Towards Fair Retrieval: Controlling Bias through a Backpack-Inspired Architecture

Amirabbas Afzali[*]
Aktus AI
amirabbas.afzali@aktus.ai

Amirreza Velae[*]
Sharif University of Technology
amirrezavelae@gmail.com

Iman Ahmadi
Sharif University of Technology
iman1234ahmadi@gmail.com

Mohammad Aliannejadi
University of Amsterdam
m.aliannejadi@uva.nl

Abstract

The presence of social biases in large language models (LLMs) has become a significant concern in AI research. These biases, often embedded in training data, can perpetuate harmful stereotypes and distort decision-making processes. When LLMs are integrated into ranking systems, they can propagate these biases, leading to unfair outcomes in critical applications such as search engines and recommendation systems. Backpack Language Models (Hewitt et al., 2023) are effective at learning diverse word senses and debiasing generative language models. Unlike traditional transformer-based models that treat text sequences as monolithic structures, Backpack-LM generates outputs as weighted combinations of non-contextual, learned word aspects, also known as senses. Leveraging this architecture, we propose a framework for debiasing ranking tasks. Our experimental results. show that this framework effectively mitigates gender bias in text retrieval and ranking with minimal degradation in performance, offering a more balanced approach to information retrieval.

1 Introduction

Ranking and retrieval are core components of modern information systems, underpinning search, recommendation, and decision-support pipelines. Recent advances in large language models (LLMs) such as LLaMA2 (Touvron et al., 2023) and T5 (Raffel et al., 2023) have substantially improved effectiveness when used as rankers or re-rankers (Ma et al., 2023; Zhuang et al., 2022). However, integrating LLMs into retrieval workflows raises well-documented concerns about fairness and bias. Because these models are trained on large-scale web corpora that encode societal stereotypes, including those related to gender and race (Nadeem et al., 2021; Kotek et al., 2023), they can propagate or even amplify harmful associations. For instance, an LLM-informed ranker may implicitly associate specific occupations with a particular gender, yielding systematically skewed results in hiring or recommendation scenarios(Chen et al., 2025).

Decades of research in psychology and sociology show that gender stereotypes shape expectations, judgments, and information processing (Burgess and Borgida, 1999; Ellemers, 2018; Heilman, 2012), contributing to biased treatment and outcomes (Swim et al., 1989) and often arising from misperceptions (Huddy and Terkildsen, 1993). Within Information Retrieval (IR), these concerns are salient because neural embeddings and representations—widely used for matching and ranking—have been shown to encode stereotypical regularities (Rekabsaz and Schedl, 2020; Sun et al., 2019; Bolukbasi et al., 2016; Fabris et al., 2020). Such biases pose practical risks by reinforcing gendered patterns in retrieved content. Prior work has both analyzed how biased embeddings affect ranked outputs (Rekabsaz and Schedl, 2020; Fabris et al., 2020) and proposed metrics to quantify gendered responses in ranked lists (Fabris et al., 2020).

Mitigation strategies in IR have largely been out-of-process: post hoc interventions that operate external to the model, including re-ranking, fairness-aware scoring, or embedding adjustments (Zehlike et al., 2017; Asudeh et al., 2019; Zerveas et al., 2022). While effective in some settings, these approaches often require additional optimization or fine-tuning, can be computationally costly for large models, and may offer limited interpretability or controllability during deployment. This motivates an in-process perspective, in which fairness control is implemented within the model's inference procedure itself—exposing explicit, interpretable levers to regulate bias without retrain-

[*]Equal contribution.

