Fairness and Transparency in Large Language Models

Gaurab Pokharel, Samuel Franklin
Department of Computer Science
George Mason University
Fairfax, VA - 22030
{gpokhare, sfrank22}@gmu.edu

***Abstract*—The pervasive use of Large Language Models across various sectors necessitates a rigorous examination of their fairness and transparency. This paper presents a comprehensive survey of the challenges and methodologies associated with addressing biases in LLMs. We begin by defining fairness within the context of LLMs and illustrate how these models can perpetuate existing societal biases, with specific examples demonstrating the real-world implications. Then, through a detailed taxonomy, we identify the critical areas where biases may occur—namely, the data, algorithms, and models used in LLMs—and discuss targeted strategies for each phase of the model lifecycle, including pre-processing, in-processing, and post-processing. We also review the literature according to this framework, shedding light on the intertwined nature of bias in LLMs and the effectiveness of various mitigation strategies. We end by exploring unresolved issues and barriers that complicate the path towards equitable LLMs. It underscores the importance of collaborative efforts spanning multiple disciplines and the inclusion of diverse global perspectives in model development processes. Our survey advocates for a foundational integration of fairness in LLM development and stresses the need for stronger regulatory frameworks to ensure these models are both transparent and inclusive. By mapping out the current landscape and proposing directions for future research, this work aims to contribute to the ethical advancement of AI technologies, promoting models that are not only effective but also just and equitable.**

***Index Terms*—Fairness, AI, Deep Learning, LLM**

I. Introduction

What does *fair* truly mean? As Machine Learning (ML) models, especially Large Language Models (LLMs), become increasingly ubiquitous, ensuring they embody fairness is essential. The rise of LLMs, built on Transformer architectures [1] and trained on massive web corpora [2], has marked a new era not only in technological advancement but in societal impact . These models demonstrate strong performance across a range of tasks such as machine translation, question answering, and dialogue, enabled by their few-shot and zero-shot learning abilities through pre-training, fine-tuning, or prompting [3].

However, the rapid deployment of LLMs has also brought to light their potential to inherit and amplify social biases present in their training data, perpetuating harms against marginalized groups ([4], [5]). Instances where LLMs like GPT-3 associate Muslims with violence [6], African-American English with toxicity [7], and women with lower occupational competence [8] highlight the critical need for fairness, transparency, and accountability in AI systems. Efforts to identify, quantify, and mitigate these biases have produced a range of techniques and evaluation metrics, yet the challenge of standardizing these efforts remains formidable.

In addressing these concerns, it is critical to systematically identify and rectify potential failures throughout a model's lifecycle. This includes pre-processing strategies to ensure bias-free, representative training data; in-processing adjustments during model training to integrate fairness directly into algorithms; and post-processing strategies to adjust outputs for fairness and clarity. By applying these strategies across pre-processing, in-processing, and post-processing stages, we ensure a comprehensive approach to mitigate biases from the data collection phase to the final model output, enhancing the fairness and transparency of LLMs throughout their operational lifecycle.

This survey aims to consolidate and categorize research on bias and fairness in LLMs, covering expanded definitions of social bias, a taxonomy of bias evaluation metrics, and techniques for bias mitigation. By unifying the literature in bias measurement and reduction for LLMs, we facilitate critical analysis and development of more equitable language technologies, focusing on English LLMs but expect the findings to generalize to models based on other languages as well. Our work contributes to the dialogue on how AI can be guided by ethical principles to ensure societal well-being, equity, and justice.

The ensuing sections will delve deeper into each problem area, exploring the challenges, methodologies, and open problems therein, and providing a comprehensive overview of the current state of the art in the field. Section II provides the necessary background for readers to understand the core concepts discussed in the literature. Section III presents taxonomies to organize the discussions, and Section IV surveys existing literature based on these taxonomies. Finally, Section V highlights potential open problems, with Section VI offering concluding remarks. Through this exploration, we aim to shed light on the complexities, challenges, and ethical dilemmas presented by ML models and LLMs in societal contexts. This survey seeks to pave the way toward a future where ML models and LLMs, guided by principles of fairness and transparency, serve as enablers of societal well-being, equity, and justice, ensuring that the benefits of these technologies are

accessible, equitable, and justifiable.

