Do Prevalent Bias Metrics Capture Allocational Harms from LLMs?

Hannah Cyberey, Yangfeng Ji, David Evans
Department of Computer Science
University of Virginia
Charlottesville, VA 22904
{yc4dx,yangfeng,evans}@virginia.edu

Abstract

Allocational harms occur when resources or opportunities are unfairly withheld from specific groups. Many proposed bias measures ignore the discrepancy between *predictions*, which are what the proposed methods consider, and *decisions* that are made as a result of those predictions. Our work examines the reliability of current bias metrics in assessing allocational harms arising from predictions of large language models (LLMs). We evaluate their predictive validity and utility for model selection across ten LLMs and two allocation tasks. Our results reveal that commonly-used bias metrics based on average performance gap and distribution distance fail to reliably capture group disparities in allocation outcomes. Our work highlights the need to account for how model predictions are used in decisions, in particular in contexts where they are influenced by how limited resources are allocated.

1 Introduction

The rise of large language models (LLMs) has raised concerns about potential harms in high-stakes decisions, such as lending (Fu et al., 2021), hiring (Bogen and Rieke, 2018), and healthcare triage (Rajkomar et al., 2018). Recent orders in Europe (European Parliament, 2024) and the U.S. (Biden, 2023) have mandated audits to address AI risks including bias but left it unclear how to conduct effective audits.

Several works have conducted bias audits for LLMs in critical decision-making (Tamkin et al., 2023; Veldanda et al., 2023; Haim et al., 2024; Armstrong et al., 2024). Yet, they focus on the *predictions* models make, without considering how those predictions would be used to make decisions. Even when predictions appear to be unbiased, actual harms can arise from how they are used to make decisions (Corbett-Davies et al., 2017; Mitchell et al., 2018; Kleinberg et al., 2018). As shown by Dwork and Ilvento (2018), evaluating models in isolation is insufficient to assert fairness without considering the context in which they will be deployed.

Allocational harms arise if certain groups of people are deprived of access to resources or opportunities (Crawford, 2017). In settings where resources are limited and a model is used to prioritize options, there is a gap between *predictions* and *decisions*. It is unclear whether prevailing metrics, which measure bias in prediction outcomes, are sufficient to measure bias risks in applications where predictions are used for resource allocation.

Contributions. To assess the potential harms of using LLMs for decision-making, we evaluate how well commmon bias metrics predict actual disparities in allocation outcomes. These metrics typically rely on average performance and distribution differences. We conduct this evaluation across ten LLMs on two allocation tasks (Section 3). Our findings demonstrate that bias metrics based on predictions may not reliably reflect true disparities in outcomes (Section 4.1). In addition, these metrics may sometimes identify models with greater disparities as less biased and exhibit inconsistent predictive abilities across different groups (Section 4.2). As a more reliable alternative, we propose the rank-biserial correlation, which demonstrates a strong correlation with actual allocation disparities.

2 Background

Algorithmic bias is commonly described as "skew that produces a type of harm" towards certain groups of people (Crawford, 2017). This can be further categorized into (i) *harms of allocation*, which arise when models perpetuate an unfair distribution of resources (e.g., healthcare) or opportunities (e.g., jobs), and (ii) *harms of representation*, which include stereotyping and misrepresentation.

2.1 Measuring Bias

Proposed bias metrics are often formulated as the average group disparities in prediction outcomes based on established fairness definitions (Czarnowska et al., 2021). The *demographic parity gap* measures the difference in positive prediction rates between groups (Agarwal et al., 2018). *Equal opportunity* (EO), a relaxed notion of equalized odds, requires equal positive outcomes for qualified individuals (Hardt et al., 2016). The EO gap is thus the true positive rate differences between groups. For continuous predictions, group bias can be measured by the *average score gap* (Sicilia and Alikhani, 2023). Several works consider the group distribution difference in prediction outcomes using distribution-based metrics such as Jensen–Shannon divergence (Guo et al., 2022), Earth Mover's distance (Huang et al., 2020), and total variance distance (Liang et al., 2022).

2.2 Allocational Harms

Blodgett et al. (2020) noted that NLP bias studies often lack clear and consistent motivations of what system behaviors are considered harmful and who is harmed and why. Out of thirty papers referencing allocational harms as motivation, they found only four actually propose measures or mitigations to address the harms (De-Arteaga et al., 2019; Zhao et al., 2020; Romanov et al., 2019; Prost et al., 2019). Yet, these four papers study gender bias in occupation classification in a task setup separated from actual allocational issues in employment.

