Addressing Bias and Hallucination in Large Language Models

Nihar Sahoo∗, Ashita Saxena∗, Kishan Maharaj∗, Arif Ahmad∗, Abhijit Mishra†, Pushpak Bhattacharyya∗
∗CFILT, Indian Institute of Technology Bombay, India,
† University of Texas at Austin, Texas, USA
{nihar, ashitasaxena, kishan, pb}@cse.iitb.ac.in, arifahmadpeace@gmail.com, abhijitmishra@utexas.edu

Abstract

In the landscape of natural language processing (NLP), addressing the challenges of bias and hallucination is paramount to ensuring the ethical and unbiased development of Large Language Models (LLMs). This tutorial delves into the intricate dimensions of LLMs, shedding light on the critical importance of understanding and mitigating the profound impacts of bias and hallucination. The tutorial begins with discussions on the complexity of bias propagation in LLM development, where we dissect its origins and far-reaching impacts along with the automatic evaluation metrics for bias measurement. We then present innovative methodologies for mitigating diverse forms of bias, including both static and contextualized word embeddings and robust benchmarking strategies. In addition, the tutorial explores the interlinkage between hallucination and bias in LLMs by shedding light on how bias can be perceived as a hallucination problem. Furthermore, we also talk about cognitively-inspired deep learning frameworks for hallucination detection which leverages human gaze behavior. Ultimately, this cutting-edge tutorial serves as a guiding light, equipping participants with indispensable tools and insights to navigate the ethical complexities of LLMs, thus paving the way for the development of unbiased and ethically robust NLP systems.

1. Introduction

Large Language Models (LLMs) represent a cutting-edge class of AI models guided by specific prompts to generate tailored outputs, revolutionizing diverse sectors worldwide. These models, exemplified by ChatGPT and Google Bard, alongside open-source counterparts like Dolly 2.0 and LLaMa2.0, have garnered immense popularity. LLMs are poised to underpin transformative advancements across developed and developing societies, including facilitating cross-language communication, personalizing education, propelling healthcare innovations, ultimately ensuring broader accessibility to digital content and services for diverse audiences. However, amidst their astounding capabilities, LLMs are not without their challenges. This tutorial provides a comprehensive overview of two critical aspects of LLMs: *bias* and *hallucination*, with a predominant focus on *bias*.

We begin the tutorial with a primer on Language Models (LLMs), providing an overview of their training methods, variations, and historical development. We also highlight the ethical considerations pertinent to their deployment in practical contexts.

Given the significant impact of bias in LLMs, we then proceed to the first segment where, we define bias formally, outlining its types and the rationale behind its study. Subsequently, we explore the origins of bias in NLP pipelines, with a particular emphasis on the role of hallucination in the propagation of biased content and its implications in different domains. To address and alleviate bias, we then present several approaches, focusing on methods for both static and contextualized word embeddings. The importance of benchmarking datasets in the identification of bias is underscored, alongside an introduction to specific benchmarks tailored for quantifying bias, including the extraction of social bias from hate speech.

We then discuss bias from the lens of hallucination, which highlights the parallel between the presence of bias and hallucination. We conclude this discussion with a glimpse of cognitively inspired hallucination detection.

We hope this tutorial acts as a beacon, providing participants with essential resources and knowledge to navigate the ethical intricacies of LLMs, thereby facilitating the creation of impartial and morally sound NLP systems. We have made all the materials of this tutorial publicly available [1].

2. Target Audience

The target audiences include researchers and industry practitioners working on NLP tasks who extensively use LLMs for research or applications. This tutorial will give them an in-depth understanding of how to develop and fine-tune efficient yet ethically sound LLMs. We will also provide application-based demos and code walkthroughs for programming enthusiasts interested in the internal workings of these techniques.

[1] Tutorial Website

LREC-COLING 2024 Tutorials, pages 73–79
20-25 May, 2024.