Some relevant conferences and journals where research in this area is generally published are AAAI Conference on Artificial Intelligence (AAAI), ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT), Neural Information Processing Systems (NeurIPS), AAAI/ACM Conference on AI, Ethics, and Society (AIES), ACM Conference on Economics and Computation (EC), International Conference on Autonomous Agents and Multiagent Systems (AAMAS), International Joint Conference on Artificial Intelligence (IJCAI), International Conference on Machine Learning (ICML).

II. Preliminaries

Let M be an LLM parameterized by θ that takes a text sequence $X = (x_1, ..., x_m) \in X$ as input and produces an output $\hat{Y} \in \hat{Y}$, where $\hat{Y} = M(X; \theta)$. Inputs may be drawn from a labeled dataset $D = \{(X^{(1)}, Y^{(1)}), ..., (X^{(N)}, Y^{(N)})\}$, or an unlabeled dataset $D = \{X^{(1)}, ..., X^{(N)}\}$. A metric $\psi(\cdot) \in \Psi$ quantifies some property of M, such as fairness, with respect to social groups [9]. For this and other notations, see Table I.

TABLE I
Summary of Key Notation

Symbol	Definition
M	LLM model
θ	Parameters of the model M
X	Text sequence input
$\hat{Y}$	Output of the model M
D	Dataset
$\psi(\cdot)$	Metric quantifying some property of M

B. Defining Fairness and Bias

Social Bias: We broadly define social bias as disparate treatment or outcomes between social groups $G \in G$ arising from historical and structural power asymmetries [10]. In the context of NLP, this manifests as representational harms (misrepresentation, stereotyping, denigration, erasure) and allocational harms (discrimination in decisions or opportunities) [11]. Table II expands on types of social bias relevant to LLMs.

Fairness: Drawing on the fairness in machine learning literature [10], we consider several group fairness criteria that LLMs should ideally satisfy with respect to a set of protected attributes A and corresponding groups G. For a model M, outcome $\hat{Y} = M(X; \theta)$, and measure M, these include:

- **Demographic parity:**
  $P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A = a')$
  for groups $a, a' \in A$. In other words, positive outcomes are independent of group membership.
- **Equalized odds:**
  $P(\hat{Y} = 1|Y = y, A = a) = P(\hat{Y} = 1|Y = y, A = a')$
  for $y \in \{0, 1\}$ and $a, a' \in A$.
  True positive and false positive rates are equal across groups.

TABLE II
Taxonomy of Social Biases in NLP

Bias Type	Description
Gender Bias	perpetuates stereotypes or inequalities based on gender.
Racial Bias	results in the misrepresentation or marginalization of racial or ethnic groups.
Socioeconomic Bias	Bias in language models that reflects or reinforces social class disparities or prejudices.
Age Bias	ageism or stereotyping based on age groups.
Cultural Bias	favors or marginalizes certain cultural groups or norms.
Ideological Bias	influenced political or ideological perspectives, leading to partisan or biased outputs.
Linguistic Bias	Bias in language models that favors certain dialects, languages, or linguistic styles over others.

- **Predictive parity:**
  $P(Y = 1|\hat{Y} = \hat{y}, A = a) = P(Y = 1|\hat{Y} = \hat{y}, A = a')$
  for $\hat{y} \in \{0, 1\}$ and $a, a' \in A$.

See [12] for additional fairness definitions. Satisfying these criteria is often impossible simultaneously [13], and their suitability depends on the specific context and task. We propose diagnostic probes of this fairness for LLMs:

- For a prompt X with gendered terms substituted, an unbiased model M should produce semantically equivalent and equally likely outputs $\hat{Y}_i, \hat{Y}_j$:

$$\hat{Y}_i = \hat{Y}_j$$
$$\arg\max_Y P(Y|X_i)P(X_i) = \arg\max_Y P(Y|X_j)P(X_j)$$

- For prompts X_i, X_j describing an occupation with different racial terms, M should generate equally high-quality candidacy justifications.
- For a prompt X with dialect variation, M should produce equally positive sentiment continuations.

Relationship between Bias & Fairness: Following [14], we formalize intrinsic and extrinsic bias to violate fairness. Consider a pre-trained M that encodes X to $z = M(X)$. The intrinsic bias of M with respect to attribute A is:

$$|E_i(z) - E_i(z')| > \epsilon_i$$

where $E_i(\cdot)$ is an intrinsic metric (e.g., cosine similarity); $z = M(X_i)$, $z' = M(X_j)$; X_i, X_j are inputs for groups G_i, $G_j \in G$; $\epsilon_i \approx 0$ in the fair case. The extrinsic bias of a fine-tuned M' is:

$$|E_e(y) - E_e(y')| > \epsilon_e$$

where $E_e(\cdot)$ is an extrinsic metric (e.g., true positive rate); $y = M'(X_i)$, $y' = M'(X_j)$; $\epsilon_e \approx 0$ is fair.