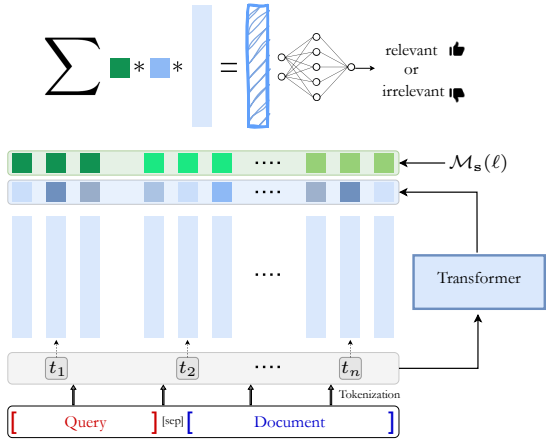

Figure 1: Overview of our proposed bias-controllable ranking framework. Each token is decomposed into multiple non-contextual sense vectors. We compute the sensitivity of each sense to a targeted aspect (e.g., gender), and apply a learned reweighting policy $\mathcal{M}_s(\ell)$ to selectively amplify or suppress individual senses. The reweighted representations are aggregated and passed through a two-layer MLP to produce a final relevance score.

ing the (potentially very large) backbone.

Backpack Language Models (Backpack-LMs) (Hewitt et al., 2023) expose interpretable, non-contextual sense vectors for each token that are linearly reweighted by context, enabling in-process control at inference. We leverage this structure for fair ranking by (i) estimating per-sense sensitivity to a targeted attribute (e.g., gender) using paired lexicons, and (ii) applying a mapping policy $\mathcal{M}_s$ that reweights senses during scoring. The intervention is training-free with respect to fairness (no additional fine-tuning) and integrates directly into our listwise ranking pipeline. Empirically, it reduces gender bias (RaB/ARaB) with minimal impact on effectiveness (MRR/NDCG), providing a practical and interpretable route to fairer retrieval.

2 Preliminaries

Ranking Systems. Given a query q_i and a finite candidate set $\mathcal{D}_i = \{d_{i1}, \ldots, d_{im}\}$, a ranking system outputs a permutation of the candidates. The ground truth for query i is an ordinal label vector $y_i = (y_{i1}, \ldots, y_{im})$ with $y_{ij} \in \{1, \ldots, m\}$, where lower values indicate better rank (ties permitted). We parameterize a scoring function f_θ that maps each querydocument pair (q_i, d_{ij}) to a real-valued score,

$$s_{ij} \;=\; f_\theta(q_i, d_{ij}) \in \mathbb{R}.$$

A full ranking is induced by sorting the scores in descending order, $\hat{\pi}_i = \mathrm{argsort}_{\downarrow}(s_{i1}, \ldots, s_{im})$. Training minimizes a loss that encourages order consistency between y_i and the predicted scores, i.e., items with better (lower) ground-truth ranks should receive higher scores.

Backpack Architecture. Let $\mathcal{V}$ denote a finite vocabulary and $x_{1:n} = (x_1, \ldots, x_n)$, with $x_i \in \mathcal{V}$, an input token sequence. The Backpack architecture maps $x_{1:n}$ to a sequence of output vectors $o_{1:n} = (o_1, \ldots, o_n)$, $o_i \in \mathbb{R}^d$. For each token $x \in \mathcal{V}$, the model stores k *sense vectors* $C(x)_1, \ldots, C(x)_k$, where $C : \mathcal{V} \to \mathbb{R}^{d \times k}$ defines a multi-vector, non-contextual embedding space that captures distinct senses or facets of a word.

Contextualization is achieved by aggregating sense vectors across the sequence with learned weights. Formally,

$$o_i \;=\; \sum_{j=1}^{n} \sum_{\ell=1}^{k} \alpha_{\ell ij}\, C(x_j)_\ell, \tag{1}$$

where $\alpha \in \mathbb{R}^{k \times n \times n}$ are context-dependent weights produced by a function $A : \mathcal{V}^n \to \mathbb{R}^{k \times n \times n}$, i.e., $\alpha = A(x_{1:n})$. Here, $C(x_j)_\ell \in \mathbb{R}^d$ denotes the ℓ-th sense vector associated with token x_j.

