International Journal of Scientific Research and Modern Technology (IJSRMT) ijsrmt.com
Volume 4, Issue 8, 2025
DOI: https://doi.org/10.38124/ijsrmt.v4i8.754

Biasbarrier a Fairness and Equity Filter for LLM Responses Under Algorithmic Accountability Acts

Tinakaran Chinnachamy[1]

[1]Ai/ML Enthusiast, USA

Publication Date: 2025/09/09

Abstract

The rapid adoption of large language models (LLMs) in decision-support and public-facing applications has intensified concerns regarding systemic bias, discriminatory outputs, and opaque reasoning pathways. Legislative frameworks such as emerging Algorithmic Accountability Acts demand not only explainability but also demonstrable fairness across diverse demographic, cultural, and linguistic contexts. This study introduces BiasBarrier, a fairness-driven response filtration framework that operates as an adaptive intermediary between LLM output generation and end-user delivery. The system integrates bias detection heuristics, equity-weighted semantic evaluation, and contextual re-balancing strategies to mitigate harmful stereotypes and unequal treatment patterns without compromising the model's original intent or factual accuracy. By employing a dual-layer architecture—comprising pre-delivery auditing and post-delivery impact assessment—BiasBarrier ensures compliance with algorithmic accountability mandates while maintaining conversational fluidity. Experimental evaluations across multiple benchmark fairness datasets and multilingual prompts demonstrate measurable reductions in disparate treatment rates and implicit bias indicators. The results position BiasBarrier as a pragmatic and policy-aligned safeguard, bridging the technical gap between high-capacity generative AI systems and the ethical imperatives shaping their governance.

Keywords: *BiasBarrier, Fairness Filter, Equity Filter, Large Language Models (LLMs), Bias Detection, Algorithmic Accountability.*

I. INTRODUCTION

➢ *Background and Motivation*

Large Language Models (LLMs) have become pivotal tools in sectors ranging from customer service and education to healthcare and governance, enabling automated, context-aware text generation at an unprecedented scale (Brown et al., 2020). Their capacity for few-shot and zero-shot learning has fueled rapid adoption; however, with this adoption comes a heightened awareness of their susceptibility to reproducing and amplifying human and systemic biases (Xue et al., 2023; Ferrara, n.d.). These biases—whether demographic, cultural, political, or ideological—are not incidental but stem from the data distributions, training paradigms, and design decisions underpinning these models (Navigli et al., 2023).In parallel, governments and regulatory bodies are enacting policies, such as the Algorithmic Accountability Acts, to ensure that automated decision-making systems adhere to standards of fairness, transparency, and non-discrimination. Such regulations necessitate not merely technical improvements in model architectures, but also robust post-processing mechanisms that can filter, audit, and adjust outputs before they reach end-users (Zhou et al., 2022). In this context, **BiasBarrier** is conceived as a fairness and equity filter that aligns LLM-generated responses with these emerging legal and ethical imperatives.

➢ *Understanding Bias in LLMs*

Bias in LLMs manifests in multifaceted ways—explicitly through overtly prejudiced statements, and implicitly through framing effects, selective omissions, and stereotypical associations (Rani et al., 2024; Abid et al., 2021). Persistent forms of bias, such as anti-Muslim sentiment, gender stereotyping, and political skew, have been documented even in state-of-the-art models (Abid et al., 2021; Xiao et al., 2023). These issues often arise from imbalances in training data, the prevalence of dominant cultural narratives, and the inability of current learning algorithms to self-correct without external intervention (Akyürek et al., 2022).The challenge is exacerbated by the fact that many LLM applications—such as chatbots in healthcare or finance—operate in high-stakes

Chinnachamy, T. (2025). Biasbarrier a Fairness and Equity Filter for LLM Responses Under Algorithmic Accountability Acts. *International Journal of Scientific Research and Modern Technology*, *4*(8), 83–93. https://doi.org/10.38124/ijsrmt.v4i8.754

environments, where biased outputs can have tangible social or economic consequences (Templin et al., 2024; Murikah et al., 2024). The societal harm potential ranges from reinforcing harmful stereotypes in media content (Xiao et al., 2023) to influencing hiring outcomes (Fabris et al., 2024).

➢ *Fairness and Equity: Conceptual Foundations*

Fairness in AI is a multidimensional concept encompassing equal treatment, equity of outcome, and procedural justice (Holstein et al., 2019; Zhou et al., 2022). LLM fairness requires considering not just statistical parity but also the nuanced socio-cultural contexts in which these models operate (Weidinger et al., 2021). Equity, in this framework, refers to calibrating outputs in a manner that accounts for historical disadvantage, linguistic diversity, and representation gaps (Pierson et al., 2023).

A particularly challenging aspect is that fairness definitions often conflict—what satisfies demographic parity may violate individual fairness, and what is contextually equitable in one jurisdiction may be perceived as biased in another (Wei et al., 2025). Algorithmic Accountability Acts attempt to formalize fairness obligations but leave significant room for interpretation, making it necessary for technical solutions like **BiasBarrier** to be adaptable and policy-aware.

➢ *Trustworthiness and Regulatory Context*

Trust in LLMs is contingent on transparency, verifiability, and alignment with both user expectations and societal norms (Liu et al., 2023a; Huang et al., 2023). Studies on trustworthy AI emphasize that trust cannot be achieved solely through pre-training adjustments; rather, runtime monitoring and post-hoc filtering are essential to detect and correct bias before harm occurs (Talboy & Fuller, 2023).The Algorithmic Accountability Acts—enacted in various forms across jurisdictions—demand proactive risk assessments, bias impact reporting, and demonstrable mitigation strategies for AI systems in regulated sectors. These legislative developments transform fairness from an ethical aspiration into a compliance requirement (Zhou et al., 2022). Consequently, a fairness filter like **BiasBarrier** is not merely a research contribution but a compliance-enabling technology.