III. Area Taxonomy

We analyze the literature through a general ML lens, which naturally extends to LLMs. This approach provides a well-established and grounded framework for examining LLM-relevant literature. To enhance the fairness and transparency of ML models, this paper presents a structured taxonomy that identifies points of failure and aligns them with solution

strategies. The taxonomy organizes potential biases and transparency issues into systematic categories, aiding in the comprehensive examination and management of these concerns. It is visually represented in Figure 1.

A. Points of Failure

This taxonomy examines three critical areas where failures may occur: The Data, The Algorithm, and The Model. Each is detailed with scenarios illustrating how biases can arise during data collection, algorithm design, and model application. This examination provides the necessary foundation to understand and tackle ethical issues in AI.

1) The Data: serves as the foundation for training LLMs. Biases in data, stemming from non-representative training sets or historically biased information, can perpetuate societal prejudices. Issues like data privacy and consent also play critical roles in data integrity.

2) The Algorithm: involves the computational methods and processes that govern how data is processed within models. Algorithmic failures can arise from the design and implementation of these methods, which may inadvertently introduce or fail to mitigate biases. Such failures often manifest as discrimination or inequity in decision-making processes.

3) The Model: refers to the output or the operational instance of an LLM. Failures here can be due to overfitting, underfitting, lack of generalizability, or transparency in the model's decision-making process. Model-based failures impact the reliability, accountability, and interpretability of LLM outputs.

B. Solution Approaches

The taxonomy categorizes interventions into three stages: pre-processing, in-processing, and post-processing, each targeting specific phases of the model lifecycle to ensure fairness and transparency.

1) Pre-Processing: This approach targets the initial stages involving The Data and The Algorithm. Pre-processing solutions aim to correct or mitigate biases in the data before it is used by the algorithm. Techniques include data augmentation, re-sampling, or the application of statistical methods to adjust data distributions. These methods are crucial for ensuring that the input to the algorithm is as unbiased and representative as possible.

2) In-Processing: solutions are integrated during the operation of all three: The Data, The Algorithm, and The Model. These solutions involve modifications directly within the algorithmic processes or model architecture to promote fairness. Examples include the implementation of fairness constraints or regularization techniques during the training of models. This approach is beneficial for dynamically addressing biases as the model interacts with data and refines its parameters.

3) Post-Processing: strategies focus on The Algorithm and The Model after a model has been trained. These approaches adjust the outputs of models to ensure fairness, often through techniques such as calibration or by altering decision thresholds for different groups. Post-processing is particularly useful for scenarios where pre- and in-processing modifications are insufficient or not feasible.

The taxonomy's structure progresses logically from data inception to model output, emphasizing interventions at critical points to prevent bias infiltration. By distinguishing solutions based on their implementation timing, the taxonomy offers a clear roadmap for effectively applying fairness and transparency measures. This approach not only identifies the potential for bias but also matches specific remedies to these vulnerabilities, facilitating a structured enhancement of ethical practices in LLM development and deployment. This systematic categorization aids in pinpointing vulnerabilities in LLMs and aligning them with effective mitigation strategies, thereby supporting the broader goal of ethical AI development.

IV. Taxonomy Based Survey

There are multiple pathways through which one can navigate the taxonomy, with options to explore either by points of failure or by solution approaches. To maintain clarity and coherence in our discussion, we will use points of failure as our primary navigation markers. This approach allows us to systematically reference relevant solution approaches as they become pertinent in our discussion. Thus, without further delay, let us delve into the detailed survey itself, examining each identified point of failure and exploring how targeted solution strategies can mitigate these risks.

A. The Data

Methods for addressing bias and ensuring fairness in data-driven models primarily focus on *pre-processing* techniques, although some, such as counterfactual logit pairing [15], are *post-processing* based. The work presented in [15], for instance, helps maintain consistency in model predictions across text pairs that only differ by the substitution of a specific identity term, thus promoting fairness in outcomes. This is almost like an extension of Demographic Parity and Predictive Parity in the context of LLMs. However, the majority of debiasing strategies involve some form of data pre-processing.