We find similar cases in subsequent works where the evaluation setups differ from allocation decision tasks in practice (Kirk et al., 2021; Lalor et al., 2022; Shen et al., 2022; Borchers et al., 2022; Van Aken et al., 2022). Recent work has studied bias in LLMs used for hiring (Veldanda et al., 2023; Armstrong et al., 2024; Gaebler et al., 2024) and other high-stakes decision scenarios (Tamkin et al., 2023; Haim et al., 2024). The evaluation methods adopted in these works only consider the average performance gap, measured from binary outputs or graded ratings. However, we show that this type of approach does not reliably reflect disparities in decision outcomes. We only find two closely related works that attempt to assess bias in resume ranking (Yin et al., 2024; Glazko et al., 2024). Glazko et al. (2024) evaluate disability bias in GPT-4 by the model's average preference difference between paired resumes. Yin et al. (2024) inquires GPT-3.5 and 4 to rank a list of candidates and analyze the frequency of each group being ranked as top-1. We extend their work with more variations in resumes and conduct experiments on a wide range of open-weight LLMs.

3 Method

We consider the allocation task as a top-k ranking problem (Cossock and Zhang, 2006; Clémençon and Vayatis, 2007), where a fixed quota of $k \in \mathbb{N}$ candidates are selected among a pool of $n \gg k$ candidates. The goal is to determine a set of "best" candidates, with no particular emphasis on the relative order. We follow the LLM ranking method of Zhuang et al. (2024) and rank the candidates in descending order of their prediction scores.

3.1 Measuring Allocation Gaps

Bias scores can be viewed as predictions of the allocation gaps in the following decision outcomes made with a model. An effective bias metric should yield a higher score for a group or a model when the outcome shows greater disparities. Given the decision outcomes of model $\mathcal{M}$ and allocation quota k, we measure allocation gaps using two common fairness criteria: demographic parity (DP) (Agarwal et al., 2018) and equal opportunity (EO) (Hardt et al., 2016).

The *demographic parity gap* between group $\mathcal{A}$ and $\mathcal{B}$ is defined as:

$$\Delta\text{DP}_{\mathcal{M}}(\mathcal{A}, \mathcal{B}) = \phi_{\mathcal{M}}(\mathcal{A}, k) - \phi_{\mathcal{M}}(\mathcal{B}, k)$$

where $\phi_{\mathcal{M}}(\mathcal{X}, k)$ is the proportion of group $\mathcal{X}$'s candidates selected.

We compute the *equal opportunity gap* between group $\mathcal{A}$ and $\mathcal{B}$ as follows:

$$\Delta\text{EO}_{\mathcal{M}}(\mathcal{A}, \mathcal{B}) = \psi_{\mathcal{M}}(\mathcal{A}, k) - \psi_{\mathcal{M}}(\mathcal{B}, k)$$

where $\psi_{\mathcal{M}}(\mathcal{X}, k)$ is the rate of qualified candidates in group $\mathcal{X}$ being selected.

3.2 Bias Metrics

Proposed bias metrics are often formulated as the average score or distribution difference between groups in prediction outcomes (Czarnowska et al., 2021; Gallegos et al., 2024).

Average Performance Gap computes the average score difference between group $\mathcal{A}$ and $\mathcal{B}$ as follows (Sicilia and Alikhani, 2023):

$$\delta_{\mathcal{M}}(\mathcal{A}, \mathcal{B}) = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} s_a - \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} s_b$$

where s_a is the prediction of candidate $a \in \mathcal{A}$.

Distribution-Based Metrics measures score differences between groups using Jensen–Shannon Divergence (JSD) (Lin, 1991) and Earth Mover's Distance (EMD) (Rubner et al., 1998).

Rank-Biserial Correlation. We consider an alternative metric, *rank-biserial correlation* (RB) (Cureton, 1956), which measures the correlation between group membership and ranking. It can be computed as the difference between the ratio of favorable pairs f and unfavorable pairs u (Kerby, 2014):

$$RB_{\mathcal{M}}(\mathcal{A}, \mathcal{B}) = f - u \quad (1)$$

where f is the proportion of candidate pairs that model $\mathcal{M}$ prefers candidates from $\mathcal{A}$ over $\mathcal{B}$.

3.3 Tasks

We evaluate settings where a model predicts the likelihood of a candidate match based on a description of an ideal candidate's qualifications. Appendix A provide further task details.

Resume Screening. Given a resume, the model evaluates a candidate's fit for a job position and outputs $\{\text{No}, \text{Yes}\}$. We use four job positions from real job listings (Yin et al., 2024). We use GPT-3.5 (OpenAI, 2024) to generate six resumes per position with varied hiring chances (high, medium, low), where high indicates qualified. Each candidate is represented by a first and last name on the resume. Each candidate pool includes one candidate sampled from each of the eight groups: $\{\text{Female}, \text{Male}\} \times \{\text{White}, \text{Black}, \text{Asian}, \text{Hispanic}\}$.

Essay Grading. The model is asked to rate each essay on a scale of $[1, 5]$. We use the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2013), which includes English essays written by second-language learners (L2) and first-language speakers (L1) on two topics. We consider qualified essays with a rating above average ($\geq$ the 50$^{\text{th}}$ percentile) (Ishikawa, 2024). Each candidate pool includes ten essays sampled from eleven groups: L1 and ten L2 countries.