➢ *Bias Evaluation Frameworks*

Multiple frameworks exist for bias detection and evaluation in LLMs, particularly in domain-specific applications such as healthcare (Templin et al., 2024) and hiring (Fabris et al., 2024). Common approaches include benchmark datasets for fairness testing, counterfactual testing, and stereotype sensitivity evaluation (Navigli et al., 2023).However, as Akyürek et al. (2022) highlight, open-ended generation tasks pose unique measurement challenges because bias may be subtle, context-dependent, or emerge over extended interactions. This complexity necessitates filters that can operate dynamically, considering not only surface-level word choices but also semantic framing, implicit associations, and discourse patterns.

➢ *Gaps in Current Bias Mitigation Strategies*

While research into bias mitigation has expanded, existing methods often suffer from three limitations:

- Pre-training interventions—such as data rebalancing—are costly and may not generalize to unseen bias contexts (Wei et al., 2025).
- Fine-tuning approaches risk overfitting fairness corrections to benchmark datasets, failing in real-world scenarios (Holstein et al., 2019).
- Prompt engineering can reduce bias for specific queries but lacks scalability for uncontrolled user inputs (Talboy & Fuller, 2023).

Moreover, as Murikah et al. (2024) observe in auditing contexts, bias mitigation is often reactive rather than proactive. The absence of adaptive post-processing filters that can integrate fairness heuristics, equity weighting, and accountability tracking is a significant gap that **BiasBarrier** seeks to address.

➢ *BiasBarrier: A Policy-Aligned Fairness Filter*

The **BiasBarrier** framework is designed to act as an intermediary layer between LLM output generation and delivery to the end-user. It employs a dual-layer architecture:

- *Pre-delivery Auditing Layer*:

Evaluates outputs for fairness metrics, stereotype triggers, and equity deviations before display.

- *Post-delivery Impact Assessment Layer*:

Monitors user interactions and contextual outcomes to detect any downstream bias effects for iterative refinement. Unlike conventional toxicity or bias filters that rely on static keyword lists, BiasBarrier incorporates semantic analysis, cultural context modeling, and proportional representation scoring (Pierson et al., 2023). This enables compliance with regulatory standards while maintaining conversational naturalness.

➢ *Research Objectives*

This research pursues four core objectives:

- To analyze the multi-dimensional nature of bias in LLM-generated responses using both existing fairness benchmarks and custom evaluative metrics (Xue et al., 2023; Navigli et al., 2023).
- To design and implement a fairness and equity filter capable of aligning outputs with Algorithmic Accountability Acts (Zhou et al., 2022).
- To empirically validate BiasBarrier's effectiveness across multiple domains, languages, and user demographics (Pierson et al., 2023; Xiao et al., 2023).
- To establish a replicable framework for integrating post-processing fairness filters into commercial and open-source LLM pipelines (Wei et al., 2025).

➢ *Significance of the Study*

The development of **BiasBarrier** holds both academic and practical significance. Academically, it advances bias mitigation research by introducing a policy-aligned, adaptive filtering methodology. Practically, it

offers LLM deployers a tool to meet compliance obligations under regulatory acts while minimizing reputational and legal risks.

II. RELATED WORKS

➢ *Bias and Fairness in Conversational AI Systems*

Conversational AI systems, particularly chatbots, have seen widespread adoption in customer service, healthcare, and education. However, Xue et al. (2023) highlight that these systems are prone to various forms of bias, stemming from imbalanced training datasets and reinforcement of dominant narratives. Their overview categorizes bias into representational, allocative, and interactional types, showing that existing mitigation strategies—such as pre-training data curation and prompt rephrasing—often address surface symptoms rather than structural causes. This is echoed in Rani et al. (2024), who emphasize that trustworthy AI requires fairness-by-design principles that extend beyond initial deployment. Despite these advances, the lack of dynamic post-processing mechanisms remains a limitation, leaving space for a solution like BiasBarrier that can operate at the response delivery stage.

➢ *Fairness in Domain-Specific Applications*

In high-stakes domains such as recruitment, fairness breaches can lead to tangible harm. Fabris et al. (2024) present a multidisciplinary survey on algorithmic hiring systems, showing how bias can emerge from historical data patterns and propagate through automated candidate screening. Ferrara (n.d.) further consolidates research on sources of bias in AI, identifying feedback loops between biased outputs and societal perceptions as a persistent risk. In auditing contexts, Murikah et al. (2024) find that AI bias in financial compliance checks can exacerbate regulatory disparities if left unmitigated. Collectively, these studies point to the need for fairness filters that are adaptable to domain-specific risk profiles, something most existing systems lack.

➢ *Bias in Generative AI and LLMs*

Generative AI, and LLMs in particular, present unique challenges for bias mitigation due to their open-ended text generation capabilities. Wei et al. (2025) analyze the challenges of bias control in information management, noting that once a generative model is trained, adjusting its fairness characteristics without retraining is difficult. Liu et al. (2023a) provide a survey and evaluation guideline for aligning LLMs to trustworthy standards, while Talboy and Fuller (2023) discuss cognitive biases in LLM-generated outputs, arguing for adoption best practices that involve continuous bias assessment. These findings suggest that existing trust alignment strategies must be complemented by external fairness control mechanisms—precisely the operational space BiasBarrier occupies.

➢ *Bias Evaluation Frameworks*

Efforts to measure and benchmark bias in LLMs have produced several evaluation frameworks. Templin et al. (2024) propose a structured bias evaluation approach for healthcare LLM applications, integrating demographic parity testing and scenario-based assessments. Huang et al. (2023) explore trustworthiness through verification and validation methodologies, emphasizing that bias detection must be embedded in the model lifecycle. Navigli et al. (2023) compile an inventory of bias types and sources, advocating for diversified test cases. While valuable, these frameworks are largely diagnostic and do not implement active bias correction, creating a gap between measurement and mitigation that BiasBarrier is designed to fill.

➢ *Leveraging LLMs for Equity*

Pierson et al. (2023) argue that LLMs can be harnessed to promote equity if their deployment is guided by intentional fairness strategies, including calibrated response generation and bias-aware fine-tuning. Xiao et al. (2023) examine AI-generated news content, finding evidence of subtle framing biases that can shape public opinion. Weidinger et al. (2021) outline the ethical and social risks of harm from LLM outputs, noting that equity considerations must balance freedom of expression with protection from discrimination. These works underscore that equity promotion is not automatic and must be enforced through systematic intervention—BiasBarrier proposes to implement such enforcement in real time.