In scenarios where text classifiers are trained on imbalanced datasets, certain identity terms may be disproportionately represented, leading to biased predictions. For example, classifiers could associate the term "gay" with toxic labels due to its frequent occurrence in negatively biased reviews [5]. In such instance, pre-processing techniques like robust word substitution [16] could prove beneficial which is a technique that enhances the diversity in the representation of sensitive terms by substituting them with synonyms or related words, thus reducing the model's sensitivity to these terms and improving its generalization capabilities across different contexts. However, doing this on a very large data-set in and of itself might be a computationally challenging task, not to mention the problem of making sure that the substitutions themselves have a neutral sentiment associated with it (which is a different problem altogether). One must think about the trade-off that comes from employing such methods. This can be done through stakeholder engagement.

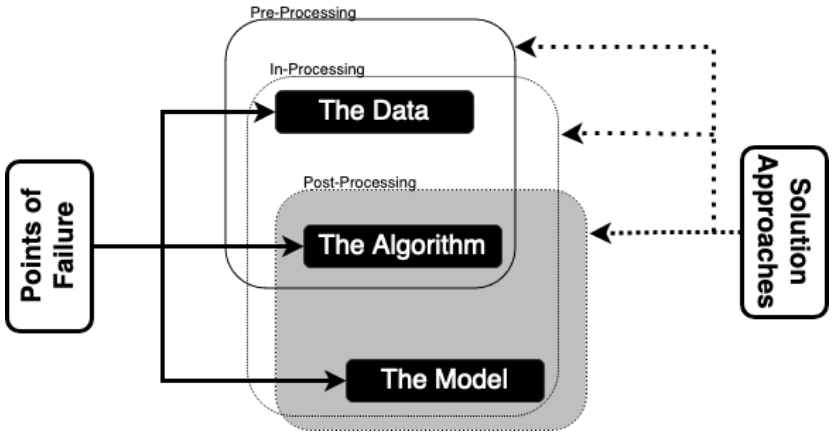

Fig. 1. Proposed taxonomy for Enhancing Fairness and Transparency in Large Language Models: This diagram categorizes the primary *points of failure*—The Data, The Algorithm, and The Model—alongside *targeted solution approaches* in pre-processing, in-processing, and post-processing stages to address biases and ensure transparency throughout the model's lifecycle.

There are other pre-processing techniques that are used commonly. In machine translation, such methods include labeling the gender of samples and creating a gender-balanced adaptation dataset [17]. For detecting toxic language, strategies range from employing transfer learning from less biased corpora [18] to relabeling samples considering dialect and race [7], automatically sensing dialects [19], and eliminating proxy words associated with identity terms [20]. In classification tasks, approaches like using an infinitesimal jackknife-based method ([21], [22]) adjust the corpus by selectively removing training data based on its impact on fairness metrics. This statistical method estimates the variance of an estimator through minor perturbations to the dataset, allowing for the assessment of variability without needing to exclude entire observations repeatedly. Such a method is particularly valuable for large datasets or when computational resources are constrained, providing a computationally efficient alternative to traditional jackknife methods while preserving accuracy, especially in scenarios with complex estimators or limited sample sizes. Here again, we run into a similar aforementioned problem of trade-offs and figuring out what might be important to the current task.

Moving along in the pre/in-processing category is Instance Weighting ([23], [24]), which involves adjusting the influence of individual training examples to mitigate biases in LLMs. This technique formalizes social bias as a shift from a non-discriminatory distribution to a discriminatory one [23] and calculates instance weights based on this shift, aiming to recover a non-discriminatory distribution from the selection bias. It is particularly useful in reducing the influence of biased instances on model attention during training for downstream tasks [24]. However, a limitation of this approach is its assumption that each instance of discrimination is drawn independently of non-discriminatory samples, which may not always hold true in real-world scenarios.

B. The Algorithm

Algorithmic issues in Large Language Models (LLMs) primarily stem from various sources that introduce biases during the model training and application phases. These biases can have significant impacts on the fairness and functionality of these models.

These prejudices have multiple important sources, including: Uncensored pre-training corpora containing damaging material or biased annotators adding personal subjectivity to labels are two examples of label bias that lead to models learning from examples including stereotypes [25]. When the training set's and the test set's demographic distributions diverge, sampling bias occurs and model biases are impacted by this distribution shift [26]. Furthermore, unanticipated biases in the language model encoding process are referred to as "semantic bias" and show up as biased semantic information in the embeddings [14].

