Evaluating Bias Mitigation Techniques in Large Language Models: A Study on the Accuracy-Fairness Trade-off in Decision-Making

Priyanka Raikwad
Member, Symbiosis Skills and Professional University

Abstract- **The rise of large language models (LLMs) has brought significant progress across many decision-making domains, from hiring recommendations to legal risk assessments. However, as these models are increasingly integrated into real-world systems, concerns around fairness and unintended bias have become more prominent. While model accuracy is often the primary benchmark during development, fairness remains an underrepresented metric—despite its crucial impact on equitable decision-making.**

This study investigates the trade-offs between optimizing for accuracy and ensuring fairness in LLMs used for decision-making. We conduct a systematic evaluation of fairness-aware interventions applied to pre-trained transformer models across real-world, demographically sensitive datasets. Through comparative analysis, the study offers insights into how different mitigation strategies impact both performance and fairness, informing more responsible and equitable deployment of LLMs in practice.

Index Terms- **Large Language Models, Fairness, Bias Mitigation, Accuracy-Fairness Trade-off, Decision-Making, Fairness-Aware Interventions, Model Optimization, Responsible AI, Fairness Metrics**

I. INTRODUCTION

The rapid development and application of Large Language Models (LLMs) such as GPT-3 and BERT (Vaswani et al., 2017) have led to their increasing deployment in a wide range of sectors, including healthcare, finance, and human resources. These models have demonstrated remarkable proficiency in generating human-like text, automating complex tasks, and providing decision-making support in fields that once required human expertise. However, their usage in critical decision-making processes raises serious ethical concerns, particularly related to fairness and the potential for bias. LLMs, trained on vast datasets sourced from the internet, inherit biases that exist in these datasets. As a result, when applied in areas such as recruitment, lending, and law enforcement, they can exacerbate discriminatory practices, leading to unfair treatment of individuals based on gender, race, or socio-economic background (Binns, 2018).

Addressing these biases is critical to ensuring equal opportunities and minimizing harm in decision-making tasks. While there has been significant progress in developing fairness-aware techniques (Zafar et al., 2017), such as pre-processing, in-processing, and post-processing interventions, the trade-off between optimizing for accuracy and fairness remains poorly understood, particularly in large-scale LLMs. Moreover, the computational cost and effectiveness of these fairness interventions, especially in resource-constrained environments, have not been adequately explored. This research seeks to empirically evaluate the impact of fairness-aware interventions on LLMs when applied to decision-making tasks. Rather than introducing new algorithms, we aim to systematically assess existing mitigation techniques across multiple datasets and architectures, offering practical insights for developing AI systems that are both accurate and fair. The study emphasizes reproducibility, statistical rigor, and cross-domain consistency to support responsible deployment of LLM

II. RESEARCH GAP

Despite the growing application of Large Language Models (LLMs) in automated decision-making systems, existing research predominantly focuses on isolated fairness interventions—such as pre-processing, in-processing, and post-processing techniques—without adequately addressing the intricate trade-off between accuracy and fairness. Many studies evaluate these interventions in controlled settings or on narrow datasets, limiting the generalizability of their findings to real-world applications. Furthermore, there is a lack of standardized metrics to simultaneously assess both accuracy and fairness in LLMs, hindering the development of comprehensive evaluation frameworks.

Additionally, the computational efficiency of fairness interventions, especially in large-scale LLMs, remains underexplored, posing challenges for their practical deployment in resource-constrained environments. To address these gaps, this research conducts a comprehensive empirical analysis of fairness interventions, systematically evaluating their impact on both accuracy and fairness across diverse LLM architectures and datasets.

III. RESEARCH OBJECTIVES

- Assess the Accuracy-Fairness Trade-off: Investigate how optimizing accuracy impacts fairness in LLMs and identify strategies for balancing both in decision-making tasks.
- Implement Fairness Interventions: Explore pre-processing, in-processing, and post-processing techniques to mitigate biases and evaluate their impact on the trade-off between accuracy and fairness.
- Create Standardized Evaluation Metrics: Develop a consistent methodology for evaluating accuracy and fairness in parallel, considering domain-specific biases.
- Foster Responsible AI: Provide insights into creating transparent, fair, and accurate AI systems that reduce bias and ensure ethical decision-making.
- Establish a Reproducible Evaluation Protocol: Design a systematic approach to apply and compare fairness interventions using open-source libraries and best practices.