The model defines a distribution over an output space $\mathcal{Y}$ via a log-linear transformation of the sequence representation $o_{1:n} \in \mathbb{R}^{d \times n}$:

$$p(y \mid o_{1:n}) \;=\; \mathrm{Softmax}\big(E(o_{1:n})\big), \tag{2}$$

where $E : \mathbb{R}^{d \times n} \to \mathbb{R}^{|\mathcal{Y}|}$ is a linear map and $y \in \mathcal{Y}$. This construction preserves a log-linear dependency on the underlying sense vectors $C(x_j)_\ell$, enabling fine-grained attribution of predictive influence to individual senses across contexts.

3 Methodology

In this section, we detail our proposed framework. To leverage the pretrained knowledge of the Backpack-LM for ranking tasks, we replace the final linear layer of the language model with a scalar output layer with sigmoid activation. We then fine-tune the pretrained model on the ranking task using the *listwise softmax cross-entropy* loss (Bruch et al., 2019), which optimizes ranking by considering the entire list rather than treating documents individually or in pairs. For a given query q_i, where y_{ij} represents the ground-truth relevance

of document d_{ij} and $\hat{y}_{ij}$ denotes our model's predicted score, the loss is defined as:

$$\mathcal{L}_{\text{Softmax}}(y_i, \hat{y}_i) = -\sum_{j=1}^{m} y_{ij} \log \left(\frac{\exp(\hat{y}_{ij})}{\sum_{j'} \exp(\hat{y}_{ij'})} \right). \quad (3)$$

As mentioned in Section 2, given an input token sequence $x_{1:n}$, the Backpack model projects each token x_j into k vectors $C(x_j)_i$, for $i \in \{1, 2, \ldots, k\}$, independently of other tokens. During the inference phase, we aim to disentangle and control different aspects of language, including potential biases like gender bias. To this end, we propose a two-step heuristic as follows:

(i) Sensitivity Estimation. We compute the sensitivity of each sense vector, denoted as $s = (s_1, \ldots, s_k) \in \mathbb{R}^k$, to identify those most relevant to the targeted aspect. This is done using an auxiliary set of paired words $\mathcal{D}^{\text{aux}} = \{(d_i^-, d_i^+)\}_i$, corresponding to a specific attribute such as gender (e.g., (*"He"*, *"She"*)). Each sensitivity score s_i is computed as:

$$s_\ell = \frac{1}{|\mathcal{D}^{\text{aux}}|} \sum_{(d^-, d^+) \in \mathcal{D}^{\text{aux}}} \langle C(d^-)_\ell, C(d^+)_\ell \rangle, \quad (4)$$

where $\langle \cdot, \cdot \rangle$ denotes the cosine similarity between the i-th normalized sense vectors of the word pair. Intuitively, lower values of s_i indicate higher sensitivity to the targeted aspect, enabling us to suppress or amplify these components through some mapping policy $\mathcal{M}_s$ and thereby control the bias present in the final output score.

(ii) BiasControlled Reweighting. We then define $\mathcal{M}_s : \{1, \cdots, k\} \to \mathbb{R}^+$, a mapping policy that assigns a positive weight to each sense index. When $\mathcal{M}_s(\ell) \geq 1$, the influence of the ℓ-th sense is amplified; otherwise, it is suppressed. Using this policy, the debiased output vector $\tilde{o}_i$ is computed as:

$$\tilde{o}_i = \sum_{j=1}^{n} \sum_{\ell=1}^{k} \mathcal{M}_s(\ell)\, \alpha_{\ell ij}\, C(x_j)_\ell, \quad (5)$$

where $\alpha_{\ell ij}$ denotes the contextual attention weights and $C(x_j)_\ell$ the ℓ-th sense vector of token x_j. Through this framework, i.e., by adjusting the mapping policy $\mathcal{M}_s$, we can directly control the influence of the targeted social aspect (e.g., gender) in the final representation. This representation, $\tilde{o}_i$, is passed through a two-layer MLP to produce a scalar score, which represents the document's predicted relevance.

As a result, our method enables a controllable ranking framework that supports fine-grained fairness interventions during inference. By selectively amplifying or suppressing specific sense vectors, we can steer the model's behavior toward fairer retrieval outcomes, without requiring additional training or structural modifications to the underlying language model.