To address these biases algorithmically, one effective technique is adversarial learning, an in-processing approach as depicted in our taxonomy in Figure 1. This method was originally popularized by Generative Adversarial Networks (GANs) [27], where a generator and a discriminator engage in a continual tug-of-war. Adversarially, for LLMs, this approach involves concealing sensitive information from the decision function through the interaction of an attacker and an encoder. The attacker attempts to detect protected attributes in the encoder's representation, while the encoder tries to prevent this detection within a given task ([28], [29]). This approach can generally be used to force the LLM to learn fairer representations. To be more precise, one way to do this is by modifying the loss function. Here, the idea is that adversarial learning trains an LLM and a discriminator for a protected attribute. The loss forces the LLM to fool the discriminator:

$$L_{adv} = \min_{\theta} \max_{\phi} \mathbb{E}x[L(D\phi(M_\theta(x), y))]$$

where L is the loss, D_ϕ is the discriminator with parameters ϕ, and M_θ is the LLM with parameters θ. This setup can

satisfy fairness notions like demographic parity and equality of odds by making the model output independent of protected attributes. Other relevant literature add regularization (discussed more in "The Model" subsection) terms to the loss to equalize probabilities of gendered words [30] or penalize higher likelihoods for stereotype-aligned text than anti-stereotypes :

$$L_{reg} = \lambda \sum_{i=1}^{N} \sum_{k=1}^{K_{BS}} \left| \log P(S_k^+ \mid S_k) - \log P(S_k^- \mid S_k) \right|$$

where λ is a hyperparameter, N_{BS} is the number of stereotype pairs, and S_k^+ and S_k^- are stereotype-aligned and anti-stereotype sentences. The model's encoding and prediction outputs may still contain sensitive information, even with the potential for significant bias reduction [31]. Future research examining these embedded biases and potential mitigation strategies is still possible.

Another similar algorithmic technique that can used are contrastive loss functions. The work by pushing the embeddings of different social groups apart while pulling same-group embeddings closer:

$$L_{con} = \sum_{(x,x^+)} \max(0, \sigma^+ - \cos(z, z^+)) + \sum_{(x,x^-)} \max(0, \cos(z, z^-) - \sigma^-)$$

where z, z^+, z^- are the representations of a main, positive, and negative input; σ^+, σ^- are margins; and (x, x^+) are same-group pairs while (x, x^-) are from different groups.

Finally, the most natural and intuitive pre/in-processing technique (especially within this context of LLMs) is modification of the model architecture itself. The nature and dexterity of LLMs allows for additional components to be "tacked on" to condition it for fairer outputs. Plug and play language models [32] combine a pre-trained LLM with a smaller tunable attribute model that shifts the generated distribution:

$$p_{PPLM}(x) \propto p_{LM}(x) \cdot p_{attr}(a|x)$$

where p_{LM} is the LLM distribution, p_{attr} is the attribute model distribution, and a is the desired attribute (e.g., non-toxic). SideNet [33] trains an external knowledge-enhanced network to re-rank an LLM's original top-k outputs to a less biased order using a regularized loss. Other approaches update a subset of the LLM parameters to mitigate bias while preserving other knowledge. Adapter-based de-biasing injects trainable adapter layers around the frozen pre-trained weights and fine-tunes only the adapters. ADEPT [34] replaces intermediate activations of sensitive neurons.

C. The Model

At the model level, numerous approaches aim to reduce bias and enhance fairness in machine learning models, with regularization being the most prominent among them. It serves both as a post- and in-processing solution to the "point of failure: model" issue. Regularization helps prevent overfitting by adding a penalty term to the objective function, which promotes simpler models that are more likely to generalize well to unseen data. In the context of fairness, regularization integrates fairness objectives into the training process by incorporating a term that penalizes causal and spurious features identified through a counterfactual framework [35], [36] which is achieved via manual tuning. However, this manual tuning of features could introduce bias from the individual performing the tuning and can be labor-intensive. While regularization techniques are pivotal for balancing fairness constraints against model performance, they can adversely affect the latter. Furthermore, these techniques may not generalize well across different datasets and domains and often require significant computational resources, potentially prolonging training times. As LLMs become larger and more complex, it remains to be seen whether regularization can provide a long-term benefit to fairness metrics in these models, or if efforts would be more efficiently allocated to other areas.