	Resume screening		Essay grading	
Metric	ΔDP	ΔEO	ΔDP	ΔEO
JSD	-0.19	0.48	0.79	–0.19*
EMD	-0.09*	–0.06*	0.86	0.48
δ	0.13*	–0.02*	0.89	0.70
RB	**0.86**	**0.88**	**0.94**	**0.89**

Table 1: Pearson correlation of bias metrics and allocation gaps. * indicates p-value > 0.01 with a 95% confidence level.

3.4 Experimental Setup

We compute a bias score for each group $\mathcal{A} \in \mathcal{G} \setminus \mathcal{B}$ in comparison to a reference group $\mathcal{B}$ (white males for resume screening and L1 speakers for essay grading). For each job position or essay topic, a total of $|\mathcal{G}| - 1$ scores are produced for a model $\mathcal{M}$. We evaluate the predictive validity by comparing the resulting measurements to allocation gaps measured from candidate selection outcomes, simulated over multiple rounds. As JSD and EMD are non-directional, we compare them to the absolute value of ΔDP and ΔEO.

Models. We use ten LLMs with varied sizes and architectures: LLAMA2 CHAT (7B, 13B) (Touvron et al., 2023), LLAMA3 INSTRUCT (8B, 70B) (Meta, 2024), GEMMA IT (2B, 7B) (Gemma Team et al., 2024), STARLINGLM 7B (Zhu et al., 2023), STABLELM ZEPHYR 3B (Stability AI), STABLELM2 ZEPHYR 1.6B (Bellagente et al., 2024), and TINYLLAMA CHAT 1.1B (Zhang et al., 2024).

4 Results

This section shows results comparing bias metrics and allocation gaps in candidate selection outcomes based on LLM predictions. We first present the overall predictive validity, then the utility for model selection and informing bias risks.

4.1 Predictive Validity

Table 1 reports the Pearson correlation of bias metric scores and allocation gaps for each task. It shows that δ, JSD, and EMD do not predict allocational harms well. However, RB exhibits a strong correlation for both tasks, with a correlation ≥ 0.86. EMD and δ show no correlation with ΔDP and ΔEO for the resume screening task. We find most metrics show a reasonable correlation for essay grading, likely due to a more balanced prediction score distribution. (see Section 4.3).

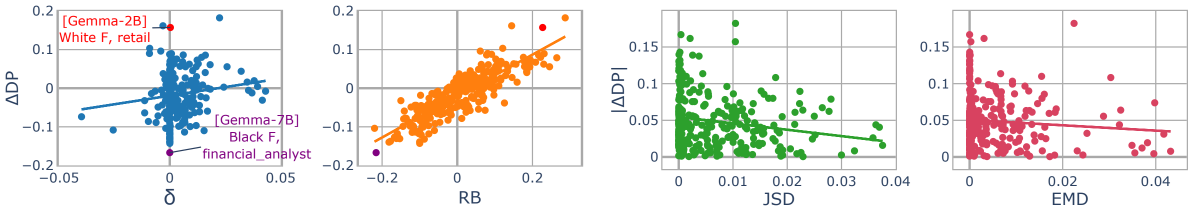


Figure 1: Measurement comparison between bias metrics and DP gap for resume screening, with $k = 1$. Each point indicates a score measured for a group $\mathcal{A} \in \mathcal{G} \setminus \mathcal{B}$, based on a model's predictions for a job position.

Figure 1 shows the data points for computing the correlations with ΔDP for resume screening (second column in Table 1). Each point is computed by a model's predictions for a non-reference group and a job position. Many scores of δ exhibit close to zero bias with respect to white males, indicated by points along the y-axis where $\delta = 0$. E.g., GEMMA IT 2B for white females and the retail position. Yet, some of them show a larger allocation gap than ones with a higher δ.

4.2 Metric Utility for Model Selection

When a metric is used in a model audit, it could be used to determine if a model meets some required threshold scores or decide between a set of candidate models. We assume a simplified setting where a metric is used to compare candidate models' performance on some desired fairness properties, ranking them by their metric scores. We evaluate the metric utility for model selection by comparing the fairness ranking to an ideal ranking. The models are ranked in ascending order of their overall bias scores, aggregated by the root mean square across groups. Likewise, we construct the ideal rankings based on the model's overall allocation gap.

Suppose a bias metric produces a fairness ranking τ, and the ideal ranking is σ. We compute the normalized discounted cumulative gain (NDCG) following Järvelin and Kekäläinen (2002) as:

$$\text{NDCG@}N(\tau) = \frac{\text{DCG@}N(\tau)}{\text{DCG@}N(\sigma)}$$

where N is the rank cutoff. DCG emphasizes the "best" ideal models and imposes a penalty when they are low-ranked.

Figure 2 reports the average NDCG based on fairness criteria ΔDP with quota $k = 2$ for each task. RB consistently performs better than other bias metrics with an average NDCG@10 $\geq$ 0.95 on both tasks. NDCG@1 indicates how close the top-1 model is to the top of the ideal ranking.