IV. METHODOLOGY

This research adopts an empirical, multi-phase approach to investigate the interplay between accuracy and fairness in large language models (LLMs) across various decision-making tasks.

A. Dataset Selection and Bias Analysis

Benchmark datasets relevant to decision-making contexts (e.g., employment screening, loan approvals) were selected and analysed for potential biases using statistical tools and fairness diagnostics.

B. Model Selection and Baseline Construction

Pre-trained LLMs such as GPT-based or BERT-based architectures were fine-tuned on the selected datasets. Baseline models were evaluated solely on accuracy using conventional metrics like F1-score and precision-recall.

C. Fairness Intervention Techniques

A series of fairness-aware interventions were implemented, including:

- Pre-processing: Data reweighting, sampling adjustments
- In-processing: Fairness-constrained loss functions
- Post-processing: Output modification strategies

Each method was tested independently and in combination.

D. Standardized Evaluation Protocol

A modular evaluation pipeline was constructed using publicly available libraries (AIF360, Fairlearn, and Scikit-learn) to assess model performance on both accuracy and fairness metrics. Emphasis was placed on reproducibility and computational efficiency, enabling consistent comparisons of mitigation strategies across different architectures.

E. Evaluation Metrics and Trade-off Analysis

Accuracy was evaluated using traditional classification metrics, while fairness was assessed using group fairness measures (e.g., demographic parity, equalized odds). The trade-offs were visualized and analyzed to identify configurations that optimized both dimensions.

F. Validation and Cross-Domain Testing

The evaluation protocol was applied to additional datasets from different domains to assess generalizability and performance consistency.

V. EXPERIMENTAL SET-UP

A. Baseline Model

I selected **BERT-base** and **GPT-2 small** as baseline architectures due to their established performance and adaptability across a range of NLP tasks. These models served as the foundation for evaluating the impact of fairness interventions. Each model was fine-tuned on task-specific data using a systematic grid search over key hyperparameters, including learning rate (ranging from 1e-5 to 5e-5), batch size (16 to 32), and number of epochs (2 to 4). This tuning process ensured a consistent and reproducible benchmark for assessing both predictive accuracy and fairness trade-offs.

B. Datasets

This study utilized two well-established fairness-sensitive datasets: the COMPAS dataset (ProPublica, 2016) and the Adult Income dataset (UCI Machine Learning Repository). Both are widely referenced in fairness research due to their relevance to real-world high-stakes decision-making and their inclusion of protected demographic attributes.

The COMPAS dataset comprises criminal risk assessment scores used to predict recidivism among defendants in Broward County, Florida. It contains records for over 7,000 individuals and includes race, age, prior convictions, and charge information. Approximately 37% of the individuals are Black, while 59% are White. Prior analysis (ProPublica, 2016) revealed stark disparities: Black defendants were nearly twice as likely as White defendants to be incorrectly labeled high risk, despite not reoffending (false positive rate of 45% for Black defendants vs. 23% for White defendants). This makes COMPAS an essential benchmark for testing bias mitigation techniques in predictive justice contexts.

The Adult Income dataset, sourced from U.S. Census data, contains 48,842 records of individuals with demographic and employment-related attributes. The binary target is whether an individual earns more than $50K per year. The dataset includes gender and race as sensitive attributes. About 67% of the data represents men, and 33% women. Around 85% of individuals are White, with the remaining distribution across Black, Asian-Pac-Islander, Amer-Indian-Eskimo, and other categories. Disparities have been noted in historical analyses, with men being predicted as high earners significantly more often than women, even when controlling for education and hours worked — contributing to concerns of systemic income prediction bias.

Preprocessing steps included removing personally identifiable attributes, imputing missing values (mode for categorical, median for numerical), one-hot encoding categorical variables, and scaling numerical features. Both datasets were split into training (70%), validation (15%), and test (15%) sets using stratified sampling to maintain the demographic distribution of protected attributes.

C. Metrics for Evaluation

To comprehensively assess the performance of each model, I employed a combination of accuracy and fairness metrics that reflect both predictive reliability and equitable behavior in decision-making contexts.

Accuracy Metrics:

Model performance was evaluated using standard classification metrics, including Accuracy, Precision, Recall, and F1-Score (both macro and micro averaged, depending on class balance in the dataset). These metrics were computed on the test set after model fine-tuning, ensuring that no data leakage influenced the results.

