Fairness in Large Language Models in Three Hours

Thang Viet Doan
Florida International University
Miami, FL, US
tdoan011@fiu.edu

Zichong Wang
Florida International University
Miami, FL, US
ziwang@fiu.edu

Nhat Nguyen Minh Hoang
Florida International University
Miami, FL, US
nhoan009@fiu.edu

Wenbin Zhang[*]
Florida International University
Miami, FL, US
wenbin.zhang@fiu.edu

Abstract

Large Language Models (LLMs) have demonstrated remarkable success across various domains but often lack fairness considerations, potentially leading to discriminatory outcomes against marginalized populations. Unlike fairness in traditional machine learning, fairness in LLMs involves unique backgrounds, taxonomies, and fulfillment techniques. This tutorial provides a systematic overview of recent advances in the literature concerning fair LLMs, beginning with real-world case studies to introduce LLMs, followed by an analysis of bias causes therein. The concept of fairness in LLMs is then explored, summarizing the strategies for evaluating bias and the algorithms designed to promote fairness. Additionally, resources for assessing bias in LLMs, including toolkits and datasets, are compiled, and current research challenges and open questions in the field are discussed. The repository is available at https://github.com/LavinWong/Fairness-in-Large-Language-Models.

CCS Concepts

• **General and reference** → **Surveys and overviews**.

Keywords

Large Language Model, Fairness, Social Sciences

ACM Reference Format:
Thang Viet Doan, Zichong Wang, Nhat Nguyen Minh Hoang, and Wenbin Zhang. 2024. Fairness in Large Language Models in Three Hours. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, Barcelona, Spain, 4 pages. https://doi.org/10.1145/3627673.3679090

1 Introduction

Large Language Models (LLMs), such as BERT [9], GPT-3 [5], and LLaMA [40], have shown powerful performance and development prospects in various tasks of Natural Language Processing due to their robust text encoding and decoding capabilities and discovered emergent capabilities (*e.g.,* reasoning) [8]. Despite their great performance, LLMs tend to inherit bias from multiple sources, including training data, encoding processes, and fine-tuning procedures, which may result in biased decisions against certain groups defined by the *sensitive attribute* (*e.g.,* age, gender, or race). The biased prediction has raised significant ethical and societal concerns, severely limiting the adoption of LLMs in high-risk decision-making scenarios such as hiring, loan approvals, legal sentencing, and medical diagnoses.

To this end, many efforts have been made to mitigate bias in LLMs [23, 44]. For example, one line of work extends traditional fairness notions—individual fairness and group fairness—to these models [6]. Specifically, individual fairness seeks to ensure similar outcomes for similar individuals [13, 45], while group fairness focuses on equalizing outcome statistics across subgroups defined by sensitive attributes [43, 42] (*e.g.,* gender or race). While these classification-based fairness notions are adept at evaluating bias in LLM's classification results [6], they fall short in addressing biases that arise during the LLM generation process [18]. In other words, LLMs demand a nuanced approach to measure and mitigate bias that emerges both in their outputs and during the generation process. This complexity motivates other lines of linguistic strategies that not only evaluate the accuracy of LLMs but also their propagation of harmful stereotypes or discriminatory language. For instance, a study examining the behavior of an LLM like ChatGPT revealed a concerning trend: it generated letters of recommendation that described a fictitious individual named Kelly (*i.e.,* a commonly female-associated name) as "warm and amiable", while describing Joseph (*i.e.,* a commonly male-associated name) as a "natural leader and role model". This pattern indicates that LLMs may inadvertently perpetuate gender stereotypes by associating higher levels of leadership with males, underscoring the need for more sophisticated mechanisms to identify and correct such biases.

These burgeoning and varied endeavors aimed at achieving fairness in LLMs [7, 16, 27] highlight the necessity for a comprehensive understanding of how different fair LLM methodologies are implemented and understood across diverse studies. Lacking clarity on these correspondences, the design of future fair LLMs can become challenging [4]. Consequently, there is a pressing need for a systematic tutorial elucidating the recent advancements in fair LLMs. However, although there are several tutorials that address fairness in machine learning algorithms, [15, 28, 37] these primarily focus on fairness in broader machine learning algorithms. There is a noticeable gap in inclusive resources that specifically address fairness within LLMs, distinguishing it from traditional models and discussing recent developments.