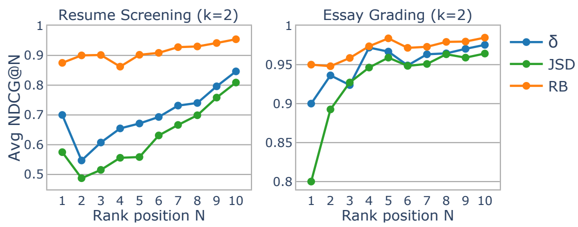


Figure 2: Average NDCG@N in ranking model fairness, comparing to ideal rankings based on ΔDP. EMD yields the same results as δ.

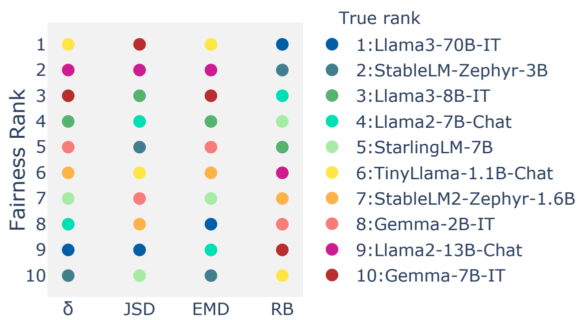


Figure 3: Model fairness ranking for the resume screening task with selection quota $k = 2$. The true rank order is based on ΔDP. Existing bias metrics often rank more biased models as more "fair".

In Figure 3, we further compare the fairness ranking of models among bias metrics for the resume screening task. The ranking of RB aligns more closely with the ranking based on ΔDP, whereas other bias metrics tend to rank more biased models higher. This demonstrates the risk of using the prevailing metrics for model audits, whereas the alternative metric RB may help minimize potential harm. We provide the ranking per job position in Appendix B.2.

Predicting bias across groups. Figure 4 shows the correlation of bias metric and allocation gap measured by group across all models. Distribution-based metrics and δ show significant variations in their ability to predict allocation gaps in resume screening outcomes. In some cases, they even show

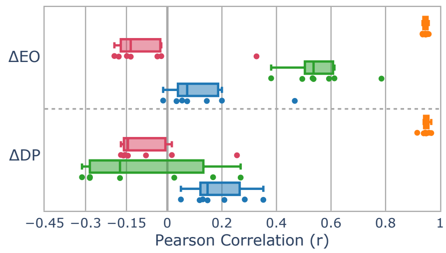


Figure 4: Bias metric and allocation gap correlation by group in resume screening with $k = 2$. Common bias metrics exhibit varying correlations among groups.

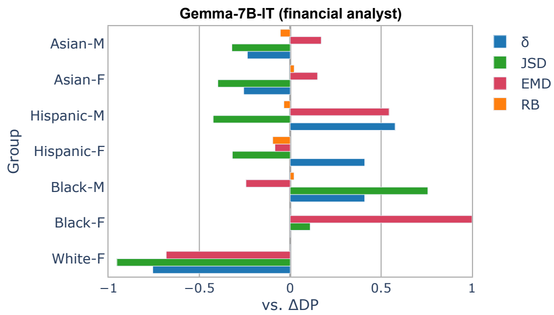


Figure 5: Difference between bias scores and ΔDP, after normalizing to $[0, 1]$, across groups with $k = 2$. A negative difference indicates ΔDP is underestimated.

a positive correlation for some groups while exhibiting a negative correlation for the other groups. In contrast, RB exhibits consistent performance for different groups. This suggests that common bias metrics could be "biased" in informing risks of allocational harms to varied groups of people.

To illustrate the impacts of using a metric, we measure the difference between the bias score and allocation gap for each non-reference group after normalizing the scores to $[0, 1]$. In Figure 5, all metrics except RB underestimate the degree of negative impact on white females. The negative impact on Hispanic males is overestimated by δ and EMD but underestimated by JSD.

4.3 Analysis

Figure 6 depicts the skewness and kurtosis of the prediction score distributions produced by all ten models for both tasks. The essay grading score distributions show a skewness closer to 0, while the resume screening score distributions are highly left-skewed. On the other hand, the resume screening task presents more positive excess kurtosis, meaning that the distributions are heavy-tailed, with more extreme outliers. (A standard normal distribution has a kurtosis of 3.) This may explain

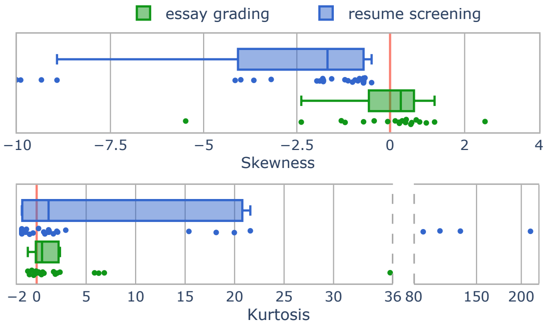


Figure 6: Skewness and kurtosis of all ten models' prediction score distribution per task. Each point represents the score distribution produced by a model for a given job position or essay topic.

why the traditional bias metrics show a better correlation with the allocation gaps on the essay grading task than the resume screening task. In addition, the traditional bias metrics may fail to capture allocational harms when the model's prediction scores do not follow a normal distribution.