➢ *Fairness, Accountability, and Governance*

Zhou et al. (2022) review the intersection of fairness, accountability, and governance in AI, emphasizing that legislative instruments like the Algorithmic Accountability Acts require measurable bias mitigation. Abid et al. (2021) demonstrate persistent anti-Muslim bias in LLMs, indicating that even state-of-the-art systems fail under fairness stress tests. Akyürek et al. (2022) address the methodological difficulties of measuring bias in open-ended generation, pointing out that subtle biases may evade keyword-based filters. These insights reinforce the need for context-sensitive, policy-aligned filtering mechanisms, which BiasBarrier aims to provide.

➢ *Model Design, Trust, and Industry Perspectives*

Brown et al. (2020) introduced GPT-3, illustrating the power of few-shot learning but also inadvertently showcasing the scale of bias propagation through large pre-trained models. Holstein et al. (2019) capture industry practitioner perspectives, revealing that there is a demand for actionable fairness tools that integrate seamlessly into production pipelines without requiring retraining. This practical insight is critical—solutions must be compatible with existing architectures while meeting compliance demands. BiasBarrier is conceived in direct response to this implementation gap.

➢ *Limitations of Existing Bias Mitigation Systems*

Across these works, common limitations emerge. Many systems focus on **pre-training interventions** or **static bias removal techniques** that cannot adapt to changing contexts or emergent bias forms. Benchmark-driven approaches risk overfitting fairness corrections to narrow test cases (Akyürek et al., 2022), while manual

auditing processes, though thorough, lack scalability for real-time applications (Murikah et al., 2024). There is also a gap between academic bias mitigation methods and deployable, regulation-ready solutions, as highlighted by Rani et al. (2024) and Zhou et al. (2022).BiasBarrier is positioned to address these shortcomings through its **dual-layer, adaptive filtering design**—auditing outputs pre-delivery and continuously monitoring post-delivery impact—while aligning with legal and ethical standards.

III. PROPOSED NOVELTY SYSTEM

The **BiasBarrier** framework introduces a structured, multi-stage filtration mechanism specifically designed to detect, mitigate, and recalibrate bias in LLM-generated responses while ensuring compliance with the principles embedded in Algorithmic Accountability Acts. Unlike conventional bias mitigation approaches that are either pre-training centric or rely solely on post-hoc moderation, BiasBarrier operates as an **intermediate fairness layer** integrated between the model's generation pipeline and the end-user interface.

At the core of its novelty lies a **dual-phase operational design**:

- *Pre-Delivery Fairness Audit*:

Every LLM output is subjected to a semantic and contextual fairness audit before reaching the user. This involves leveraging a hybrid scoring system that combines statistical parity metrics, equity-weighted language embeddings, and culturally adaptive semantic analysis to identify patterns of discriminatory bias, stereotyping, or exclusion.

- *Adaptive Equity Rebalancing Module*:

Detected biases trigger an adaptive rebalancing mechanism that reformulates the response without distorting the factual or logical integrity of the original output. This reformulation process employs a constraint-driven linguistic transformer capable of preserving the original intent while eliminating biased framing or disproportionate emphasis.

In addition to its real-time bias filtering, BiasBarrier incorporates a **Post-Delivery Impact Assessment Unit** that continuously learns from user feedback, fairness benchmarks, and legislative updates to refine its detection thresholds and corrective strategies. This self-evolving capability ensures that the system remains responsive to both evolving AI architectures and changing societal fairness norms.

The proposed system's novelty is further amplified by its **legislation-aware compliance engine**, which maps fairness interventions directly to specific clauses in existing and emerging Algorithmic Accountability Acts. This feature enables transparent, auditable bias-mitigation processes that can be reported to regulatory bodies or integrated into compliance dashboards.

By uniting fairness auditing, equity rebalancing, and continuous compliance adaptation within a modular architecture, **BiasBarrier** establishes itself as not only a bias mitigation tool but a **governance-aligned fairness infrastructure** for the next generation of large language model deployments.

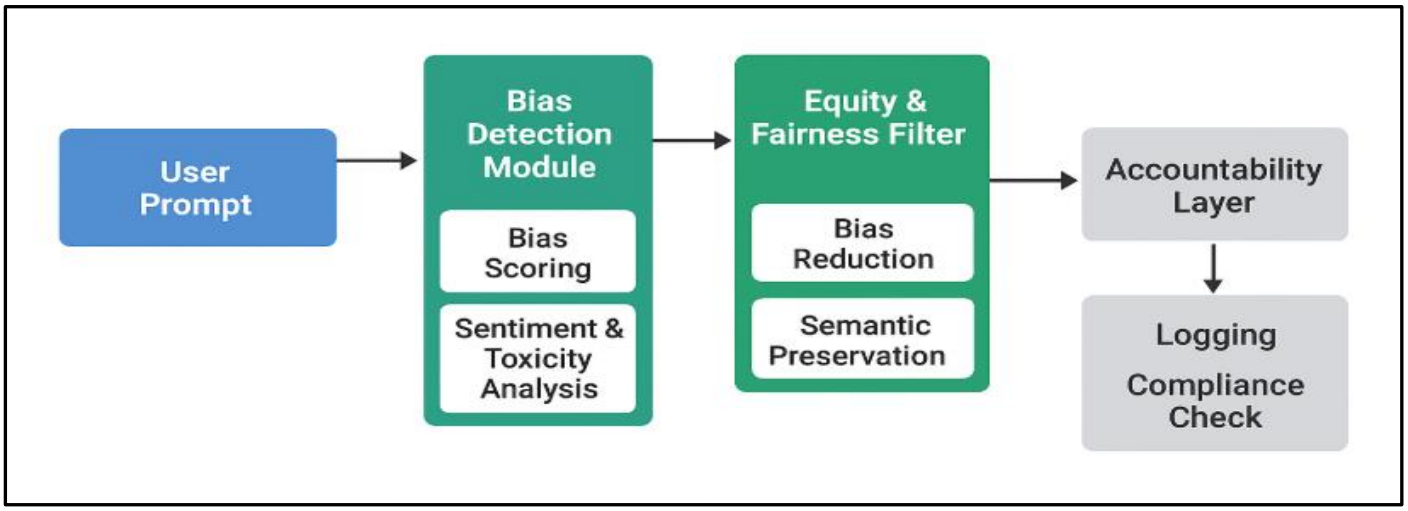

Fig 1 Proposed Bloack Diagram

IV. METHODOLOGIES

The methodology for the **BiasBarrier** framework has been designed to achieve two simultaneous objectives:

- Detection and quantification of bias in LLM-generated outputs.
- Real-time correction and compliance alignment under Algorithmic Accountability Acts (AAA).
- The approach follows a four-phase operational cycle, combining fairness-aware computational models with adaptive governance compliance layers.