4 Experiments

Experimental Setup. Our experiments focus on controlling gender bias in document ranking. We begin by evaluating the general ranking performance of our proposed architecture, which integrates a 170M-parameter pretrained Backpack-based backbone[1] with a scalar output layer. We compare its performance against a similarly sized GPT-2 model, demonstrating that fine-tuning a decoder-only ranker can achieve competitive results on ranking tasks. All models are fine-tuned on the MS MARCO dataset (Bajaj et al., 2018) using the listwise softmax cross-entropy loss. Training is performed for 4 epochs with a learning rate of 1×10^{-5}. Next, we evaluate our model on the dataset introduced by (Rekabsaz and Schedl, 2020) to assess gender bias in the ranking outputs. The evaluation is conducted across multiple cut-off values (5, 10, 20, 30, and 40), and the results are presented in Table 2. We use RaB and ARaB to measure gender fairness, and MRR@10, NDCG@5, and NDCG@10 to assess ranking performance[2].

Results. Table 1 presents the ranking performance of different backbone variants. The results show that our approach consistently outperforms the GPT-2 baseline across all metrics. These findings confirm that a decoder-only ranker, when fine-tuned appropriately, can achieve strong ranking performance.

To evaluate the controllability of gender bias in our framework, we introduce a hyperparameter α. We identify the two most gender-sensitive sense vectors (i.e., those with the lowest sensitivity scores) and scale their influence by α. Lower values of α correspond to stronger suppression of gender-related information. Table 2 reports RaB and ARaB metrics at cut-off values $t \in \{10, 20, 30, 40\}$ for our method and several base-

[1] https://huggingface.co/stanfordnlp/backpack-gpt2

[2] Further details about the dataset, baselines, and the measurement metrics are provided in Appendices A, B, and C.

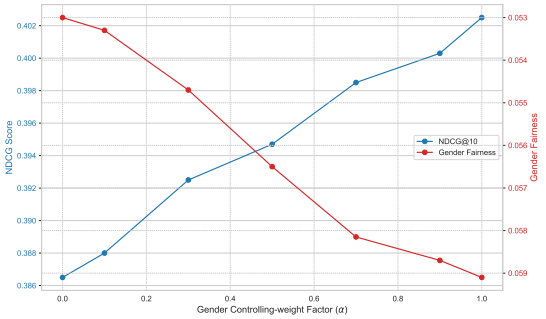

Figure 2: Effect of the controlling weight α on the trade-off between ranking performance (NDCG@10) and gender fairness. Lower values of α improve fairness with minimal loss in performance.

Table 1: MS MARCO ranking performance across different backbone variants. The best scores are highlighted in bold. As expected, decreasing the value of α slightly reduces ranking performance, but the impact is minimal, indicating that our framework maintains robust effectiveness even under fairness constraints.

Backbone	MRR@10	NDCG@5	NDCG@10
GPT-2	0.3156	0.3448	0.3843
BackPack ($\alpha = 0.5$)	0.3252	0.3563	0.3947
BackPack ($\alpha = 0.7$)	0.3298	0.3601	0.3985
BackPack ($\alpha = 1.0$)	**0.3343**	**0.3640**	**0.4025**

lines. Bold values indicate the lowest (i.e., best) bias scores. Figure 2 illustrates the trade-off between gender fairness and ranking performance as the value of α is varied. As expected, decreasing α improves fairness by reducing bias, with only a marginal decrease in ranking quality.

5 Related Work

LLMs have become foundational in applications such as search engines (Xiong et al., 2024) and machine translation (Thirunavukarasu et al., 2023). In ranking and retrieval systems, LLMs enhance contextual understanding, with models like RAG (Lewis et al., 2020) improving response relevance by grounding generation in retrieved content (Salemi and Zamani, 2024). Encoder-decoder architectures, including `MonoT5` and `ListT5` (Ge et al., 2021; Yoon et al., 2024b), further boost ranking performance through joint modeling of queries and documents.