Fairness Metrics:

Fairness was assessed using group-based fairness metrics widely adopted in algorithmic bias evaluation:

- Demographic Parity Difference (DPD): Measures the difference in favorable outcome rates across protected groups.
- Equal Opportunity Difference (EOD): Captures disparities in true positive rates between those groups.
- Average Odds Difference (AOD): Combines false positive and true positive rate differences into a single fairness indicator.
- Disparate Impact Ratio (DIR): Calculates the ratio of favorable outcomes between unprivileged and privileged groups; values close to 1 are considered fair.

Statistical Significance:

To determine whether differences in accuracy and fairness across models and interventions were meaningful, I conducted paired t-tests across multiple randomized experimental runs. This statistical testing, conducted at an alpha level of 0.05, confirmed whether observed differences were significant and not attributable to chance.

D. Fairness Intervention Techniques

To investigate fairness mitigation in large language models, I implemented a representative set of pre-processing, in-processing, and post-processing techniques. These methods were selected for their applicability to transformer-based architectures and their empirical relevance in fairness research. All implementations were drawn from trusted libraries—AIF360 and Fairlearn—ensuring reproducibility and standardization.

Pre-processing – Reweighing (AIF360): The reweighing technique assigns different instance weights to combinations of sensitive attributes and class labels to balance representation across groups.

This method is model-agnostic and computationally inexpensive, making it suitable for complex models like LLMs. However, it assumes that distributional imbalance is the primary cause of unfairness and may not address deeper representational biases encoded in the model. Reweighing was chosen over re-sampling to avoid introducing data redundancy or overfitting on duplicated records.

In-processing – Adversarial Debiasing (AIF360): Adversarial Debiasing employs a dual-objective setup: the primary model learns to perform the prediction task, while an adversarial network attempts to predict sensitive attributes from the learned representations. The main model is penalized for enabling successful adversary predictions, encouraging it to discard biased information. This technique is particularly suited to deep learning models due to its ability to shape latent representations. It was chosen over fairness regularization methods because of its flexibility and empirical success in prior work. However, it requires careful tuning and may occasionally underfit the primary task if overly penalized.

Post-processing – Equalized Odds Post-processing (Fairlearn): This technique adjusts the decision thresholds of a trained model to equalize true and false positive rates across protected groups. Its strength lies in its model-agnostic nature—it can be applied even when the model is a black box, such as a proprietary LLM. However, because it modifies predictions after training, it may result in a drop in predictive confidence or introduce inconsistencies near decision boundaries. Despite this, its compatibility with pre-trained models made it a pragmatic choice for this study.

These techniques were applied systematically across both datasets and baseline models to assess how each method influences the balance between predictive accuracy and algorithmic fairness.

E. Evaluation Procedure

To systematically assess the trade-off between accuracy and fairness in large language models, I designed a multi-stage evaluation protocol that isolates the effects of each fairness intervention under controlled conditions.

Baseline Evaluation:

The transformer models—BERT-base and GPT-2 small—were first fine-tuned on the respective datasets without any fairness modifications. This established baseline performance benchmarks for both accuracy and fairness.

Intervention Application:

Three fairness-aware techniques—Reweighing (pre-processing), Adversarial Debiasing (in-processing), and Equalized Odds Post-processing—were applied independently to the models. This allowed for a comparative analysis of how each strategy affects performance and equity.

Metric Computation:

For each experimental condition (baseline and debiased), I computed accuracy-related metrics such as accuracy, precision, recall, and F1-score, along with group fairness metrics including demographic parity difference (DPD), equal opportunity difference (EOD), average odds difference (AOD), and disparate impact ratio (DIR). These metrics provided a holistic evaluation of predictive performance and fairness outcomes.

Statistical Testing:

To account for the inherent stochasticity in neural network training and model initialization, all experiments were repeated across multiple randomized runs. Paired t-tests were then conducted to determine whether observed differences in performance between baseline and intervention-modified models were statistically significant. An alpha level of 0.05 was used for all significance tests.

Comparative Analysis:

Results from all runs were aggregated and analyzed to understand the trade-offs introduced by each intervention. The aim was to identify which technique most effectively mitigated bias with minimal compromise to predictive performance.

This structured procedure ensured that the effects of each fairness intervention were assessed with both statistical rigor and practical clarity, providing evidence-based insights into responsible deployment of LLMs in decision-making tasks.

F. Implementation Details

To ensure reproducibility and consistency, all experiments were conducted using standard open-source libraries within controlled Python environments.