This discussion is not exhaustive of model-based fairness techniques; others, such as the use of auxiliary classifiers [37], [38] have been omitted from this taxonomy due to time constraints.

Table III gives a short summary overview of Problem Areas, Solution Approaches, and Examples in Enhancing Fairness and Transparency within LLMs that we have detailed in this section. It categorizes each identified point of failure within LLMs and aligns them with targeted solution strategies, offering a succinct summary of the surveyed literature and practical interventions.

TABLE III
SUMMARY OF PROBLEMS, SOLUTION APPROACHES, ALONG WITH EXAMPLES

Problem Area	Solution Approach	Examples
The Data	Pre-Processing	Robust word substitution [16] Gender-balanced dataset creation [17] Infinitesimal jackknife method ([21], [22])
	In-Processing	Instance weighting ([23], [24])
	Post-Processing	Counterfactual logit pairing [15]
The Algorithm	Pre-Processing	Data augmentation and cleansing Elimination of proxy words [20]
	In-Processing	Adversarial learning ([28], [29]) Regularization techniques for fairness Contrastive loss functions
	Post-Processing	Calibration and threshold adjustments
The Model	Pre-Processing	Data curation and validation
	In-Processing	Modification of model architecture Plug and play models [32] Adapter-based de-biasing [39]
	Post-Processing	Output re-ranking [33] Auxiliary classifiers

V. OPEN PROBLEMS AND CHALLENGES

Achieving equitable large language models (LLMs) necessitates an interdisciplinary research approach that spans technical, social, and humanistic disciplines.

1) Power Imbalances: LLMs are predominantly developed by Western industry labs and elite universities, embedding Anglo-American cultural norms into training data, benchmarks, and models. This trend obscures transparency in data origin and model specifications and often sidelines marginalized communities most at risk from biases. To counteract this,

participatory frameworks should involve impacted populations in all stages of development—from problem formulation to deployment [40]. Additionally, we need stronger representation from Global South perspectives and robust governance structures like national AI regulations and third-party audits.

2) Conceptual Foundations and Reliability of Bias Metrics: Current bias metrics simplify the complexities of identity by reducing them to binary comparisons, which fails to account for non-binary and intersectional identities. More nuanced evaluation methods and a shift towards prioritizing extrinsic metrics are recommended [41]. Collaboration between NLP, sociolinguistics, HCI [42], and science and technology studies [43] can deepen our understanding of how language functions as a vector for bias and inequality.

3) Evaluating Fairness in Large-sized LLMs: The fairness evaluation of large-sized LLMs is challenging due to limited accessibility and reliance on human judgment, which is resource-intensive and potentially biased. We advocate for automated, statistically grounded measurement techniques and the creation of diverse, large-scale datasets specifically designed for evaluating large-sized LLMs across various tasks.

4) Quantifying and Understanding the Causes of Bias: Bias in LLMs is multifaceted, originating from data-driven, psychological, political, and historical factors. Expanding research to include these diverse perspectives will provide a deeper understanding of biases and support the development of more equitable systems. Even in state-of-the-art automated techniques like adversarial methods, there is work to be done regarding the embedded biases in the learned models. We recommend comprehensive methodologies that integrate multiple analytical perspectives to better understand and quantify biases.

5) Efficient De-biasing Strategies for Large-sized LLMs: Current de-biasing methods are resource-intensive and not always scalable. We need to develop low-cost, scalable debiasing methods such as advanced prompt engineering and prompt tuning techniques. Moreover, incorporating fairness from the beginning of the data processing and model architecture stages during development can fundamentally reduce biases, making fairness a foundational aspect of LLM design.

6) Theoretical Limits and Mitigation Strategies: There is a crucial need for theoretical guarantees on the fairness-utility tradeoff. Studying LLM inductive biases and employing causal analyses [44] can inform proactive de-biasing strategies. Additionally, formal verification of fairness constraints [45] and ensuring the robustness and security of de-biasing methods against potential subversion are essential for maintaining the integrity of fairness measures [46].

By addressing these challenges, we can pave the way for developing more fair, transparent, and inclusive LLM technologies.