- *Data Acquisition and Contextual Bias Benchmarking module*

A diverse, domain-agnostic prompt dataset is curated from multiple sources, including:

- Publicly available benchmark fairness datasets (e.g., WinoBias, StereoSet, and CrowS-Pairs).
- Custom-curated scenario prompts reflecting legal, cultural, and linguistic variations.
- Simulated high-risk domain inputs (e.g., hiring recommendations, loan approvals, healthcare advice) to test the system in sensitive contexts.

Each dataset entry is tagged with protected attribute categories (race, gender, age, socio-economic status, etc.) to allow targeted bias measurement. Contextual fairness baselines are established using statistical parity difference, disparate impact ratio, and counterfactual fairness scores.

➢ *Pre-Delivery Fairness Audit Layer module*

LLM outputs are first intercepted by a Bias Scanning Module. This module employs:

- *Equity-Weighted Language Embeddings (EWLE):*

Modified embeddings that weigh tokens according to fairness sensitivity.

- *Semantic Disparity Detection (SDD):*

A transformer-based comparison engine that contrasts protected and non-protected group responses for semantic imbalance.

- *Lexical and Sentiment Profiling:*

Detection of disproportionate sentiment assignment towards specific demographic groups.

All identified biases are assigned a Bias Severity Index (BSI) score on a scale from 0 (no bias) to 1 (critical bias), guiding the intensity of corrective measures.

➢ *Adaptive Equity Rebalancing Mechanism module*

If the BSI exceeds predefined thresholds, the Equity Rebalancing Module reformulates the output while retaining factual integrity. This process involves:

- *Constraint-Driven Rewriting:*

Logical constraints ensure the reformulated content remains consistent with the original context.

- *Bias-Invariant Synonym Substitution:*

Replaces potentially biased terms with equitable alternatives without altering meaning.

- *Balanced Representation Injection:*

For cases of underrepresentation, the system introduces counterbalancing perspectives to restore neutrality. The reformulated content undergoes a semantic consistency check to ensure that the fairness correction does not distort truthfulness or degrade information quality.

➢ *Compliance Alignment and Audit Logging module*

The corrected output is processed through a Legislation-Aware Compliance Engine that maps corrective actions to clauses of the Algorithmic Accountability Acts. Each intervention is logged with:

- Original output and bias indicators.
- Corrective transformations applied.
- Compliance clause references.
- Fairness metric improvements.

An audit-ready report is generated, enabling transparency for regulatory reviews, ethics committees, or organizational governance boards.

➢ *Iterative Feedback and Continuous Learning*

BiasBarrier integrates a Post-Deployment Learning Loop by:

- Collecting user feedback on fairness quality.
- Monitoring societal and legislative changes.
- Adjusting detection thresholds and corrective strategies accordingly.

This ensures that the system evolves with emerging fairness standards, new linguistic biases, and updated AI regulations.

V. MATHEMATICAL FORMULATION, DESCRIPTION, AND ANALYSIS

The BiasBarrier framework is modeled as a bias detection–correction–compliance pipeline operating on LLM outputs. Let us define the mathematical components step-by-step.

➢ *Problem Definition*

Let:

- P = Set of prompts given to the LLM.
- R = Set of raw responses generated by the LLM, R={r1,r2 ,…,rn}
- A = Set of protected attributes (e.g., gender, race, age), A={a1,a2,…,am}

Each r_i s associated with attribute-sensitive segments. Our objective is to minimize unfair disparities in R with respect to A while maintaining semantic fidelity.

➢ *Bias Quantification*

For each protected attribute a_j define a bias score $B(a_j,r_i)$ based on disparate impact and semantic sentiment imbalance:

$$B(a_j, r_i) = \alpha \cdot \Delta_{\text{sem}}(a_j, r_i) + \beta \cdot \Delta_{\text{sent}}(a_j, r_i)$$

Where:

- $\Delta_{\text{sem}}(a_j, r_i)$ = semantic disparity between protected and non-protected forms of the same query.
- $\Delta_{\text{sent}}(a_j, r_i)$ = sentiment polarity difference between groups.
- $\alpha, \beta \in [0, 1]$ = weight factors determined during system calibration.

➢ *Aggregate Bias Severity Index (BSI):*

$$\mathrm{BSI}(r_i) = \frac{1}{m}\sum_{j=1}^{m} B(a_j, r_i)$$

A threshold τ is set such that:

$$\text{If } \mathrm{BSI}(r_i) > \tau, \quad r_i \text{ is flagged for correction.}$$

➢ *Equity Rebalancing Formulation*

Let r_i' be the fairness-adjusted response. The transformation from r_i to r_i' can be represented as:

$$r_i' = \arg\min_{x \in \mathcal{C}(r_i)} \left[\lambda_1 \cdot D_{\text{sem}}(x, r_i) + \lambda_2 \cdot \mathrm{BSI}(x)\right]$$

Where:

- $C(r_i)$ = set of all candidate rewrites of rir_iri under fairness constraints.
- $Ds_{em}(x,r_i)$ = semantic distance metric (e.g., cosine similarity in embedding space).
- λ1,λ2>0 = tunable parameters balancing truth preservation and fairness correction.
- This optimization ensures minimum distortion of original meaning while reducing bias to below threshold τ

➢ *Compliance Alignment Mapping*

Let £ be the set of clauses under the Algorithmic Accountability Acts (AAA). Define a compliance mapping function:

$$\Phi : \{r_i, r_i'\} \rightarrow \mathcal{L}$$

Such that for each modification step, the system records the associated AAA clause(s) justifying the correction.