3. Outline

Duration: Half Day

3.1. Introduction to LLMs

[Duration: 20 mins]

1. Language modeling: Task and Types
2. LLM paradigms: Dataset, training, evaluation
3. Evolution of LLMs
4. Ethical concerns

3.2. Understanding of Bias in LLMs

[Duration: 15 mins]

1. Bias definition and its types
2. Sources of bias in LLM development pipelines
3. Hallucination as a reason for bias
4. Downstream impact

3.3. Approaches for Bias detection

[Duration: 40 mins]

1. Bias Metrics: WEAT, SEAT, and MAC
2. Bias assessment in static word embeddings: Using PCA and Nullspace projection
3. Identifying Undesirable associations in Transformers: multi-headed attention Layer analysis
4. Intersectional biases across social axes: Gender and Race, Gender and Religion
5. Datasets and source of biases within data
6. Popular multilingual approaches: Few-shot, continuous pretraining, and prompting

Tea Break

3.4. Approaches for bias mitigation

[Duration: 40 mins]

1. Word embeddings: Soft and Hard debiasing
2. Debiasing context-representations
3. Designing Fairness-oriented loss functions
4. Counter-narratives based Debasing
5. Debiasing using prompting

3.5. Bias benchmarking Datasets

[Duration: 25 mins]

1. Importance of benchmarking datasets
2. Benchmarks for bias quantification: Stereoset, Crows-Pairs, BBQ, BIOS, and IndiBias

3.6. Bias from the lens of Hallucination

[Duration: 10 mins]

1. Parallels between the presence of bias and hallucination in machine-generated text
2. Possible causes of biases in hallucinated content

3.7. Cognitively inspired approaches for Hallucination detection

[Duration: 10 mins]

1. Basics of cognitively inspired deep learning methods
2. Behavioural insights related to hallucination and attention bias
3. Cognitively inspired deep learning architecture for hallucination detection

3.8. Open Problems and Future scope

[Duration: 10 mins]

3.9. Conclusion and Closing Remarks

[Duration: 10 mins]

4. Outline Description

4.1. Introduction to LLMs

The introduction section, spanning 20 minutes, outlines the fundamental aspects of Language Models (LLMs) by discussing language modeling as a task and the various types of such models. It further highlights the key paradigms governing LLMs, including dataset, training, and evaluation, while tracing their evolutionary trajectory. Lastly, the segment underscores the ethical considerations associated with the use of LLMs.

4.2. Understanding of Bias in LMs

In this section, spanning 30 minutes, the focus is on comprehending bias in Language Models (LMs). The discussion includes an elucidation of bias and its various types, such as gender, racial, and cultural biases (Singh et al., 2022; Crawford, 2017). We will also discuss data-bias, algorithmic and user-interaction driven biases (Hovy and Spruit, 2016; Vig et al., 2020) and highlight the role of hallucination as a contributing factor, followed by the downstream impacts of bias across various sensitive domains such as healthcare.

4.3. Approaches for Bias Detection

This section of 45 minutes covers NLP-based bias detection methods. Initially, we discuss the methodologies that quantify text data bias using WEAT (Caliskan et al., 2017), SEAT (Liang et al., 2020), and MAC (Manzini et al., 2019) metrics. Then we discuss the methods for detecting biases at various levels of text-processing, *e.g.,* word-embeddings (Bolukbasi et al., 2016) followed by contextualized sentence embeddings (Zhao et al., 2019; Garimella et al., 2021). The section also discusses intersectional biases (Tan and Celis, 2019; Lalor et al.,

2022) in different languages and cultures. The importance of dataset biases and bias detection methods for multilingual LLMs (Sahoo et al., 2023), including few-shot and continuous pretraining, will also be highlighted.

4.4. Approaches for bias mitigation

This segment covers various techniques for mitigating bias, including strategies such as soft and hard debiasing in word embeddings (Bolukbasi et al., 2016), and debiasing context-representations in Transformer based models. We will also delve into modern zero-shot techniques such as debiasing via prompts that guide models to produce unbiased results at inference time (Guo et al., 2022; Schick et al., 2021). Some other relevant topics such as Fairness-oriented Loss Functions (Zhang et al., 2018), counter-narratives (Sahoo et al., 2024a) based language rectification and debiasing (Sahoo et al., 2022) will also be highlighted.