[*] Corresponding author

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679090

Our tutorial aims to bridge this gap by providing an up-to-date and comprehensive review of existing work on fair LLMs. It begins with a general overview of LLMs, followed by an analysis of the sources of bias inherent in their training processes. We then delve into the specific concept of fairness as it applies to LLMs, summarizing the strategies and algorithms employed to assess and enhance fairness. The tutorial also offers practical resources, including toolkits and datasets, that are essential for evaluating bias in LLMs. Furthermore, we explore the unique challenges of fairness in LLMs, such as those presented by word embeddings and the language generation process. Finally, the tutorial concludes by addressing the current research challenges and proposing future directions for this field.

Previous tutorial. To the best of our knowledge, no other tutorial on fairness in LLMs has been presented at CIKM or other similar venues.

2 Tutorial Outline

We plan to give a half-day tutorial (3 hours plus breaks). To ensure our tutorial remains engaging and interactive, we intend to accomplish as follows: **i) Case Studies Introduction.** We'll start with a series of case studies that highlight specific instances of bias within LLMs. By grounding our discussion in real-world examples, we aim to help contextualize the discussion and make it more relatable for the audience. We aim to encourage participants to share their thoughts on these cases and foster dialogue. **ii) Interactive Bias Discussion.** An integral part of our tutorial will involve presenting participants with various LLM outputs and prompts. We'll then facilitate a discussion to identify and analyze potential biases within these examples. **iii) Fair LLMs Discussion.** We will explore strategies and algorithms for developing fairer LLMs through practical examples. Following this, a presentation of useful tools and datasets for assessing fairness in LLMs will take place to provide participants with concrete tools and methodologies for fairness in LLMs. **iv) Q&A Discussion.** The tutorial will culminate in a Q&A session, allowing participants to ask questions and seek clarifications on any aspects of the session. Additionally, we will make tutorial materials, such as the description, presentation slides, and pre-recorded videos, available for post-tutorial access and dissemination.

2.1 Agenda

The outline of the tutorial is as follows:

- Part I: Background on LLMs (**30 minutes**)
  - Introduction to LLMs
  - Training Process of LLMs
  - Root Causes of Bias in LLMs
- Part II: Quantifying Bias in LLMs (**60 minutes**)
  - Demographic representation [5, 30, 31]
  - Stereotypical association [1, 5, 30]
  - Counterfactual fairness [29, 30]
  - Performance disparities [30, 41]
- Part III: Mitigating Bias in LLMs (**40 minutes**)
  - Pre-processing [14, 25, 44]
  - In-training [24, 34, 35]
  - Intra-processing [2, 19, 32]
  - Post-processing [11, 22, 39]
- Part IV: Resources for Evaluating Bias (**30 minutes**)
  - Toolkits [3, 21, 38]
  - Datasets [10, 20, 26, 33, 36]
- Part V: Challenges and Future Directions (**20 minutes**)
  - Formulating Fairness Notions
  - Rational Counterfactual Data Augmentation
  - Balancing Performance and Fairness in LLMs
  - Fulfilling Multiple Types of Fairness
  - Developing More and Tailored Datasets

2.2 Content

Background on LLMs. We start by providing the audience with fundamental knowledge about LLMs. Next, we briefly explain the key steps required to train LLMs, including 1) data preparation and preprocessing, 2) model selection and configuration, 3) instruction tuning, and 4) alignment with humans. By examining the training process in detail, we identify and discuss three primary sources contributing to bias in LLMs: i) training data bias, ii) embedding bias, and iii) label bias.

Quantifying Bias in LLMs. To evaluate bias in LLMs, the primary method involves analyzing bias associations in the model's output when responding to input prompts. These evaluations can be conducted through various strategies including demographic representation, stereotypical association, counterfactual fairness, and performance disparities [12].

Demographic representation [5, 30, 31] evaluation method assesses bias by analyzing the frequency of demographic word references in the text generated by a model in response to a given prompt [27]. In this context, bias is defined as a systematic discrepancy in the frequency of mentions of different demographic groups within the generated text.

Stereotypical association [1, 5, 30] method assesses bias by measuring the disparity in the rates at which different demographic groups are linked to stereotyped terms (*e.g.,* occupations) in the text generated by the model in response to a given prompt [30]. In this context, bias is defined as a systematic discrepancy in the model's associations between demographic groups and specific stereotypes, which reflects societal prejudices.

Counterfactual fairness [29, 30] evaluates bias by replacing terms characterizing demographic identity in the prompts and then observing whether the model's responses remain invariant [27]. Bias in this context is defined as the model's sensitivity to demographic-specific terms, measuring how changes to these terms affect its output.

Performance disparities [30, 41] method assesses bias by measuring the differences in model performance across various demographic groups on downstream tasks. Bias in this context is defined as the systematic variation in accuracy or other performance metrics when the model is applied to tasks involving different demographic groups.