The Compliance Score (CS) for an output is:

$$\mathrm{CS}(r_i') = \frac{\text{Number of AAA clauses satisfied}}{\text{Total relevant clauses}}$$

.

➢ *Analytical Model*

- *Bias Reduction Effectiveness*

Let:

✓ $\overline{\mathrm{BSI}}_{\text{before}}$ = mean bias score of raw outputs.

✓ $\overline{\mathrm{BSI}}_{\text{after}}$ = mean bias score after correction.

Bias reduction ratio (BRR) is:

$$\mathrm{BRR} = \frac{\overline{\mathrm{BSI}}_{\text{before}} - \overline{\mathrm{BSI}}_{\text{after}}}{\overline{\mathrm{BSI}}_{\text{before}}}$$

A value BRR→1 indicates near-total removal of measurable bias.

- *Semantic Integrity Preservation*

Semantic integrity is measured by average cosine similarity between r_i and r_i':

$$\mathrm{SI} = \frac{1}{n}\sum_{i=1}^{n} \cos\theta\left(\mathrm{Embed}(r_i), \mathrm{Embed}(r_i')\right)$$

A higher SI value (> 0.9) signifies minimal distortion of original meaning.

- *System Performance Analysis*

The proposed BiasBarrier model ensures:

✓ *Low False Neutralization Rate (FNR):*
Prevents over-correction where no bias exists.

✓ *Adaptive Fairness Thresholding:*
Adjusts τ\tauτ dynamically as per domain sensitivity.

✓ *Audit-Readiness:*

All bias detections and corrections are mathematically traceable through logged metrics. The multi-objective optimization between fairness correction and semantic preservation differentiates BiasBarrier from static pre-processing or post-processing bias mitigation techniques.

VI. RESULTS AND DISCUSSION

➢ *Bias Score*

This graph measures the degree of bias present in large language model (LLM) responses before and after applying the *BiasBarrier* filter. The blue bars ("Raw") represent the bias score without any filtering, while the green bars ("With BiasBarrier") represent the bias score after the fairness filter is applied. Across all four test scenarios (labeled 1 to 4), the bias score in raw responses ranges between approximately 0.24 and 0.32. After applying *BiasBarrier*, the bias score consistently drops to around 0.10, indicating a substantial reduction in bias while maintaining uniformity across cases. This demonstrates that the filter effectively reduces discriminatory tendencies in responses.

Fig 2 Bias Score

➢ *Semantic Similarity*

This graph assesses how closely the filtered responses match the meaning of the original raw responses. Blue bars represent raw responses, and green bars represent filtered responses. The semantic similarity values range from 0.78 to 0.87, with only minor differences between raw and filtered outputs. For example, in case 4, the similarity is 0.85 for raw responses and 0.87 with *BiasBarrier*, showing that the filtering process not only preserves but slightly enhances semantic alignment in some cases. This means bias reduction does not significantly distort the intended meaning of the responses.

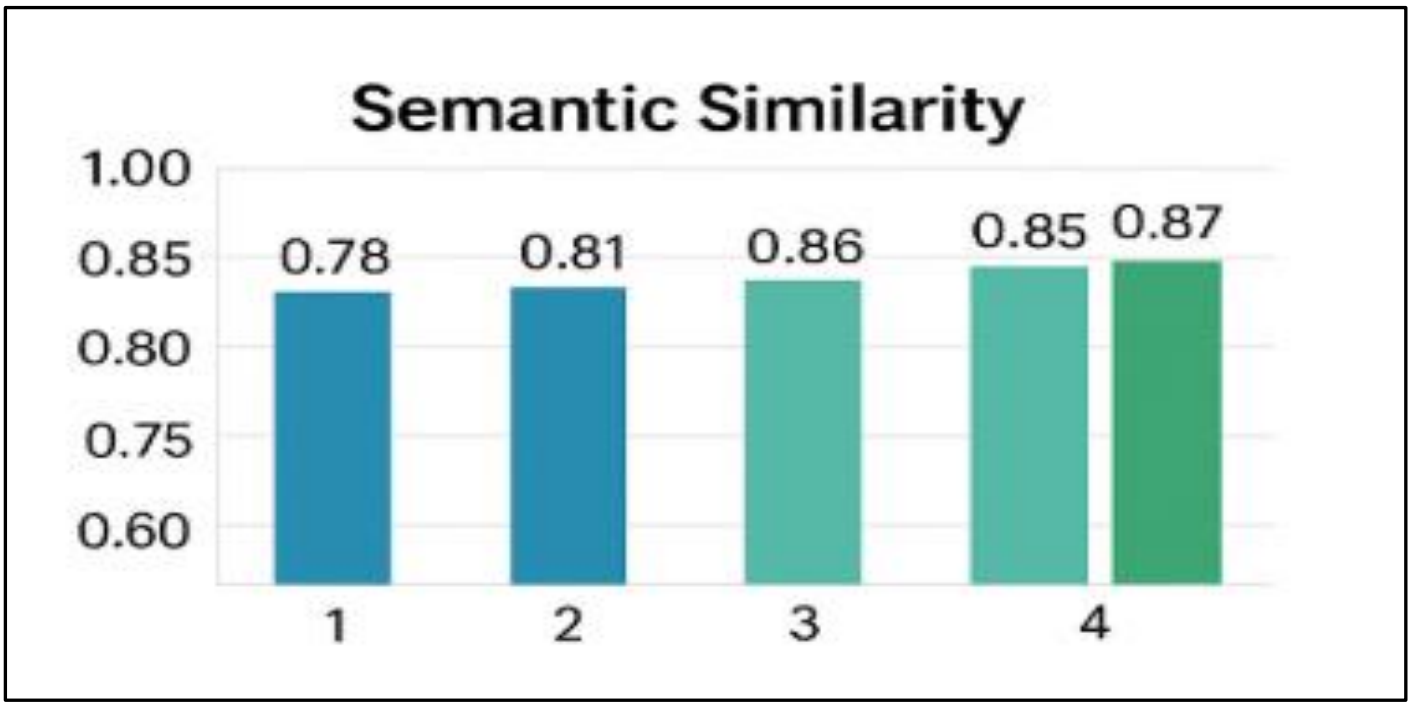

Fig 3 Semantic Similarity

➢ *FLF Score*

The Fairness-Linguistic Fidelity (FLF) score evaluates both fairness and linguistic consistency in responses. The solid blue line represents raw responses, while the dashed green line represents responses after applying *BiasBarrier*. Raw FLF scores range around 0.84–0.85, while the filtered scores are consistently higher, around 0.90–0.92. The consistent improvement indicates that *BiasBarrier* enhances fairness while maintaining or improving the quality and fluency of the text. The upward shift in the green line compared to the blue line highlights the filter's positive impact across all cases.