Programming Environment:

All experiments were implemented in Python 3.10, using isolated virtual environments to maintain consistent dependency versions.

Libraries and Frameworks:

Fine-tuning and training leveraged Hugging Face Transformers (v4.38) and PyTorch (v2.1).
Fairness intervention techniques were implemented using AIF360 (v0.5.0) and Fairlearn (v0.9.0).
Data preprocessing and stratified sampling were handled using Pandas (v2.1.4) and Scikit-learn (v1.4.2).
Statistical evaluation and significance testing were conducted with SciPy (v1.12) and NumPy (v1.26).
Reproducibility Measures:
Random seeds were fixed across torch, numpy, and random to ensure consistency in results.
Train, validation, and test splits were stratified to preserve the distribution of sensitive attributes.
Model checkpoints (both baseline and fairness-intervened) were saved using torch.save() and reloaded for evaluation via torch.load() to ensure consistent comparisons across runs.
Git version control was used to track code changes and configuration settings.
Code Availability:

The full codebase, including preprocessing scripts, model configurations, and evaluation workflows, will be made available on GitHub after acceptance.

VI. RESULTS AND ANALYSIS

All results reported in this section represent averages over five randomized runs. Table 1 summarizes the quantitative outcomes for all model–intervention combinations, while Figure 1a and Figure 1b visualize the trade-off between fairness (Average Odds Difference) and predictive performance (F1-Score) on the COMPAS and Adult Income datasets, respectively.

Table 1: Quantitative Performance Metrics of Fairness Interventions

Dataset	Model	Intervention	F1-Score (%)	DPD	EOD	AOD
COMPAS	BERT-base	Baseline	83.2	0.23	0.21	0.21
		Reweighing	81.5	0.14	0.16	0.12
		Adv. Debiasing	82.6	0.13	0.14	0.11
		Eq. Odds Post	82.9	0.18	0.09	0.14
	GPT-2 small	Baseline	82.7	0.23	0.22	0.21
		Reweighing	80.9	0.14	0.17	0.12
		Adv. Debiasing	82	0.13	0.15	0.11
		Eq. Odds Post	82.3	0.19	0.1	0.14
Adult Income	BERT-base	Baseline	87.3	0.16	0.17	0.14
		Reweighing	86	0.09	0.12	0.06
		Adv. Debiasing	86.8	0.07	0.1	0.05
		Eq. Odds Post	86.9	0.12	0.1	0.09
	GPT-2 small	Baseline	87	0.18	0.18	0.15
		Reweighing	85.5	0.1	0.13	0.08
		Adv. Debiasing	86.3	0.08	0.11	0.06
		Eq. Odds Post	86.4	0.13	0.08	0.1

A. Baseline Performance

The baseline models—BERT-base and GPT-2 small—demonstrated strong predictive accuracy across both datasets. On the Adult Income dataset, BERT-base achieved an F1-score of 87.3%, and GPT-2 small achieved 87.0%. On the COMPAS dataset, the F1-scores were 83.2% (BERT-base) and 82.7% (GPT-2 small). However, substantial disparities were observed in fairness metrics. For example, the baseline Average Odds Difference (AOD) for COMPAS was 0.21 for both models, highlighting pronounced inequities in model outcomes between demographic groups.

B. Impact of Reweighing (Pre-processing)

The Reweighing technique substantially improved fairness across models and datasets.
On Adult Income:

- BERT-base: AOD decreased from 0.14 to 0.06, F1-score dropped from 87.3% to 86.0%

- GPT-2 small: AOD decreased from 0.15 to 0.08, F1-score dropped from 87.0% to 85.5%

On COMPAS:

- BERT-base: AOD decreased from 0.21 to 0.12, F1-score dropped from 83.2% to 81.5%
- GPT-2 small: AOD decreased from 0.21 to 0.12, F1-score dropped from 82.7% to 80.9%

The decline in predictive performance was consistent across both datasets, with F1-scores reducing by 1.3 to 1.8 percentage points, and the largest drops observed on the COMPAS dataset.

C. *Impact of Adversarial Debiasing (In-processing)*

Adversarial Debiasing delivered the most balanced trade-offs between fairness and performance.