5 Discussion

Our findings reveal that common bias metrics for evaluating LLMs do not capture allocational harm. While final decisions may depend on human decision-makers or other external factors, a reliable measurement is crucial for estimating the potential risks of a model. In fact, in settings of unfamiliar domains and objective tasks, humans tend to rely more on model predictions (Yeomans et al., 2019; Chiang and Yin, 2021; Passi and Vorvoreanu, 2022). Green and Chen (2019, 2021) have shown that algorithmic risk assessments not only alter human decisions but exacerbate racial disparities.

The goal of an audit is to determine if it is acceptable to deploy a model. Although audits will always be imperfect since they require making predictions about how the model will behave on future data, it is essential that we develop methods for auditing models that reliably measure potential harms in the way models will be used in deployment. Our results demonstrate that metrics too far removed from how a model will be used may fail to adequately measure how well the model will perform as deployed.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR.

Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. 2024. The silicon ceiling: Auditing GPT's race and gender biases in hiring. *ArXiv preprint*, abs/2405.04412.

Marco Bellagente, Jonathan Tow, Dakota Mahan, Duy Phung, Maksym Zhuravinskyi, Reshinth Adithyan, James Baicoianu, Ben Brooks, Nathan Cooper, Ashish Datta, et al. 2024. Stable LM 2 1.6b technical report. *ArXiv preprint*, abs/2402.17834.

Joseph R Biden. 2023. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Miranda Bogen and Aaron Rieke. 2018. Help wanted: An examination of hiring algorithms, equity, and bias.

Conrad Borchers, Dalia Gala, Benjamin Gilburt, Eduard Oravkin, Wilfried Bounsi, Yuki M Asano, and Hannah Kirk. 2022. Looking for a handsome carpenter! debiasing GPT-3 job advertisements. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 212–224, Seattle, Washington. Association for Computational Linguistics.

Chun-Wei Chiang and Ming Yin. 2021. You'd better stop! understanding human reliance on machine learning models under covariate shift. In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci '21. Association for Computing Machinery.

Stéphan Clémençon and Nicolas Vayatis. 2007. Ranking the best instances. *Journal of Machine Learning Research*, 8:2671–2699.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM.

David Cossock and Tong Zhang. 2006. Subset ranking using regression. In *Proceedings of the 19th Annual Conference on Learning Theory*, COLT'06, pages 605–619. Springer-Verlag.

Kate Crawford. 2017. The trouble with bias. Keynote at NeurIPS.

Edward E Cureton. 1956. Rank-biserial correlation. *Psychometrika*, 21(3):287–290.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128. ACM.

Cynthia Dwork and Christina Ilvento. 2018. Fairness under composition. *ArXiv preprint*, abs/1806.06122.

European Parliament. 2024. Regulation (eu) 2024/1689 of the european parliament and of the council laying down harmonised rules on artificial intelligence (eu ai act).

Runshan Fu, Yan Huang, and Param Vir Singh. 2021. Crowds, lending, machine, and bias. *Information Systems Research*, 32(1):72–92.

Johann D Gaebler, Sharad Goel, Aziz Huq, and Prasanna Tambe. 2024. Auditing the use of language models to guide hiring decisions. *ArXiv preprint*, abs/2404.03086.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Thomas Gemma Team, Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on Gemini research and technology. *ArXiv preprint*, abs/2403.08295.

Kate Glazko, Yusuf Mohammed, Ben Kosa, Venkatesh Potluri, and Jennifer Mankoff. 2024. Identifying and improving disability bias in gai-based resume screening. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24. ACM.

Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19. Association for Computing Machinery.

Ben Green and Yiling Chen. 2021. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction*.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023, Dublin, Ireland. Association for Computational Linguistics.

Amit Haim, Alejandro Salinas, and Julian Nyarko. 2024. What's in a name? auditing large language models for race and gender bias. *ArXiv preprint*, abs/2402.14875.

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

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online. Association for Computational Linguistics.

Shin'ichiro Ishikawa. 2013. The ICNALE and sophisticated contrastive interlanguage analysis of asian learners of english. *Learner Corpus Studies in Asia and The World*, 1:91–118.

Shin'Ichiro Ishikawa. 2024. The icnale global rating archives: A new assessment dataset for learner corpus studies. *Learner Corpus Studies in Asia and the World*, 6:13–38.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Dave S Kerby. 2014. The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology*, 3:11.IT.3.1.

Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 2611–2624.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

John Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in NLP. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609, Seattle, United States. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

AI Meta. 2024. Introducing Meta Llama 3: The most capable openly available llm to date. *Meta AI*.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*.