Mitigating Bias in LLMs. We systematically categorize bias mitigation algorithms based on their intervention stage within the processing pipeline.

Pre-processing methods change the data given to the model, like training data and prompts. They do this by using methods like data augmentation [44] and prompt tuning [14, 25].

In-training methods aim to alter the training process to minimize bias. This includes making modifications to the optimization process by adjusting the loss function [34] and incorporating auxiliary modules [24, 35].

Intra-processing methods mitigate bias in pre-trained or fine-tuned models during inference without additional training. This technique includes a range of methods, such as model editing [2, 32] and decoding modification [19].

Post-processing methods modify the results generated by the model to reduce biases, which is crucial for closed-source LLMs where direct modification is limited. We use methods such as chain-of-thought [11, 22] and rewriting [39] as illustrative approaches to convey this concept.

Resource for Evaluating Bias. In this part, we introduce existing resources for evaluating bias in LLMs. First, we present three essential tools: Perspective API [21], developed by Google Jigsaw, detects toxicity in text; AI Fairness 360 (AIF360) [3], an open-source toolkit with various algorithms and tools; and Aequitas [38], another open-source toolkit, audits fairness and bias in LLMs, aiding data scientists and policymakers.

Next, we summarize worth-noting datasets referenced in the literature, categorized into probability-based and generation-based. Probability-based datasets, like WinoBias [36], BUG [26], and CrowS-Pairs [33], use template-based formats or counterfactual-based sentences. Generation-based datasets, such as RealToxicityPrompts [20] and BOLD [10], specify the first few words of a sentence and require a continuation. Besides, we will introduce TabLLM [17], a general framework to leverage LLMs for the classification of tabular data. That approach aims to address the challenge of using LLMs on structured tabular datasets, which are used in high-stakes domains for classification tasks.

Challenges and future directions. The tutorial concludes by exploring open research problems and future directions. Firstly, we discuss the challenges of ensuring fairness in LLMs. Defining fairness in LLMs is complex due to diverse forms of discrimination requiring tailored approaches to quantify bias, where definitions can conflict. Rational counterfactual data augmentation, a technique to mitigate bias, often produces inconsistent data quality and unnatural sentences, necessitating more sophisticated strategies. In addition, balancing performance and fairness involves adjusting the loss function with fairness constraints, but finding the optimal trade-off is challenging due to high costs and manual tuning.

For future directions, it is imperative to address multiple types of fairness concurrently, as bias in any form is undesirable. . Additionally, there is a pressing need for more tailored benchmark datasets, as current datasets follow a template-based methodology that may not accurately reflect various forms of bias.

3 Target audience and prerequisites for the tutorial

The tutorial is designed for researchers and practitioners in data mining, artificial intelligence, social science and other interdisciplinary areas, aiming to cater to individuals with varying degrees of expertise. The prerequisites include basic knowledge of probability, linear algebra, and machine learning, while prior knowledge of algorithmic fairness or specific algorithms is not a prerequisite, ensuring accessibility to beginners. This tutorial is designed for 40% novice, 30% intermediate, and 30% expert in order to achieve a good balance between the introductory and advanced materials. To foster a dynamic and participatory learning environment, the tutorial will intersperse lectures with discussion sessions, encouraging attendees to engage, ask questions, and share insights. Furthermore, to extend the tutorial's reach and impact, all materials, ranging from descriptions and slides to pre-recorded videos, will be available for post-tutorial access, supporting continued education and exploration of fairness in LLMs across diverse audiences.

4 Tutors' short bio and expertise related to the tutorial

Thang Viet Doan is a Ph.D. student in the Knight Foundation School of Computing and Information Sciences at Florida International University. He holds a Bachelor's degree in Computer Science from Hanoi University of Science and Technology (HUST). His current research interests are mainly focused on detecting and mitigating social bias in natural language systems.

Zichong Wang is currently pursuing his Ph.D. in the Knight Foundation School of Computing and Information Sciences at Florida International University. His research is centered on mitigating inadvertent disparities resulting from the interaction of algorithms, data, and human decisions in policy development. His work has been honored with the Best Paper Award at FAccT'23 and is a candidate for the Best Paper Award at ICDM'23. Additionally, he actively contributes as a member of the Program Committee/Reviewers for esteemed conferences and journals, including KDD, IJCAI, ICML, ICLR, FAccT, ECML-PKDD, ECAI, PAKDD, Machine Learning, and Information Sciences.