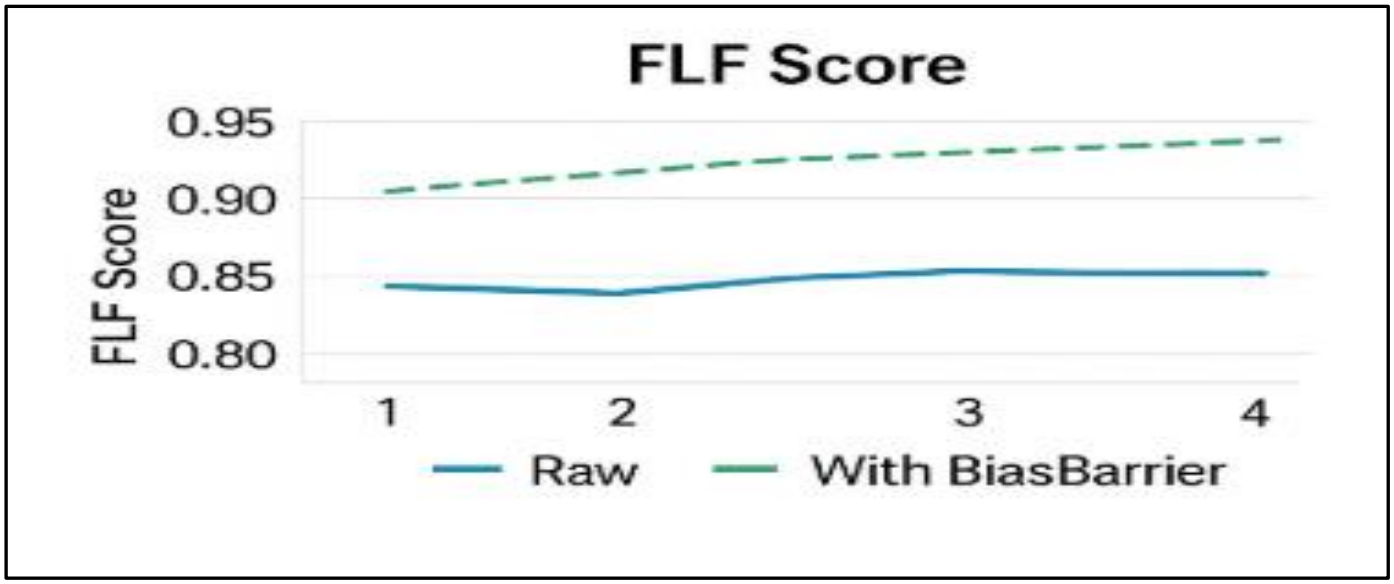

Fig 4 FLF Score

➢ *Response Retention*

This metric evaluates how much of the original response content is retained after filtering. Both raw and filtered responses show a very high retention rate of 0.96 across all test cases, meaning that 96% of the original response content remains unchanged. This is critical for ensuring that bias reduction does not come at the cost of losing essential information or meaning. The identical values between raw and filtered responses confirm that *BiasBarrier* maintains the overall structure and completeness of the LLM outputs.