4.5. Bias benchmarking datasets

In this section, we will discuss the significance of benchmarking datasets for bias evaluation. Several benchmarking datasets, such as Stereoset (Nadeem et al., 2021), Crows-Pairs (Nangia et al., 2020), BBQ (Parrish et al., 2022), BIOS (De-Arteaga et al., 2019), and IndiBias (Sahoo et al., 2024b), have emerged as valuable tools for measuring and assessing bias in language models. These benchmarks facilitate a standardized approach to assessing and comparing the performance of models in terms of bias mitigation and awareness.

Then we will discuss the biased behavior of the model from the lens of hallucination and conclude the overall tutorial with open questions, Q&A with audience followed by closing remarks.

4.6. Bias from the lens of Hallucination

In this section, we will highlight the presence of bias in hallucinated content. Hallucination is a challenging problem in this era of LLMs. The hallucinated content often contain biases. We will talk about the causes of biases and hallucinations and their similarities in this section.

4.7. Cognitively inspired approaches for Hallucination detection

In this section, we will draw parallels between human cognitive behaviour and deep learning methodologies for addressing the problem of hallucination detection (Mahowald et al., 2023; Maharaj et al., 2023). We will delve into the diverse cognitive insights and advantages that arise from integrating cognitive signals such as human eye-tracking data into deep learning-based architectures for hallucination assessment.

5. Diversity Considerations

We acknowledge the critical importance of incorporating diverse perspectives in the discussion of bias and hallucination within LLMs. This tutorial emphasizes the significance of including voices from underrepresented communities and diverse backgrounds, recognizing the nuanced impact of cultural and linguistic diversity on the understanding and mitigation of bias and hallucination. Notably, all presenters hail from different regions of India and the USA, representing a rich tapestry of language and cultural backgrounds, fostering a comprehensive exploration of these intricate NLP challenges from various global viewpoints.

6. Reading List

We intend to make the tutorial self-contained. The tutorial materials such as the slides and video recordings will published for later reference. Further reading materials beyond the content of this tutorial will be provided in the slides itself.

7. Presenters

Nihar Sahoo is a PhD student in the Computer Science department of IIT Bombay, supervised by Prof. Pushpak Bhattacharyya. His research interest lies in Ethical AI, social biases/toxicity in languages, and explainability in NLP. He has given a tutorial on *social bias detection and mitigation in NLP* at ICON. He has published papers on bias detection at conferences such as BMVC, LREC, CoNLL, NAACL, AAAI, ACL.

Ashita Saxena is a 3rd year MS by Research (CSE) student at IIT Bombay guided by Prof. Pushpak Bhattacharyya. Her research focuses on hallucination detection and mitigation in NLP tasks and her work is published in EMNLP. She has worked as a Research Intern at IBM Research on Natural Language Generation (NLG).

Kishan Maharaj is an MS (by Research) student at IIT Bombay (CSE), guided by Prof. Pushpak Bhattacharyya. His research focuses on cognitively inspired natural language processing, specifically hallucination detection and mitigation. His work was published in EMNLP. He is currently working with IBM research on prompt-based hallucination mitigation. Formerly, he worked with Turtle Mint and TATA Sons on various data science problems.

Arif Ahmad is currently in the final year of a BTech/MTech dual degree in Electrical Engineering and AI at IIT Bombay. He is working in the area of Fairness and Bias in NLP systems and

Models, under the supervision of Prof. Pushpak Bhattacharyya at the CFILT Lab in IIT Bombay.

Dr. Abhijit Mishra an Assistant Professor of Practice at the School of Information, University of Texas at Austin, boasts extensive experience in ML and NLP, spanning over a decade. Formerly a Research Scientist at Apple Inc. and IBM Research, his contributions to NLP-based products like Siri and Watson are noteworthy. With notable publications at key AI and NLP conferences such as ACL, EMNLP, and AAAI, he has demonstrated expertise in various NLP domains, including multilingual and multimodal Natural Language Understanding and Generation, Sentiment Analysis, and Cognitive NLP with eye-tracking. Dr. Mishra's recent focus on ethical LLM development aligns closely with the theme of the tutorial.