VI. Conclusion

In conclusion, our Taxonomy Based Survey methodically explores the multifaceted landscape of bias mitigation in Large Language Models through a well-defined taxonomy. By distinguishing between points of failure—The Data, The Algorithm, and The Model—we've delineated a clear path for navigating through various de-biasing strategies. This approach has enabled us to not only identify specific sources of bias but also to link these to targeted solution approaches, whether they be pre-processing, in-processing, or post-processing. Our detailed examination reveals that while significant strides have been made in developing techniques to address biases at different stages of model development, challenges remain. Continuous efforts to refine these techniques and develop new ones are essential as models evolve in complexity and application. By maintaining a focus on both emerging and established methods, we contribute to the broader goal of creating fairer and more transparent AI systems. This work lays a foundation for future explorations and improvements in the field, promising a more equitable technological future.

Acknowledgements

We are profoundly grateful to Professor Mingrui Liu for his engaging class, which not only enlightened us with intriguing insights into machine learning but also provided us the opportunity to explore these through our survey paper. This project allowed us to connect our personal interests with state-of-the-art ML models, deepening what we learned in his lectures. We also thank our classmates for a semester filled with vibrant discussions and valuable feedback on our presentations, which enriched our learning experience through a dynamic exchange of ideas.

References

[1] A. Vaswani, A. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, I. Polosukhin, L. Kaiser, and I. Polosukhin, "Attention is all you need," vol. 30, 2017, pp. 599–609.

[2] D. Bawden, A. Clement, and S. Drinkwater, *Managing digital library collections*. Library Association Publishing, 2008.

[3] T. Schick and H. Schütze, "Few-shot learning with language models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 935–944.

[4] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 610–623.

[5] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, "Measuring and mitigating unintended bias in text classification," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '18. New York, NY, USA: Association for Computing Machinery, Dec. 2018, p. 67–73. [Online]. Available: https://dl.acm.org/doi/10.1145/3278721.3278729

[6] A. Abid, M. Farooqi, and J. Zou, "Persistent anti-muslim bias in large language models," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 298–306.

[7] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The risk of racial bias in hate speech detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1668–1678. [Online]. Available: https://aclanthology.org/P19-1163

[8] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "Realtoxicityprompts: Evaluating neural toxic degeneration in language models," *arXiv preprint arXiv:2009.11462*, 2020.

[9] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[10] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

[11] A. Birhane, "Algorithmic injustice: a relational ethics approach," *Patterns*, vol. 2, no. 2, 2021.

[12] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the international workshop on software fairness*, 2018, pp. 1–7.

[13] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[14] R. Bansal, "A survey on bias and fairness in natural language processing," no. arXiv:2204.09591, Mar. 2022, arXiv:2204.09591 [cs]. [Online]. Available: http://arxiv.org/abs/2204.09591

[15] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel, "Counterfactual fairness in text classification through robustness," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '19. New York, NY, USA: Association for Computing Machinery, Jan. 2019, p. 219–226. [Online]. Available: https://dl.acm.org/doi/10.1145/3306618.3317950

[16] Y. Pruksachatkun, S. Krishna, J. Dhamala, R. Gupta, and K.-W. Chang, "Does robustness improve fairness? approaching fairness with word substitution robustness methods for text classification," in *ACL-IJCNLP 2021*, 2021. [Online]. Available: https://www.amazon.science/publications/does-robustness-improve-fairness-approaching-fairness-with-word-substitution-robustness-methods-for-text-classification

[17] E. Vanmassenhove, C. Hardmeier, and A. Way, "Getting gender right in neural machine translation," *ArXiv*, vol. abs/1909.05088, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:76654723

[18] J. H. Park, J. Shin, and P. Fung, "Reducing gender bias in abusive language detection," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2799–2804. [Online]. Available: https://aclanthology.org/D18-1302

[19] X. Zhou, M. Sap, S. Swayamdipta, Y. Choi, and N. Smith, "Challenges in automated debiasing for toxic language detection," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 3143–3155. [Online]. Available: https://aclanthology.org/2021.eacl-main.274

[20] S. Panda, A. Kobren, M. Wick, and Q. Shen, "Don't just clean it, proxy clean it: Mitigating bias by proxy in pre-trained models," in *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 5073–5085. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.372

[21] P. Sattigeri, S. Ghosh, I. Padhi, P. Dognin, and K. R. Varshney, "Fair infinitesimal jackknife: Mitigating the influence of biased training data points without refitting," no. arXiv:2212.06803, Dec. 2022, arXiv:2212.06803 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2212.06803

[22] R. G. Miller, "The jackknife–a review," *Biometrika*, vol. 61, no. 1, p. 1–15, 1974. [Online]. Available: https://www.jstor.org/stable/2334280