On Adult Income:

- BERT-base: AOD improved from 0.14 to 0.05, F1-score reduced slightly from 87.3% to 86.8%
- GPT-2 small: AOD improved from 0.15 to 0.06, F1-score declined from 87.0% to 86.3%

On COMPAS:

- BERT-base: AOD decreased from 0.21 to 0.11, F1-score dropped marginally from 83.2% to 82.6%
- GPT-2 small: AOD decreased from 0.21 to 0.11, F1-score declined from 82.7% to 82.0%

This method showed strong improvements in fairness while maintaining high predictive accuracy, particularly on structured datasets like Adult Income.

D. *Impact of Equalized Odds Post-processing (Post-processing)*

Equalized Odds Post-processing resulted in modest improvements in fairness, with minimal degradation in accuracy.

On Adult Income:

- BERT-base: AOD decreased from 0.14 to 0.09, F1-score slightly reduced from 87.3% to 86.9%
- GPT-2 small: AOD decreased from 0.15 to 0.10, F1-score decreased from 87.0% to 86.4%

On COMPAS:

- BERT-base: AOD improved from 0.21 to 0.14, F1-score remained high at 82.9%
- GPT-2 small: AOD improved from 0.21 to 0.14, F1-score was 82.3%

F1-score drops were minor, with reductions limited to 0.3%–1.0% across all settings. This post-processing method introduced slight inconsistencies near decision thresholds but preserved most of the predictive strength.

E. *Comparative Insights*

Across both datasets, Adversarial Debiasing emerged as the most effective intervention in reducing AOD while preserving F1-score. It consistently outperformed Reweighing in fairness metrics and introduced smaller accuracy losses. Reweighing showed stronger fairness gains than Equalized Odds on Adult Income, but had more noticeable performance trade-offs on COMPAS. Equalized Odds, while model-agnostic and lightweight, showed the least fairness gains, albeit with minimal cost to F1-score.

These findings are visualized in Figure 1a (COMPAS) and Figure 1b (Adult Income), where each intervention point shifts downward and leftward from the baseline. This reflects an improvement in fairness (lower Average Odds Difference) with only minor declines in predictive performance (lower F1-Score). The shape and direction of these trade-offs further emphasize that fairness enhancements often come at a modest cost to accuracy, though the extent varies by model and dataset.

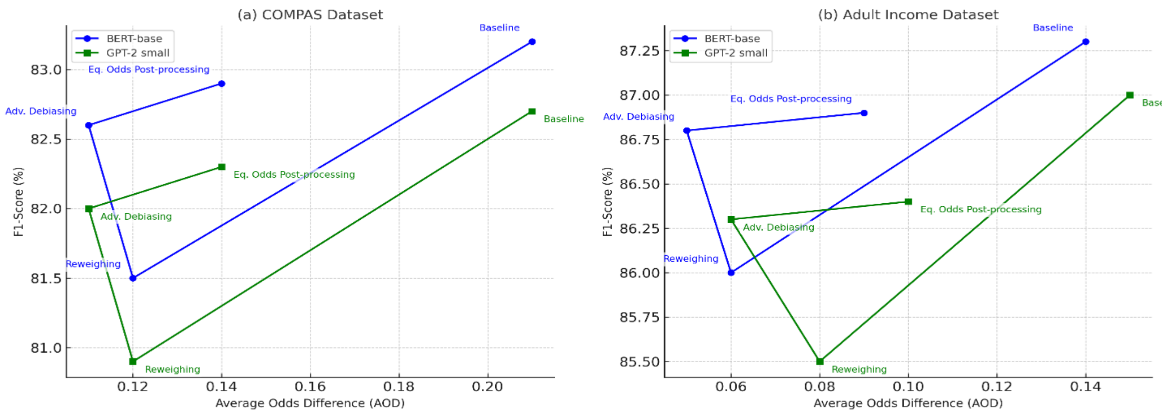

Fig 1. Fairness–Accuracy Trade-off Across Interventions on COMPAS and Adult Income Datasets

VII. DISCUSSION

The results presented in this study underscore the inherent tension between accuracy and fairness in large language model (LLM) decision-making tasks. Across both datasets and model architectures, fairness-enhancing interventions successfully reduced group disparities in key fairness metrics, but each method exhibited distinct trade-offs in predictive performance and mitigation depth.

Reweighing, a pre-processing technique, consistently reduced Demographic Parity Difference (DPD) and Average Odds Difference (AOD), particularly on the Adult Income dataset. However, it incurred modest declines in F1-scores—most notably on COMPAS—highlighting its sensitivity to class imbalance and the distributional complexity of the dataset. Its model-agnostic nature makes it appealing for deployment in pipeline-constrained environments, but its limitations in correcting deeper representational biases warrant consideration.