However, LLMs trained on large-scale web data often inherit societal biases (Nadeem et al., 2021; Kotek et al., 2023), leading to unfair associations, such as linking nurse with women and doctor with men (Zhao et al., 2018)—and reinforcing stereotypes (Navigli et al., 2023; Omiye et al., 2023). When deployed in retrieval tasks, these biases risk amplifying disparities, especially in sensitive domains like hiring and healthcare (Bigdeli et al., 2021; Otterbacher, 2018; Venkatasubramanian et al., 2020; Sarr and Appert, 2021).

Table 2: Evaluation of gender bias in retrieval rankings measured by Rank Bias (RaB) and Average Rank Bias (ARaB) at cutoffs 10, 20, 30, and 40. Values closer to zero indicate less gender bias in the top-ranked documents. Our model with $\alpha = 0.5$ consistently achieves the lowest bias across all metrics compared to baselines.

	TF		Boolean	
Model	**RaB**	**ARaB**	**RaB**	**ARaB**
Cut-off: 10				
BM25	0.062	0.063	0.048	0.044
PACRR	0.080	0.084	0.062	0.063
MP	0.065	0.072	0.052	0.056
KNRM	0.067	0.064	0.051	0.051
ConvKNRM	0.080	0.077	0.064	0.060
Ours ($\alpha = 1$)	0.064	0.064	0.047	0.047
Ours ($\alpha = 0.5$)	**0.053**	**0.056**	**0.039**	**0.042**
Cut-off: 20				
BM25	0.060	0.062	0.048	0.046
PACRR	0.073	0.081	0.058	0.061
MP	0.063	0.068	0.052	0.054
KNRM	0.068	0.066	0.054	0.052
ConvKNRM	0.071	0.075	0.058	0.059
Ours ($\alpha = 1$)	0.058	0.062	0.045	0.046
Ours ($\alpha = 0.5$)	**0.051**	**0.053**	**0.040**	**0.041**
Cut-off: 30				
BM25	0.058	0.060	0.048	0.047
PACRR	0.070	0.078	0.057	0.060
MP	0.059	0.066	0.049	0.053
KNRM	0.068	0.067	0.055	0.053
ConvKNRM	0.069	0.074	0.057	0.059
Ours ($\alpha = 1$)	0.061	0.061	0.048	0.046
Ours ($\alpha = 0.5$)	**0.052**	**0.053**	**0.041**	**0.041**
Cut-off: 40				
BM25	0.057	0.060	0.048	0.047
PACRR	0.066	0.076	0.055	0.059
MP	0.055	0.064	0.045	0.051
KNRM	0.067	0.067	0.056	0.054
ConvKNRM	0.068	0.073	0.056	0.059
Ours ($\alpha = 1$)	0.059	0.061	0.048	0.047
Ours ($\alpha = 0.5$)	**0.053**	**0.053**	**0.043**	**0.041**

6 Conclusion

In this work, we introduced a bias-controllable ranking framework based on Backpack Language Models. By leveraging the interpretable structure of sense vectors, our method enables in-process gender bias mitigation during inference through a simple reweighting mechanism. Experimental results demonstrate that our approach significantly reduces gender bias while preserving ranking performance, offering a practical and effective solution for fair information retrieval.

7 Limitation

A key limitation concerns the scale of the Backpack model used in this study. For fair and meaningful comparison across ranking systems, it is important that models operate at a similar capacity and parameter budget. Discrepancies in model size can confound performance differences, making it difficult to attribute gains or losses to the ranking methodology itself rather than to model capacity. Additionally, while we address specific forms of social bias such as gender, broader categories of social bias in ranking—such as those related to race, age, or socioeconomic status—lack well-established and generalizable evaluation metrics. This limits the ability to comprehensively assess fairness interventions and their downstream impacts.

