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, m - \Phi(q, {d_i}^+) + \Phi(q, {d_j}^-)) \tag{3}$$

4 Experiments

Research Questions. Our experiments are structured around four Research Questions (RQs): **RQ1.** Does curriculum sampling effectively mitigate biases in neural rankers? Specifically, we investigate whether structuring the training sequence based on bias levels can reduce bias in ranking outputs while maintaining model effectiveness. **RQ2.** How does the model perform relative to the state-of-the-art bias reduction baseline methods? **RQ3.** Does the choice of probability distribution hyperparameters for bias-aware sampling influence the model's ranking performance and bias mitigation effectiveness? **RQ4.** Is the model's performance consistent across different language models? We run experiments on 3 language models, BERT-mini [14, 28], MiniLM [29], and ELECTRA [12] to assess the generalizability of our approach across LLMs.

Datasets and Setup. We train the neural rankers on the MS-MARCO passage ranking dataset [19], with 200,000 queries and 8.8 million passages. A random sample of 3,000,000 triples is used for training over one epoch, using the Adam optimizer with a sigmoid activation. We follow OpenMatch [25] architecture, implementation, and hyperparameters. Full implementation details and source code are available on GitHub: https://shorturl.at/e3ggk.

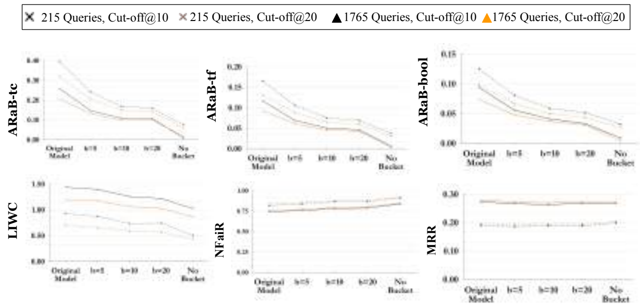

Figure 3: Impact of bucket size on model performance.

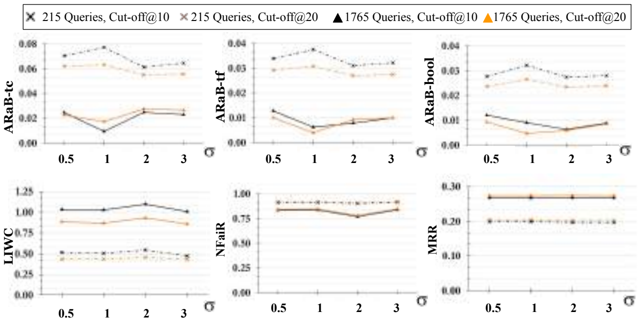

Figure 4: Impact of σ on model performance.

Bias Measure and Bias Test Datasets. We consider ARaB-tf [24] as the function Ψ in Equation 2 for measuring bias of the documents. To evaluate performance and bias reduction, we focus on *gender bias* using two bias query datasets: (a) *Gender-neutral queries*: These queries assess whether the model introduces gender stereotypes in neutral contexts. We use the query set from [23], which includes 1,765 gender-neutral queries, selected from MS MARCO queries. (b) *Socially sensitive queries*: This set includes 215 queries that may contribute to societal inequality if bias is present.

Evaluation Metrics. For ranking, we use Mean Reciprocal Rank (MRR) [19]. For bias, we use three metrics: *Average Rank Bias (ARaB)* [24], which quantifies biased word occurrences in documents using Term Count (TC), Term Frequency (TF), and Boolean metrics; *NFaiRR* [23], measuring document-level fairness, with higher values indicating fairer rankings; and *Linguistic Inquiry and Word Count (LIWC)* [21], which assesses gender mentions in text as suggested by [9].

Baseline Methods. To benchmark our approach, we compare it against five established baselines representing diverse bias mitigation strategies: (1) `AdvBert` [23] uses adversarial debiasing in the ranker's intermediate layers; (2) `Bias-aware Loss` [26] integrates a bias penalty in the loss function for targeted bias reduction during training; (3) `CODER` [34]^0 applies a neutrality regularization term in a transformer model; (4) `Light-Weight Sampling Strategy (LWS)` selects biased documents as negative samples, training the model to recognize and mitigate bias.

Findings. In **RQ1**, we assess whether our proposed curriculum sampling approach reduces bias. We conduct experiments under

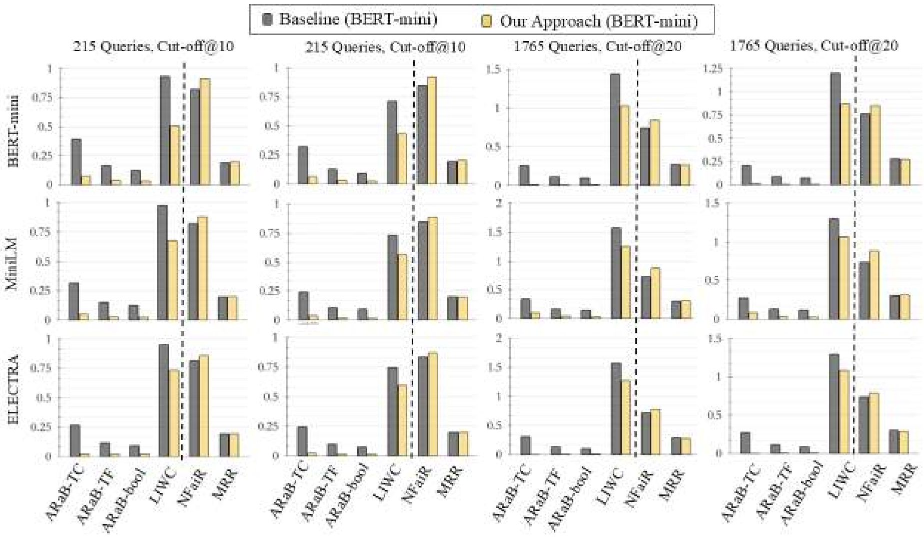

Figure 5: Generalizability of our proposed approach on different LLMs.