Minh Nhat Hoang Nguyen is a Ph.D. student at the Knight Foundation School of Computing and Information Sciences, Florida International University. He earned his Bachelor's degree in Data Science and Artificial Intelligence from Hanoi University of Science and Technology (HUST). His research focuses on detecting potential bias in machine learning algorithms, data quality and applying bias mitigation handling methods to deliver fairness in social application.

Wenbin Zhang is an Assistant Professor in the Knight Foundation School of Computing and Information Sciences at Florida International University, and an Associate Member at the Te Ipu o te Mahara Artificial Intelligence Institute. His research investigates the theoretical foundations of machine learning with a focus on societal impact and welfare. In addition, he has worked in a number of application areas, highlighted by work on healthcare, digital forensics, geophysics, energy, transportation, forestry, and finance. He is a recipient of best paper awards/candidates at FAccT'23, ICDM'23, DAMI, and ICDM'21, as well as the NSF CRII Award and recognition in the AAAI'24 New Faculty Highlights. He also regularly serves in the organizing committees across computer science and interdisciplinary venues, most recently Travel Award Chair at AAAI'24, Volunteer Chair at WSDM'24 and Student Program Chair at AIES'23.

5 Potential Societal Impacts

This tutorial possesses significant potential for positive societal impacts: i) By illuminating the nuances of fairness in LLMs, it endeavors to ignite research interest and catalyze efforts aimed at advancing fairness within this domain. Given the early stages of current initiatives addressing fairness in LLMs, this tutorial stands as a pivotal milestone in galvanizing further exploration and innovation in the field. ii) Through the exploration of new challenges that remain unaddressed in existing literature, this tutorial has the potential to inspire innovative approaches within the realm of LLMs fairness. By shedding light on these issues, it aims to stimulate critical discourse and foster the development of comprehensive solutions that address the complexities inherent in ensuring fairness within LLMs. iii) In addition to addressing fairness issues, this tutorial emphasizes the importance of developing new datasets that reflect diverse and representative forms of bias. By highlighting gaps in current datasets, it encourages the creation of new ones, aiming to support more accurate and equitable LLM training processes. iv) Beyond its immediate focus on fairness in LLMs, this tutorial endeavors to extend its impact on related research topics by uncovering new problems and elucidating their interconnectedness with fairness considerations. By identifying emerging issues, it seeks to foster interdisciplinary collaboration and facilitate holistic advancements in understanding and addressing societal concerns surrounding LLMs, thus contributing to broader societal progress and well-being.

Acknowledgement

This work was supported in part by the National Science Foundation (NSF) under Grant No. 2245895.

References

[1] Abubakar Abid, Maheen Farooqi, and James 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.

[2] Afra Feyza Akyürek et al. “DUnE: Dataset for unified editing”. In: *arXiv preprint arXiv:2311.16087* (2023).

[3] Rachel KE Bellamy et al. “AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias”. In: *IBM Journal of Research and Development* 63.4/5 (2019), pp. 4–1.

[4] Su Lin Blodgett et al. “Language (technology) is power: A critical survey of" bias" in nlp”. In: *arXiv preprint arXiv:2005.14050* (2020).

[5] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[6] Garima Chhikara et al. “Few-Shot Fairness: Unveiling LLM's Potential for Fairness-Aware Classification”. In: *arXiv preprint arXiv:2402.18502* (2024).

[7] Zhibo Chu, Zichong Wang, and Wenbin Zhang. “Fairness in Large Language Models: A Taxonomic Survey”. In: *ACM SIGKDD explorations newsletter* (2024).

[8] Zhibo Chu et al. “History, Development, and Principles of Large Language Models-An Introductory Survey”. In: *arXiv preprint arXiv:2402.06853* (2024).

[9] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[10] Jwala Dhamala et al. “Bold: Dataset and metrics for measuring biases in open-ended language generation”. In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 862–872.

[11] Harnoor Dhingra et al. “Queer people are people first: Deconstructing sexual identity stereotypes in large language models”. In: *arXiv preprint arXiv:2307.00101* (2023).

[12] Thang Viet Doan et al. *Fairness Definitions in Language Models Explained*. 2024. arXiv: 2407.18454 [cs.CL]. URL: https://arxiv.org/abs/2407.18454.

[13] Cynthia Dwork et al. “Fairness through awareness”. In: *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012, pp. 214–226.

[14] Zahra Fatemi et al. “Improving gender fairness of pre-trained language models without catastrophic forgetting”. In: *arXiv preprint arXiv:2110.05367* (2021).

[15] Rayid Ghani et al. “Addressing bias and fairness in machine learning: A practical guide and hands-on tutorial”. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023, pp. 5779–5780.