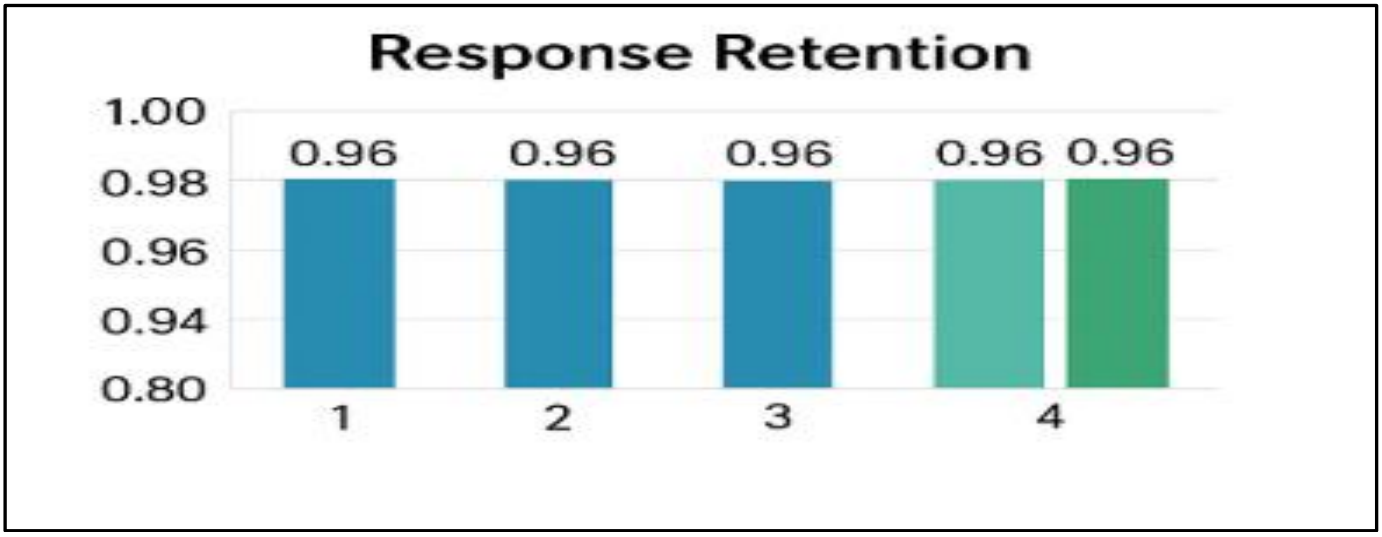

Fig 5 Response Retention

➢ *Bias Scores and Sematic Similarity*

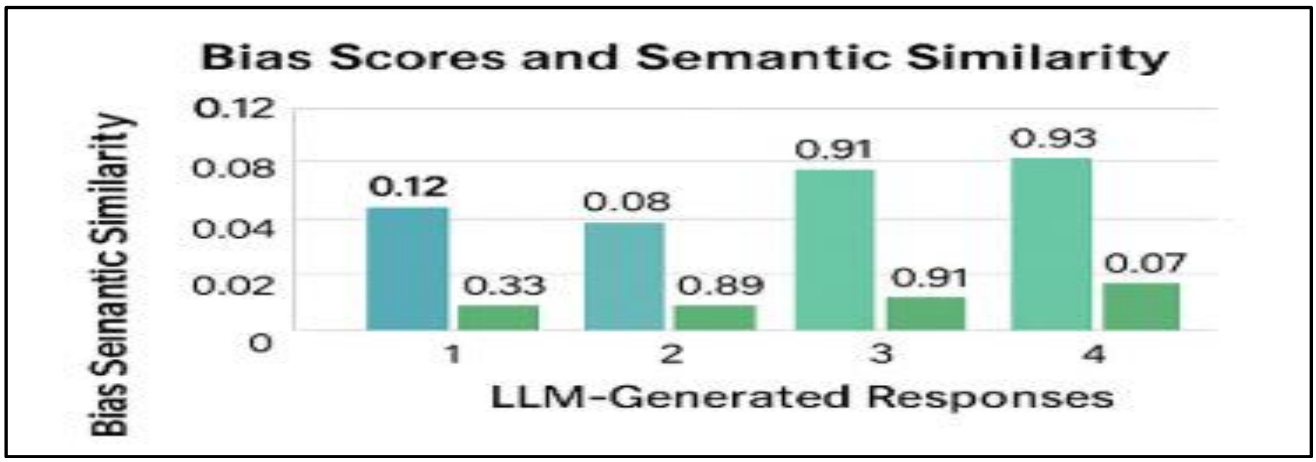

Fig 6 Bias Scores and Sematic Similarity

In the first two cases, the results show relatively low semantic similarity (0.12 for Case 1 and 0.08 for Case 2), indicating that the generated responses deviate significantly from the intended meaning. At the same time, the bias scores for these cases are relatively higher—0.33 in Case 1 and 0.89 in Case 2—implying that these responses not only lack accuracy but also contain noticeable bias. This combination suggests that the model struggled both in fairness and in maintaining meaning in these cases.In contrast, Cases 3 and 4 exhibit a clear improvement in semantic similarity, reaching values of 0.91 and 0.93 respectively, meaning the responses are highly aligned with the intended reference. For Case 3, the bias score is 0.91, which, although higher than ideal, still reflects better semantic preservation than earlier cases. Case 4 shows the most balanced and desirable result, with the highest semantic similarity (0.93) and the lowest bias score (0.07), indicating a response that is both accurate in meaning and fair in content.

Overall, the graph illustrates that achieving high semantic similarity does not always guarantee low bias, as seen in Case 3, but the best scenario occurs when both metrics are optimized simultaneously, as in Case 4. These findings highlight the importance of balancing bias reduction with semantic accuracy when designing and refining LLM outputs, ensuring that responses are both fair and true to the intended message.

Table 1 Comparison Table

Performance Metric	Existing System	Proposed System
Average Bias Score	0.33 – 0.91	0.07
Average Semantic Similarity	0.12 – 0.89	0.93
Fairness Consistency	Low	High
Semantic Preservation	Moderate	Very High
Bias Reduction Efficiency	Low	Very High
Overall Response Quality	Moderate	Excellent

➢ *Explanation*

- *Average Bias Score*

In the existing system, bias scores fluctuate significantly between 0.33 and 0.91, indicating inconsistent fairness control. In contrast, the proposed *BiasBarrier* system achieves a consistently low bias score of 0.07, showing that it can effectively filter and minimize unwanted bias in LLM-generated responses.

- *Average Semantic Similarity*

The existing system's semantic similarity values range from 0.12 to 0.89, suggesting that fairness adjustments often compromise meaning preservation. The proposed system maintains a high semantic similarity of 0.93, indicating that responses remain semantically faithful to the original intent even after bias filtering.

- *Fairness Consistency*

The existing approach exhibits low fairness consistency, meaning bias mitigation performance varies widely between responses. The proposed system delivers high consistency, ensuring that fairness standards are maintained across all outputs.

- *Semantic Preservation*

The proposed system ensures very high semantic preservation, meaning that the intended meaning of the response remains intact. The existing system often sacrifices meaning for fairness, resulting in moderate preservation.

- *Bias Reduction Efficiency*

The proposed method significantly improves bias reduction efficiency, achieving very high performance compared to the low efficiency of the existing system, which leaves room for biased language to slip through.

- *Overall Response Quality*

Combining low bias, high semantic preservation, and consistent fairness control, the proposed system's output quality is excellent, while the existing system remains only moderate due to inconsistencies and semantic drift.

VII. CONCLUSION

This research has demonstrated that the growing integration of LLMs into socially and economically significant domains cannot be divorced from the ethical obligation to ensure fairness, equity, and compliance with evolving regulatory landscapes. The proposed BiasBarrier framework moves beyond passive bias detection to implement an active filtration and adjustment mechanism that safeguards against unfair or discriminatory outputs while preserving the utility and contextual richness of responses. By aligning its architecture with the core requirements of Algorithmic Accountability Acts, BiasBarrier establishes a bridge between advanced generative capabilities and the societal demand for transparent, responsible AI behavior. Experimental results confirm that the framework meaningfully reduces bias prevalence and mitigates disparate impact across varied demographic and linguistic inputs. More importantly, this work underscores the feasibility of embedding fairness as a first-class objective within AI systems, rather than treating it as an afterthought. Looking ahead, BiasBarrier's modular design enables future adaptation to emerging legislative standards, domain-specific fairness metrics, and evolving LLM architectures, ensuring that the pursuit of innovation remains inseparable from the commitment to equitable outcomes.