[23] G. Zhang, B. Bai, J. Zhang, K. Bai, C. Zhu, and T. Zhao, "Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 4134–4145. [Online]. Available: https://aclanthology.org/2020.acl-main.380

[24] X. Han, T. Baldwin, and T. Cohn, "Balancing out bias: Achieving fairness through balanced training," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 11 335–11 350. [Online]. Available: https://aclanthology.org/2022.emnlp-main.779

[25] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, p. 1–35, Jul. 2022. [Online]. Available: https://dl.acm.org/doi/10.1145/3457607

[26] D. S. Shah, H. A. Schwartz, and D. Hovy, "Predictive biases in natural language processing models: A conceptual framework and overview," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, p. 5248–5264. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.468

[27] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," no. arXiv:1406.2661, Jun. 2014, arXiv:1406.2661 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1406.2661

[28] S. Ravfogel, M. Twiton, Y. Goldberg, and R. D. Cotterell, "Linear adversarial concept erasure," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 18 400–18 421. [Online]. Available: https://proceedings.mlr.press/v162/ravfogel22a.html

[29] P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. Chi, "Fairness without demographics through adversarially reweighted learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 728–740.

[30] Y. Qian, U. Muaz, B. Zhang, and J. W. Hyun, "Reducing gender bias in word-level language models with a gender-equalizing loss function," *arXiv preprint arXiv:1905.12801*, 2019.

[31] Y. Elazar and Y. Goldberg, "Adversarial removal of demographic attributes from text data," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 11–21. [Online]. Available: https://aclanthology.org/D18-1002

[32] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu, "Plug and play language models: A simple approach to controlled text generation," no. arXiv:1912.02164, Mar. 2020, arXiv:1912.02164 [cs]. [Online]. Available: http://arxiv.org/abs/1912.02164

[33] X. Han, T. Baldwin, and T. Cohn, "Decoupling adversarial training for fair nlp," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 471–477.

[34] K. Yang, C. Yu, Y. R. Fung, M. Li, and H. Ji, "Adept: A debiasing prompt framework," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 9, 2023, pp. 10 780–10 788.

[35] S. Park, K. Choi, H. Yu, and Y. Ko, "Never too late to learn: Regularizing gender bias in coreference resolution," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. Singapore Singapore: ACM, Feb. 2023, p. 15–23. [Online]. Available: https://dl.acm.org/doi/10.1145/3539597.3570473

[36] Z. Wang, K. Shu, and A. Culotta, "Enhancing model robustness and fairness with causality: A regularization approach," in *Proceedings of the First Workshop on Causal Inference and NLP*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 33–43. [Online]. Available: https://aclanthology.org/2021.cinlp-1.3

[37] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg, "Null it out: Guarding protected attributes by iterative nullspace projection," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7237–7256. [Online]. Available: https://aclanthology.org/2020.acl-main.647

[38] H. Liu, W. Jin, H. Karimi, Z. Liu, and J. Tang, "The authors matter: Understanding and mitigating implicit bias in deep text classification," no. arXiv:2105.02778, May 2021, arXiv:2105.02778 [cs]. [Online]. Available: http://arxiv.org/abs/2105.02778

[39] A. Lauscher, T. Lüken, and G. Glavas, "Sustainable modular debiasing of language models," in *Conference on Empirical Methods in Natural Language Processing*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237440429

[40] E. M. Smith, M. Hall, M. Kambadur, E. Presani, and A. Williams, "" i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset," *arXiv preprint arXiv:2205.09209*, 2022.

[41] U. Hébert-Johnson, M. Kim, O. Reingold, and G. Rothblum, "Multicalibration: Calibration for the (computationally-identifiable) masses," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1939–1948.

[42] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse *et al.*, "Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned," *arXiv preprint arXiv:2209.07858*, 2022.

[43] R. Benjamin, *Race after technology: Abolitionist tools for the new Jim code*. John Wiley & Sons, 2019.

[44] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. M. Shieber, "Causal mediation analysis for interpreting neural nlp: The case of gender bias," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 388–12 401.

[45] O. Bastani, X. Zhang, and A. Solar-Lezama, "Formal verification of fairness properties via concentration," in *Proceedings of the 2019 ACM SIGPLAN International Conference on Programming Language Design and Implementation*, 2019, pp. 384–398.

[46] A. Havens, Z. Jiang, and S. Sarkar, "Robustness to adversarial attacks in learning-based control," in *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020, pp. 2452–2457.