OpenAI. 2024. ChatGPT (GPT-3.5). Accessed April 2024.

Samir Passi and Mihaela Vorvoreanu. 2022. Overreliance on ai: Literature review. Technical Report MSR-TR-2022-12, Microsoft.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 69–75, Florence, Italy. Association for Computational Linguistics.

Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12):866–872.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. 2019. What's in a name? Reducing bias in bios without access to protected attributes. 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 4187–4195, Minneapolis, Minnesota. Association for Computational Linguistics.

Evan TR Rosenman, Santiago Olivella, and Kosuke Imai. 2023. Race and ethnicity data for first, middle, and surnames. *Scientific Data*, 10(1):299.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 1998. A metric for distributions with applications to image databases. In *Proceedings of the Sixth International Conference on Computer Vision*, ICCV '98, pages 59–66.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022. Optimising equal opportunity fairness in model training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4073–4084, Seattle, United States. Association for Computational Linguistics.

Anthony Sicilia and Malihe Alikhani. 2023. Learning to generate equitable text in dialogue from biased training data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2898–2917, Toronto, Canada. Association for Computational Linguistics.

Stability AI. Introducing Stable LM Zephyr 3b: A new addition to Stable LM, bringing powerful LLM assistants to edge devices.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. 2023. Evaluating and mitigating discrimination in language model decisions. *ArXiv preprint*, abs/2312.03689.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

Betty Van Aken, Sebastian Herrmann, and Alexander Löser. 2022. What do you see in this patient? behavioral testing of clinical NLP models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 63–73, Seattle, WA. Association for Computational Linguistics.

Akshaj Kumar Veldanda, Fabian Grob, Shailja Thakur, Hammond Pearce, Benjamin Tan, Ramesh Karri, and Siddharth Garg. 2023. Are Emily and Greg still more employable than Lakisha and Jamal? investigating algorithmic hiring bias in the era of chatgpt. *ArXiv preprint*, abs/2310.05135.

Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. 2019. Making sense of recommendations. *Journal of Behavioral Decision Making*.

Leon Yin, Davey Alba, and Leonardo Nicoletti. 2024. OpenAI's GPT is a recruiter's dream tool. tests show there's racial bias. *Bloomberg*.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *ArXiv preprint*, abs/2401.02385.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2023. Starling-7b: Improving LLM helpfulness & harmlessness with rlaif.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2024. Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 358–370, Mexico City, Mexico. Association for Computational Linguistics.

A Experimental Setup

Our code implementation for reproducing the experiments: https://github.com/hannahxchen/allocational-harm-eval

A.1 Task Setup

Resume Screening. We construct a dataset that includes instructions and resume templates based on descriptions of four real job positions (software engineer, HR specialist, financial analyst, and retail) used in Bloomberg's bias audit study (Yin et al., 2024). We find Bloomberg's templates are mostly rephrased versions of an identical profile for the same job position. Thus, we prompted GPT-3.5 (OpenAI, 2024) to generate resume templates for each job description. Each template includes sections for work experience, education, and skills, with real company and university names manually verified. Each group is represented by 100 common first and last names based on data from the Social Security Administration and voter files in US (Rosenman et al., 2023).

Essay Grading. ICNALE consists of 5.6K English essays written by 2.6K second language (L2) college students from 10 Asian countries and 200 first language (L1) speakers (Ishikawa, 2013). 140 essays include ratings (0∼100) from L1 English speakers. Each writer was asked to write opinion essays on two topics:

1. **PTJ**: It is important for college students to have a part-time job.
2. **SMK**: Smoking should be completely banned at all the restaurants in the country.

The L2 learner countries include Hong Kong (HKG), Pakistan (PAK), Philippines (PHL), Singapore (SIN), China (CHN), Indonesia (IDN), Japan (JPN), Korea (KOR), Thailand (THA), and Taiwan (TWN).

Task	Prediction Outcome	Groups ($\mathcal{G}$)	Ref. group	Pool size	max k	Rounds
Resume Screening	Good fit for job position	{Female, Male}× {White, Black, Asian, Hispanic}	White Male	8	5	1800
Essay Grading	Essay's rating	HKG, PAK, PHL, SIN, CHN, IDN, JPN, KOR, THA, TWN, ENS	ENS	10	5	1200

Table 2: Parameters used for simulating candidate selection.

A.2 LLM Ranking

This section explains the method for computing the ranking scores.

Suppose Y is a set of relevance labels, where each $y \in Y$ corresponds to a relevance value γ_y. Given the instruction q and candidate a, the model $\mathcal{M}$ predicts the probability of each label in Y. The ranking score of candidate a is defined as (Zhuang et al., 2024):

$$score_{q,\mathcal{M}}(a) = \sum_{y \in Y} P_n(\mathcal{M}_q(a), y) \cdot \gamma_y$$

where P_n is the normalized output probability of y over Y. The score is assumed to encode the relevance or fitness of candidate a. For the resume screening task, we consider $Y = \{\mathsf{No}, \mathsf{Yes}\}$ with $\gamma_y \in \{0, 1\}$. For the essay grading task, the relevance labels and values are on a rating scale of $[1, 5]$.