In addition, although our biascontrol re-ranking successfully alters the scores with respect to gender (see Appendix D), the simplified recruitment example shows that the overall ranking remains largely unchanged—similar to outcomes observed in other ranking experiments. This suggests that when absolute score differences are small, bias adjustment alone may be insufficient to meaningfully impact final rankings. To overcome this limitation, future work could explore more advanced ranking architectures, such as Rank-T5 (Zhuang et al., 2022) or other transformer-based rankers, and incorporate bias-adjusted scores from the Backpack-LM re-ranker to recalibrate the final rankings, potentially leading to fairer outcomes.

8 Ethical Considerations

This work addresses the ethical challenge of social bias in AI-powered ranking systems, with a particular focus on gender bias. Our proposed method aims to promote fairness and transparency in information retrieval by enabling controllable debiasing without retraining. While our evaluations focus on gender, we acknowledge that bias in AI spans multiple dimensions, including race, age, and socioeconomic status. Our method is designed to be generalizable, but further work is needed to adapt it to other forms of bias.

We emphasize that no demographic or identity-specific user data was used in this study. All benchmarks were conducted on publicly available datasets, and our experiments respect user privacy and data protection standards. Additionally, we caution that fairness interventions may interact in complex ways with model performance and societal norms, and their deployment should be accompanied by stakeholder input and domain-specific considerations.

References

Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das. 2019. Designing fair ranking schemes. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, page 12591276, New York, NY, USA. Association for Computing Machinery.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

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*, page 216224, Berlin, Heidelberg. Springer-Verlag.

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. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Sebastian Bruch, Xuanhui Wang, Michael Bendersky, and Marc Najork. 2019. An analysis of the softmax cross entropy loss for learning-to-rank with binary relevance. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, page 7578, New York, NY, USA. Association for Computing Machinery.

Diana Burgess and Eugene Borgida. 1999. Who women are, who women should be: Descriptive and prescriptive gender stereotyping in sex discrimination. *Psychology, Public Policy, and Law*, 5:665–692.

Yuen Chen, Vethavikashini Chithrra Raghuram, Justus Mattern, Rada Mihalcea, and Zhijing Jin. 2025. Causally testing gender bias in LLMs: A case study on occupational bias. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4984–5004, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 126134, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Naomi Ellemers. 2018. Gender stereotypes. *Annual Review of Psychology*, 69:275–298.

Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. 2020. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Preprint*, arXiv:2009.01334.

Jin Ge, Xuezhe Ma, Trevor Cohn, and Mark Johnson. 2021. Monot5: Mono-encoder pretraining for text generation and ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Madeline Heilman. 2012. Gender stereotypes and workplace bias. *Research in Organizational Behavior*, 32:113135.

John Hewitt, John Thickstun, Christopher D. Manning, and Percy Liang. 2023. Backpack language models. *Preprint*, arXiv:2305.16765.

Leonie Huddy and Nayda Terkildsen. 1993. Gender stereotypes and the perception of male and female candidates. *American Journal of Political Science*, 37:119–147.

Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. 2017. PACRR: A position-aware neural IR model for relevance matching. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1049–1058, Copenhagen, Denmark. Association for Computational Linguistics.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422446.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vassilis Christodoulou, and Sebastian Ruder. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: An easy-to-use python toolkit to support replicable ir research with sparse and dense representations. *Preprint*, arXiv:2102.10073.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023. Fine-tuning llama for multi-stage text retrieval. *Preprint*, arXiv:2310.08319.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage re-ranking with bert. *Preprint*, arXiv:1901.04085.

Jesutofunmi A Omiye, Jenna C Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *NPJ Digital Medicine*, 6(1):195.

Jahna Otterbacher. 2018. Addressing social bias in information retrieval. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 121–127. Springer.

Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text matching as image recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. volume 14, pages 1532–1543.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

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*, SIGIR 20. ACM.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333389.

Alireza Salemi and Hamed Zamani. 2024. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2395–2400.

Djamé Sarr and Pascal M. G. Appert. 2021. Debiasing ranking algorithms for fairness. *Proceedings of the 2021 International Conference on Learning Representations*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *Preprint*, arXiv:1906.08976.