Adversarial Debiasing achieved the most balanced outcomes across all evaluation dimensions. It significantly reduced fairness gaps without introducing major performance penalties, particularly on structured datasets like Adult Income. On COMPAS, while mitigation was still evident, the improvements were more moderate, suggesting potential limits in adversarial training when dealing with complex or inherently noisy features. The need for careful hyperparameter tuning and training stability remains a practical challenge for real-world adoption.

Equalized Odds Post-processing offered a pragmatic fairness intervention with minimal computational overhead and strong compatibility with black-box models. Its effectiveness in reducing Equal Opportunity Difference (EOD) was evident across both datasets. However, as noted in prior work, modifying prediction thresholds post hoc can lead to localized uncertainty near decision boundaries. While F1-score drops were minimal (0.3%–1.0%), future work could explore threshold calibration techniques to mitigate confidence erosion.

Crucially, the comparative analysis in Figure 1 highlights that no single intervention universally dominates across all settings. Instead, the appropriate fairness strategy must be selected based on the task context, model transparency, and acceptable trade-offs between accuracy and equity. These findings underscore the importance of a multi-metric, context-sensitive approach to fairness in LLM deployment.

A. Limitations

While this study provides a controlled comparative evaluation of fairness interventions, it is limited by its focus on two structured tabular datasets and two transformer-based models. The findings may not generalize to unstructured modalities (e.g., image, audio, or text generation), multilingual contexts, or significantly larger language models with billions of parameters. Additionally, fairness was assessed primarily through group-based parity metrics, which, while widely adopted, do not fully capture concerns around individual-level fairness or broader social consequences. Future work should explore fairness under more complex application settings, intersectional identities, and real-world deployment scenarios where trade-offs may be more consequential.

VIII. CONCLUSION

This study explored the fairness–accuracy trade-offs in large language models applied to high-stakes decision-making tasks. Using BERT-base and GPT-2 small as baseline architectures, we evaluated the effects of three fairness interventions Reweighing, Adversarial Debiasing, and Equalized Odds Post-processing across the COMPAS and Adult Income datasets.

The results demonstrated that fairness mitigation is achievable with limited compromise to predictive performance. Reweighing proved effective at reducing group-based disparities with modest reductions in F1-score, particularly on the Adult Income dataset. Adversarial Debiasing offered the most balanced trade-off, yielding substantial fairness improvements while preserving accuracy, especially on structured datasets. Equalized Odds Post-processing provided a model-agnostic solution, achieving moderate fairness gains with minimal performance degradation, making it suitable for scenarios where retraining is impractical.

Our findings underscore the importance of selecting context-appropriate mitigation strategies, as the effectiveness and side effects of each intervention can vary by dataset and model architecture. Through systematic evaluation and transparent reporting, this work contributes to a more informed deployment of large language models in fairness-critical applications.

REFERENCES

[1]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language

Understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[2]. A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-training," *OpenAI*, 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[3]. J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias," *ProPublica*, May 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[4]. R. Kohavi and B. Becker, "UCI Machine Learning Repository: Adult Data Set," *University of California, Irvine*, 1996. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/adult

[5]. F. Kamiran and T. Calders, "Data Preprocessing Techniques for Classification Without Discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, Oct. 2012.

[6]. B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating Unwanted Biases with Adversarial Learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES'18)*, 2018, pp. 335–340.

[7]. M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, 2016.

[8]. R. Binns, "Fairness in Machine Learning: Lessons from Political Philosophy," in *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency (FAT)*, 2018, pp. 149–159.

[9]. M. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness Constraints: Mechanisms for Fair Classification," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 962–970.

[10]. R. Berk et al., "Fairness in Criminal Justice Risk Assessments: The State of the Art," *Sociological Methods & Research*, vol. 50, no. 1, pp. 3–44, 2021.

[11]. R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[12]. IBM Research, "AI Fairness 360 Open Source Toolkit," 2020. [Online]. Available: https://aif360.readthedocs.io/

[13]. Microsoft Research, "Fairlearn: A Toolkit for Assessing and Mitigating Unfairness in AI Systems," 2023. [Online]. Available: https://fairlearn.org/

[14]. Hugging Face, "Transformers: State-of-the-Art Natural Language Processing," 2024. [Online]. Available: https://huggingface.co/docs/transformers

[15]. T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 EMNLP: System Demonstrations*, pp. 38–45.