five conditions: (1) shuffled training (no curriculum), (2) High to Low bias curriculum sampling, (3) Low to High bias curriculum sampling, (4) training on neutral samples (bias-free), and (5) training on biased samples. Figure 2(a) shows the trade-off between fairness (NFaiR [23]) and effectiveness (MRR). Training on biased samples yields the worst fairness and effectiveness, performing worse than the shuffled baseline. Neutral-sample training improves fairness but reduces performance by up to 10%. The High to Low strategy improves fairness but still lowers effectiveness (up to 7%). In contrast, our proposed *Low to High curriculum strategy* achieves the best fairness-effectiveness trade-off, as further confirmed in Figure 2(b), where lower LIWC values indicate reduced bias. Empirical results show that *Low to High curriculum sampling* best balances bias reduction and retrieval effectiveness. By prioritizing low-bias samples early, the model establishes a robust relevance foundation before gradually incorporating higher-bias samples. This controlled exposure prevents early bias internalization, allowing the model to learn relevance signals more effectively. Without loss of generality and to save space, we adopt this strategy for reporting subsequent RQs.
In **RQ2**, we compare the performance of our proposed approach with the state-of-the-art baseline methods. Tables 1 and 2 show this comparison. We observe that our proposed method outperforms the bias-aware loss, Light-Weight-Sampling, and ADVBERT methods in terms of bias reduction, while having higher MRR. The other baseline method, CODER, although is able to reduce the bias more than our proposed approach, but it significantly reduces retrieval effectiveness to 0.0014 for cut-off 10, and 0.0001 for cut-off 20 on the 215 query set, effectively making unhelpful retrieval.
In **RQ3,** we examine the impact of probability distribution parameters on model performance, focusing on bucket size b and standard deviation σ in sampling probabilities. Figure 3 shows the effect of varying bucket sizes using a Normal Distribution ($\mu = 0, \sigma = 1$) and BERT-mini as the baseline. The analysis covers two query sets (215 and 1,765 queries) at cut-offs 10 and 20. The results reveal: *(1) Bias Reduction:* Increasing b from 5 to 20 reduces bias across all metrics (ARaB-TC, ARaB-TF, ARaB-Bool, LIWC, and NFaiR). *(2) Effect of Larger Buckets:* Larger bucket sizes further enhance bias reduction, with $b = 20$ outperforming $b = 5$. *(3) No-Bucket Strategy:* Treating each sample as an individual bucket maximizes performance and minimizes bias, aligning with the trend that larger bucket sizes reduce bias more effectively. Figure 4 examines the effect of varying σ on model stability and bias metrics. Our results show *(i) MRR Stability:* MRR remains stable across sigma values (0.27 on 1,765 queries, 0.22 on 215 queries), showing minimal impact on retrieval effectiveness. *(ii) Consistency in Bias Metrics:* Bias measures (ARaB-TC, ARaB-TF, ARaB-BOOL, LIWC, NFaiR) exhibit less than 5% variation across sigma values.

In **RQ4**[1]**,** we assess whether our approach generalizes across different LLMs while reducing bias and maintaining retrieval effectiveness. We repeat experiments with MiniLM and ELECTRA alongside BERT-mini. Since Figure 3 indicates the best results occur in the no-bucket scenario, we train models with no bucket, and $\sigma = 1$. Figure 5 presents bias reduction and ranking performance across the LLMs. The first, second, and third rows show results for BERT-mini, MiniLM, and ELECTRA, respectively. Metrics to the left of the dotted line measure bias (lower is better), while those on the right assess fairness and effectiveness (higher is better). Our approach significantly reduces bias compared to the baseline (no curriculum sampling) while increasing the NFaiR fairness metric. Additionally, MRR remains comparable to the original model, confirming that bias reduction does not compromise ranking effectiveness in our proposed curriculum sampling approach.

5 Concluding Remarks

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

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

References

[1] Amin Abolghasemi, Leif Azzopardi, Arian Askari, Maarten de Rijke, and Suzan Verberne. 2024. Measuring Bias in a Ranked List Using Term-Based Representations. In *European Conference on Information Retrieval*. Springer, 3–19.

[2] 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.

[3] 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.

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

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

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

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

[11] 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.

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

[13] Corinna Cortes. 1995. Support-Vector Networks. *Machine Learning* (1995).

[14] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[15] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. 2020. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International conference on machine learning*. PMLR, 2803–2813.

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

[17] Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. 2020. Training curricula for open domain answer re-ranking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 529–538.

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

[19] 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

[20] Gustavo Penha and Claudia Hauff. 2020. Curriculum learning strategies for IR: An empirical study on conversation response ranking. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42*. Springer, 699–713.

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

[22] Gideon Popoola and John Sheppard. 2024. Investigating and Mitigating the Performance–Fairness Tradeoff via Protected-Category Sampling. *Electronics* 13, 15 (2024), 3024.

[23] 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.

[24] 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.

[25] 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

[26] 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.

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

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

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

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

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

[32] 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.

[33] Hansi Zeng, Hamed Zamani, and Vishwa Vinay. 2022. Curriculum learning for dense retrieval distillation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1979–1983.

[34] 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).

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

[36] 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).