Janet Swim, Eugene Borgida, Geoffrey Maruyama, and David Myers. 1989. Joan mckay versus john mckay: Do gender stereotypes bias evaluations? *Psychological Bulletin*, 105:409.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Suresh Venkatasubramanian, Anish Athalye, and Devavrat Shah. 2020. Fairness in ranking: A survey. *ACM Computing Surveys*, 53(4):1–38.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 5564, New York, NY, USA. Association for Computing Machinery.

Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. 2024. When search engine services meet large language models: visions and challenges. *IEEE Transactions on Services Computing*.

Soyoung Yoon, Eunbi Choi, Jiyeon Kim, Yireun Kim, Hyeongu Yun, and Seung won Hwang. 2024a. Listt5: Listwise reranking with fusion-in-decoder improves zero-shot retrieval. *ArXiv*, abs/2402.15838.

Soyoung Yoon, Eunbi Choi, Jiyeon Kim, Hyeongu Yun, Yireun Kim, and Seung won Hwang. 2024b. Listt5: Listwise reranking with fusion-in-decoder improves zero-shot retrieval. *Preprint*, arXiv:2402.15838.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo

Baeza-Yates. 2017. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM 17. ACM.

George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. 2022. Mitigating bias in search results through contextual document reranking and neutrality regularization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 25322538, New York, NY, USA. Association for Computing Machinery.

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*.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2022. Rankt5: Fine-tuning t5 for text ranking with ranking losses. *Preprint*, arXiv:2210.10634.

Appendix

A Datasets

We leverage the official training split of the MS MARCO passage ranking dataset (Bajaj et al., 2018) , which contains approximately 530,000 training queries and 6,800 development queries. The candidate passages are drawn from a large corpus of over 8.8 million documents, with each query annotated using binary relevance labels—1 indicating relevance and 0 indicating non-relevance. For evaluation, we utilize the official development set, using the Top-100 BM25-retrieved passages, obtained from the Pyserini (Lin et al., 2021) repository's prebuilt index.

B Bias Measurement Benchmark

(Rekabsaz and Schedl, 2020) proposes an effective framework for measuring gender bias in retrieval systems by assessing the bias in documents retrieved for gender-neutral queries. They introduce metrics to quantify this bias and compare neural ranking models with BM25. Their findings indicate that BM25 exhibits less gender bias according to both proposed metrics.

Baseline Models.

- **BM25** (Robertson and Zaragoza, 2009): default parameters $k_1 = 0.9$, $b = 0.4$.
- **KNRM** (Xiong et al., 2017), **MatchPyramid** (Pang et al., 2016), **PACRR** (Hui et al., 2017), **ConvKNRM** (Dai et al., 2018): all use 300-dimensional GloVe embeddings (Pennington et al., 2014) and follow hyperparameters from their original codebases.
- **BERT-Base/Large** (Nogueira and Cho, 2020; Devlin et al., 2019): fine-tuned on MS MARCO passages for 2 epochs, learning rate 2×10^{-5}, batch size 16, max sequence length 512.

Evaluation Metrics. This framework consists of two main components:

(i) Document Gender Magnitude Measurements: Gender magnitude is defined using predefined sets of gender-specific words. For instance, $G_f = \{\textit{she, woman, her}\}$ represents female-associated terms, and $G_m = \{\textit{he, man, him}\}$ represents male-associated terms. The gender magnitude of a document d is computed using two variants.

1. Term Frequency (TF): The magnitude is calculated as the sum of the logarithmic frequencies of gender-specific words:

$$\text{mag}_f(d) = \sum_{w \in G_f} \log(\#(w, d)) \tag{6}$$

2. Boolean: variant, the magnitude is set to 1 if any gender-specific word is present in the document, and 0 otherwise:

$$\text{mag}_f(d) = \begin{cases} 1, & \text{if } \sum_{w \in G_f} \#(w, d) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

Here, $\#(w, d)$ denotes the count of word w in document d. The same formulation is applied to compute $\text{mag}_m(d)$ using the male word set G_m.