[16] Vipul Gupta et al. “Sociodemographic Bias in Language Models: A Survey and Forward Path”. In: (2024).

[17] Stefan Hegselmann et al. “Tabllm: Few-shot classification of tabular data with large language models”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 5549–5581.

[18] Taojun Hu and Xiao-Hua Zhou. “Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions”. In: *arXiv preprint arXiv:2404.09135* (2024).

[19] Po-Sen Huang et al. “Reducing sentiment bias in language models via counterfactual evaluation”. In: *arXiv preprint arXiv:1911.03064* (2019).

[20] Yue Huang, Qihui Zhang, Lichao Sun, et al. “Trustgpt: A benchmark for trustworthy and responsible large language models”. In: *arXiv preprint arXiv:2306.11507* (2023).

[21] Google Jigsaw. *Perspective API*. 2017. URL: https://www.perspectiveapi.com/.

[22] Masahiro Kaneko et al. “Evaluating Gender Bias in Large Language Models via Chain-of-Thought Prompting”. In: *arXiv preprint arXiv:2401.15585* (2024).

[23] Hadas Kotek, Rikker Dockum, and David Sun. “Gender bias and stereotypes in large language models”. In: *Proceedings of The ACM Collective Intelligence Conference*. 2023, pp. 12–24.

[24] Anne Lauscher, Tobias Lueken, and Goran Glavaš. “Sustainable modular debiasing of language models”. In: *arXiv preprint arXiv:2109.03646* (2021).

[25] Brian Lester, Rami Al-Rfou, and Noah Constant. “The power of scale for parameter-efficient prompt tuning”. In: *arXiv preprint arXiv:2104.08691* (2021).

[26] Shahar Levy, Koren Lazar, and Gabriel Stanovsky. “Collecting a large-scale gender bias dataset for coreference resolution and machine translation”. In: *arXiv preprint arXiv:2109.03858* (2021).

[27] Y Li et al. “A survey on fairness in large language models. arXiv. doi: 10.48550”. In: *arXiv preprint arXiv.2308.10149* (2023).

[28] Yunqi Li, Yingqiang Ge, and Yongfeng Zhang. “Tutorial on fairness of machine learning in recommender systems”. In: *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 2021, pp. 2654–2657.

[29] Yunqi Li and Yongfeng Zhang. “Fairness of chatgpt”. In: *arXiv preprint: 2305.18569* (2023).

[30] Percy Liang et al. “Holistic evaluation of language models”. In: *arXiv preprint arXiv:2211.09110* (2022).

[31] Justus Mattern et al. “Understanding stereotypes in language models: Towards robust measurement and zero-shot debiasing”. In: *arXiv preprint arXiv:2212.10678* (2022).

[32] Eric Mitchell et al. “Fast model editing at scale”. In: *arXiv preprint arXiv:2110.11309* (2021).

[33] Aurélie Névéol et al. “French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022, pp. 8521–8531.

[34] SunYoung Park et al. “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*. 2023, pp. 15–23.

[35] Shauli Ravfogel et al. “Null it out: Guarding protected attributes by iterative nullspace projection”. In: *arXiv preprint arXiv:2004.07667* (2020).

[36] Rachel Rudinger et al. “Gender bias in coreference resolution”. In: *arXiv preprint arXiv:1804.09301* (2018).

[37] Pedro Saleiro, Kit T Rodolfa, and Rayid Ghani. “Dealing with bias and fairness in data science systems: A practical hands-on tutorial”. In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2020, pp. 3513–3514.

[38] Pedro Saleiro et al. “Aequitas: A bias and fairness audit toolkit”. In: *arXiv preprint arXiv:1811.05577* (2018).

[39] Ewoenam Kwaku Tokpo and Toon Calders. “Text style transfer for bias mitigation using masked language modeling”. In: *arXiv preprint arXiv:2201.08643* (2022).

[40] Hugo Touvron et al. “Llama: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971* (2023).

[41] Yuxuan Wan et al. “Biasasker: Measuring the bias in conversational ai system”. In: *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 2023, pp. 515–527.

[42] Zichong Wang et al. “: Fairness-aware graph generative adversarial networks”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2023, pp. 259–275.

[43] Zichong Wang et al. “Preventing discriminatory decision-making in evolving data streams”. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 2023, pp. 149–159.

[44] Vithya Yogarajan et al. “Tackling Bias in Pre-trained Language Models: Current Trends and Under-represented Societies”. In: *arXiv preprint arXiv:2312.01509* (2023).

[45] Wenbin Zhang et al. “Individual fairness under uncertainty”. In: *ECAI 2023*. IOS Press, 2023, pp. 3042–3049.