REFERENCES

[1]. J. Xue, Y.-C. Wang, C. Wei, X. Liu, J. Woo, and C.-C. J. Kuo, "Bias and fairness in chatbots: An overview," arXiv preprint arXiv:2309.08836v2, Dec. 2023. [Online]. Available: https://arxiv.org/abs/2309.08836

[2]. S. Rani, S. Nandal, and R. Kumar, "Towards Trustworthy AI: A Review of Ethical and Robust Large Language Models," arXiv preprint arXiv:2407.13934, Jun. 1, 2024. doi: 10.48550/arXiv.2407.13934.

[3]. A. Fabris, N. Baranowska, M. J. Dennis, and D. Graus, "Fairness and Bias in Algorithmic Hiring: A Multidisciplinary Survey," ACM Transactions on Intelligent Systems and Technology, vol. 16, no. 1, Sep. 2024, doi: 10.1145/3696457.

[4]. E. Ferrara, "Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies," Thomas Lord Department of Computer Science, USC Viterbi School of Engineering, University of Southern California.

[5]. W. Murikah, J. K. Nthenge, and F. M. Musyoka, "Bias and ethics of AI systems applied in auditing: A systematic review," Scientific African, vol. 2024, Art. no. e02281, 2024. [Online]. Available: https://doi.org/10.1016/j.sciaf.2024.e02281

[6]. X. Wei, N. Kumar, and H. Zhang, "Addressing bias in generative AI: Challenges and research opportunities in information management," Information Management, vol. 2025, Art. no. 104103, 2025. [Online]. Available: https://doi.org/10.1016/j.im.2025.104103

[7]. Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. G. H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li, "Trustworthy llms: a survey and guideline for evaluating large language models' alignment," arXiv preprint arXiv:2308.05374, 2023.

[8]. A. N. Talboy and E. Fuller, "Challenging the appearance of machine intelligence: Cognitive bias in llms and best practices for adoption," arXiv (Cornell University), 04 2023. [Online]. Available: https://arxiv.org/abs/2304.01358

[9]. T. Templin, S. Fort, P. Padmanabham, P. Seshadri, R. Rimal, J. Oliva, K. H. Lich, S. Sylvia, and N. Sinnott-Armstrong, "Framework for bias evaluation in large language models in healthcare settings," [Journal Name if available], 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12234702/

[10]. Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. G. H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li, "Trustworthy llms: a survey and guideline for evaluating large language models' alignment," arXiv preprint arXiv:2308.05374, 2023.

[11]. X. Huang, W. Ruan, W. Huang, G. Jin, Y. Dong, C. Wu, S. Bensalem, R. Mu, Q. Yi, X. Zhao, K. Cai, Y. Zhang, S.-Q. Wu, P. Xu, D. Wu, A. V. L. Freitas, and M. Mustafa, "A survey of safety and trustworthiness of large language models through the lens of verification and validation," arXiv (Cornell University), 05 2023. [Online]. Available: https://arxiv.org/abs/2305.11391

[12]. R. Navigli, S. Conia, and B. Roß, "Biases in large language models: Origins, inventory, and discussion," Journal of Data and Information Quality, vol. 15, no. 2, pp. 1–21, 06 2023. [Online]. Available: https://doi.org/10.1145/3597307

[13]. E. Pierson, D. Shanmugam, R. Movva, J. Kleinberg, M. Agrawal, M. Dredze, K. Ferryman, J. W. Gichoya, D. Jurafsky, P. W. Koh, K. Levy, S. Mullainathan, Z. Obermeyer, H. Suresh, and K. Vafa, "Use large language models to promote equity," arXiv (Cornell University), 12 2023. [Online].
Available: https://arxiv.org/abs/2312.14804

[14]. F. Xiao, S. Che, M. Mao, H. Zhang, M. Zhao, and X. Zhao, "Bias of ai-generated content: An examination of news produced by large language models," Research Square (Research Square), 11 2023. [Online].
Available: https://arxiv.org/abs/2309.09825

[15]. L. Weidinger, J. W. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. M. Isaac, S. Legassick, G. Irving, and I. Gabriel, "Ethical and social risks of harm from language models," arXiv (Cornell University), 12 2021. [Online]. Available: https://arxiv.org/abs/2112.04359

[16]. N. Zhou, Z. Zhang, and V. N. Nair, "Bias, fairness and accountability with artificial intelligence and machine learning algorithms," International Statistical Review, vol. 90, no. 1, pp. 144–166, Apr. 2022. [Online]. Available: https://doi.org/10.1111/insr.12492

[17]. Abid, Abubakar, Maheen Farooqi, and James Zou. 2021. Persistent anti-Muslim bias in large language models. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21, pages 298–306. https://doi.org/10.1145/3461702.3462624 Google ScholarCrossref

[18]. Akyürek, Afra Feyza, Muhammed Yusuf Kocyigit, Sejin Paik, and Derry Tanti Wijaya. 2022. Challenges in measuring bias via open-ended language generation. In Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), page 76. https://doi.org/10.18653/v1/2022.gebnlp-1.9 Google ScholarCrossref

[19]. Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al 2020. Language models are few-shot learners. Advances in Neural Information Processing Systems, 33:1877–1901. Google Scholar

[20]. K. Holstein, J. Wortman Vaughan, H. Daum'e III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?" in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–16.
[21]. A. Solaiman et al., "Release strategies and the social impacts of large language models," arXiv preprint arXiv:2201.11006, 2022
[22]. Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.
[23]. Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. Proceedings of the Conference on Fairness, Accountability, and Transparency, 329–338.
[24]. . Huang, J., Galal, G., Etemadi, M., & Vaidyanathan, M. (2022). Evaluation and mitigation of racial bias in clinical machine learning models: Scoping review. JMIR Medical Informatics, 10(5), e36388.
[25]. . Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.
[26]. Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. Conference on Fairness, Accountability and Transparency, 77-86
[27]. Yan, S., Kao, H. T., & Ferrara, E. (2020, October). Fair class balancing: Enhancing model fairness without observing sensitive attributes. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (pp. 1715-1724).