(ii) Retrieval Gender Bias Metrics: To quantify gender bias in ranking results, two metrics are used.

1. Rank Bias (RaB): metric measures the difference between the female and male gender magnitudes in the top t ranked documents for a given query:

$$\text{RaB}_t(q) = \frac{1}{t} \sum_{i=1}^{t} \left(\text{mag}_f(d(q)_i) - \text{mag}_m(d(q)_i)\right) \tag{8}$$

2. Average Rank Bias (ARaB): aggregates RaB over the top t positions:

$$\text{ARaB}_t(q) = \frac{1}{t} \sum_{x=1}^{t} \text{RaB}_x(q) \tag{9}$$

Here, q denotes the query, $d(q)_i$ is the i-th document in the ranked list for query q, and t is the cutoff threshold. The final bias scores RaB_t and ARaB_t are computed by averaging these values across all queries.

C Ranking Evaluation

We measure performance using Mean Reciprocal Rank (MRR@10) (Voorhees and Tice, 2000), which evaluates how high the first relevant document appears in the top 10 results, and Normalized Discounted Cumulative Gain (NDCG@5, NDCG@10)(Järvelin and Kekäläinen, 2002) which assesses ranking quality while accounting for both document relevance and position.

D Additional Experiments

To illustrate the presence of gender bias and to evaluate our model's effectiveness in mitigating it, we present a simplified recruitmentstyle retrieval task. Below, we provide the query prompt and a concise table of candidate profiles employed in this experiment.

RecruitmentStyle Retrieval Task: The following prompt and candidate documents define our recruitmentstyle retrieval task. The query is designed to identify the most qualified candidates for a senior software engineering position based on their professional experience and achievements. Each document summarizes a candidate's background. We systematically vary the biascontrol parameter α to demonstrate the model's capacity to adjust rankings and reduce gender bias.

```
query = "Who are the most qualified candidates
    for a senior software engineering role
    based on experience and achievements?"
```

Listing 1: Query and documents used for ranking experiment

Table 3: Profile summary of candidates (with explicit gender labels).

Name	Gender	Key Qualifications
Sarah	Female	12+ yrs systems
John	Male	9 yrs full-stack
Emily	Female	15 yrs ML

Note that candidates 1 and 3 are female and, despite having more professional experience, the rankers do not prioritize them for a male-dominated role. However, varying the bias parameter α alters the scores, and the score differences for male-biased documents are almost always greater than those for female-biased documents. The resulting ranking outcomes are illustrated in Figure 3.

Although the model successfully reranks candidates independently of gender, the relative ordering remains unchanged in this scenario. To further enhance sensitivity to qualification differentials, one could adopt a more advanced ranking architecture (e.g., RankT5 (Zhuang et al., 2022; Yoon et al., 2024a) or another transformerbased ranker) and integrate the biasadjusted scores produced by the BackpackLM reranker into the final scoring mechanism.

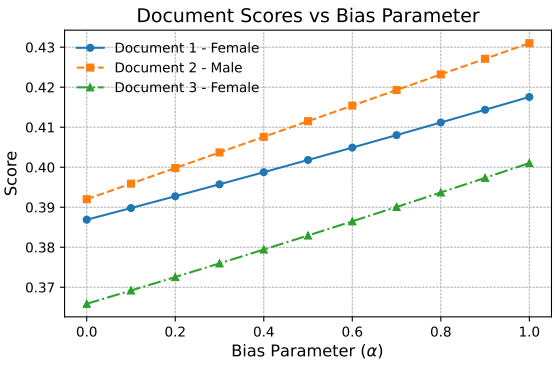

Figure 3: Ranking results of the candidates based on their qualifications.

For example, in this experiment the biascontrol adjustment for the first and third candidates differs by approximately 0.031 and 0.035, respectively, whereas the second candidate's score shifts by approximately 0.039. Leveraging these differential adjustments to recalibrate final ranking scores may yield a biasmitigated ordering but with a better accuracy.