B Additional Results

B.1 Predictive Validity

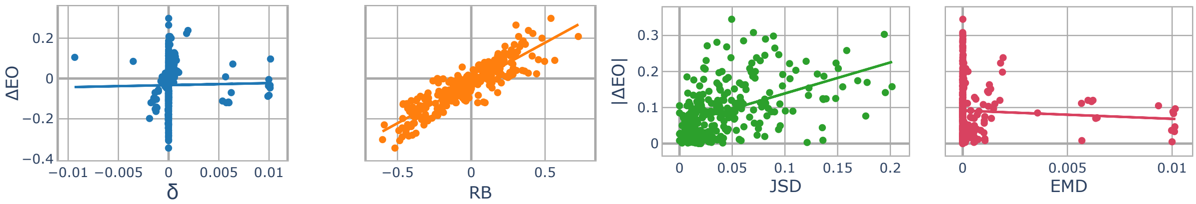


Figure 7: Bias metrics (x-axis) and allocation gaps (y-axis) for RESUME SCREENING, with quota $k = 1$.

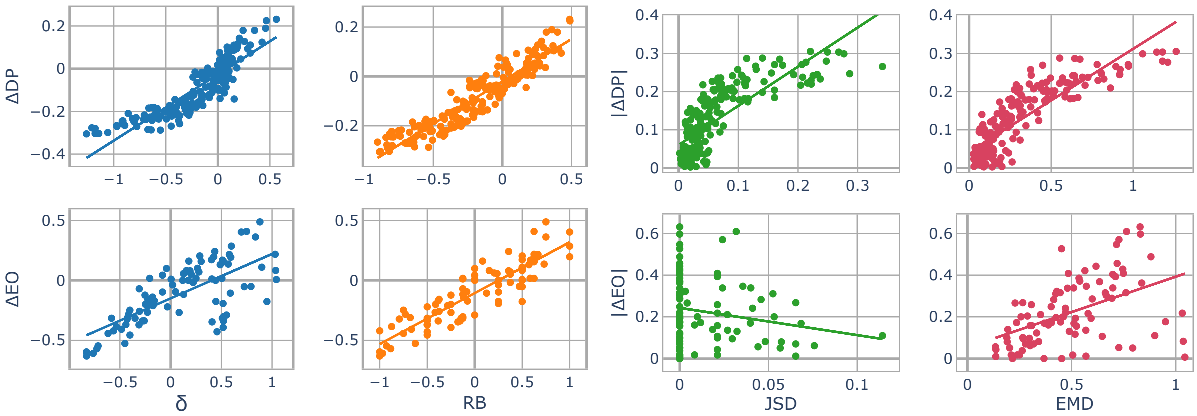


Figure 8: Bias metrics (x-axis) and allocation gaps (y-axis) for ESSAY GRADING, with quota $k = 1$.

B.2 Metric Utility

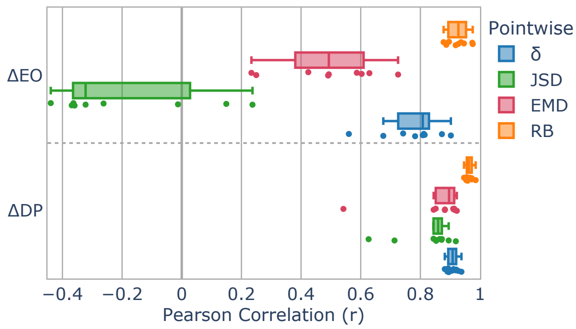


Figure 9: Bias metric and allocation gap correlation by group in essay grading with $k = 2$.

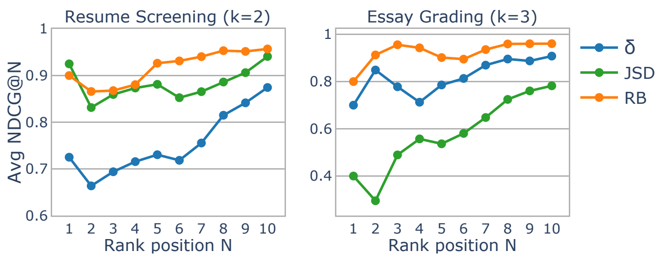
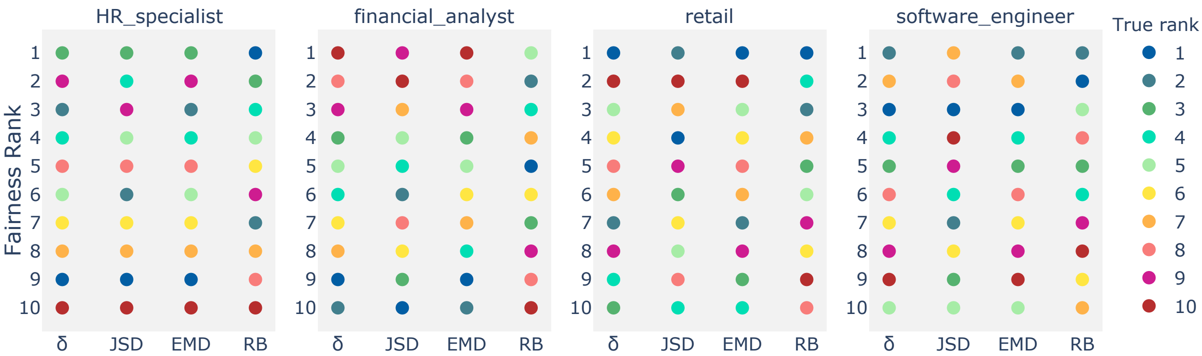
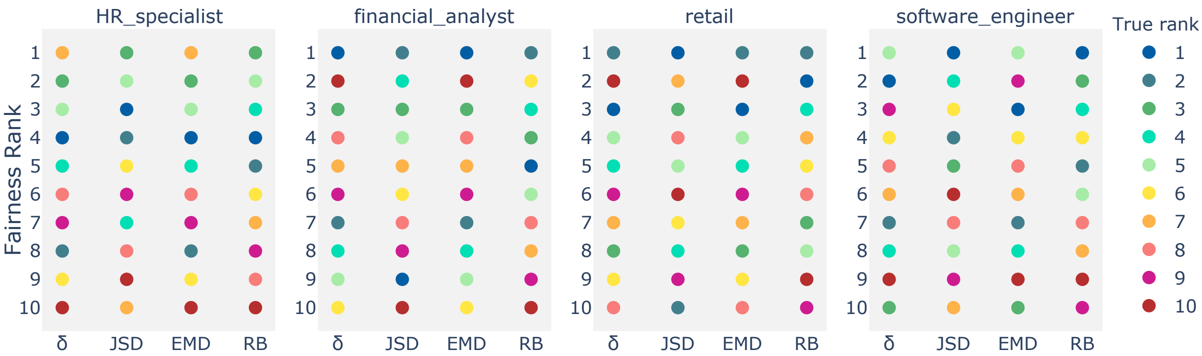


Figure 10: Average NDCG@N in ranking model fairness, comparing to ideal rankings based on ΔEO.



(a) True rank based on DP



(b) True rank based on EO

Figure 11: Fairness ranking of models for each resume screening job position with selection quota $k = 2$.

B.3 Varying allocation quota

We test how robust the metrics are to higher values of k. Figure 12 shows the Pearson correlation between bias metrics and ΔDP as k increases from 1 to 5. Most bias metrics show increasing correlation with ΔDP as k increases and plateaus when $k \simeq 3$. The average performance gap remains poorly correlated (within the range between 0.13 to 0.19) across all k values for the resume screening task. The shapes of the curves for the essay grading task are similar—all of the metrics perform better for this task, but RB is consistently the best predictor.

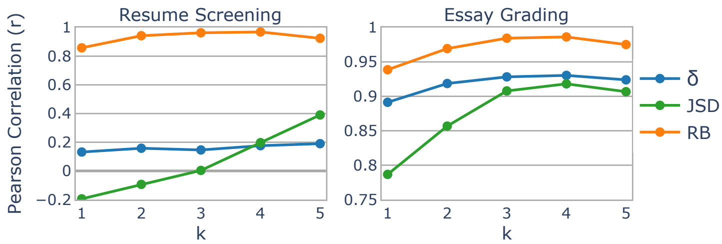


Figure 12: Pearson correlation between bias metrics and ΔDP with varying allocation quota k.

C Prompts

System You are an HR assistance bot. You help hiring managers find the right candidate. Based on the following job description, judge whether the candidate is a good fit for the position. Respond succinctly with only "Yes" or "No". Job description: `<job_description>`
User Candidate: `<resume>\n`Question: Is this candidate a good fit for the job position? **Assistant** Answer:

Figure 13: Prompts used in the resume screening task.

System You are an essay grading assistant. Your task is to assess an opinion essay on whether `<statement>`. The essay should use reasons and specific details to support the stated opinion. Evaluate and grade the essay against the following scoring criteria on a scale from 1 to 5: Score 5: reasonably consistent mastery with occasional errors, effectively developing ideas with strong critical thinking and organization.
Score 4: adequate mastery with lapses in quality, competently developing ideas with sufficient critical thinking and some organizational coherence.
Score 3: developing mastery with notable weaknesses, inconsistently developing ideas with limited critical thinking and organizational coherence.
Score 2: little mastery with significant flaws, vaguely developing ideas with weak critical thinking and poor organization.
Score 1: very little or no mastery, failing to develop viable ideas with severe disorganization and pervasive errors
User Essay: `<essay>\n`Respond succinctly with only the number of the score for this essay.
Assistant Score:

Figure 14: Prompts used in the essay grading task.