Prof. Pushpak Bhattacharyya is a Professor of Computer Science and Engineering at IIT Bombay. Educated in the IIT System (B.Tech IIT Kharagpur, M.Tech IIT Kanpur, PhD IIT Bombay), Dr. Bhattacharyya has done extensive research in Natural Language Processing and Machine Learning. He has published more than 350 research papers, has authored/co-authored 6 books including a textbook on machine translation, and has guided more than 350 students for their PhD, Masters and Undergraduate thesis. He has received many Research Excellence Awards- Manthan award from Ministry of IT, H.H. Mathur and P.K.Patwardhan awards from IIT Bombay, VNMM award from IIT Roorkee, and substantial research grants from Government and industry. Prof. Bhattacharyya holds the Bhagat Singh Rekhi Chair Professorship of IIT Bombay, is a Fellow of National Academy of Engineering, Abdul Kalam National Fellow, Distinguished Alumnus of IIT Kharagpur, past Director of IIT Patna and past President of ACL.

8. Other Information

We anticipate the active participation of approximately 100 individuals, estimated based on the past engagement with similar tutorials and the current outreach efforts. This estimate takes into account the projected interest within the NLP community, specifically on responsible LLM development and aligns with our preparation for interactive sessions and engaging discussions.

9. Ethics Statement

At the core of our tutorial on "Addressing Bias and Hallucinations in Large Language Models" lies a commitment to addressing the ethical concerns of NLP. We recognize that NLP technologies have profound societal impacts, and as educators and researchers, we have a responsibility to raise awareness about potential issues, promote ethical practices, and foster a deeper understanding of bias and hallucination in NLP systems.

10. Bibliographical References

Ayush Agrawal, Lester Mackey, and Adam Tauman Kalai. 2023. Do language models know when they're hallucinating references? *arXiv preprint arXiv:2305.18248*.

Afra Feyza Akyürek, Sejin Paik, Muhammed Yusuf Kocyigit, Seda Akbiyik, Şerife Leman Runyun, and Derry Wijaya. 2022. On measuring social biases in prompt-based multi-task learning.

Md Abdul Aowal, Maliha T Islam, Priyanka Mary Mammen, and Sandesh Shetty. 2023. Detecting natural language biases with prompt-based learning.

Simran Arora, Avanika Narayan, Mayee F Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. 2022. Ask me anything: A simple strategy for prompting language models. *arXiv preprint arXiv:2210.02441*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroglu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049. Association for Computational Linguistics.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pages 803–811. PMLR.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual*

Meeting of the Association for Computational Linguistics, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Kate Crawford. 2017. The trouble with bias.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William W Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. *arXiv preprint arXiv:1906.01081*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754*.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. *arXiv preprint arXiv:2011.10819*.

Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *arXiv preprint arXiv:2104.08455*.

Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. 2022. Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4295–4309.

Fanton, Margherita. 2021. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870.

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, Anandhavelu N, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? On Mitigating Social Biases in Language Modelling and Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545, Online. Association for Computational Linguistics.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. *arXiv preprint arXiv:2010.05478*.

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

Tianxing He, Jingzhao Zhang, Zhiming Zhou, and James Glass. 2019. Exposure bias versus self-recovery: Are distortions really incremental for autoregressive text generation? *arXiv preprint arXiv:1905.10617*.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Citation: A key to building responsible and accountable large language models. *arXiv preprint arXiv:2307.02185*.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

Masahiro Kaneko and Danushka Bollegala. 2021. Dictionary-based debiasing of pre-trained word embeddings. *ArXiv*, abs/2101.09525.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3206–3219, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*.

Kishan Maharaj, Ashita Saxena, Raja Kumar, Abhijit Mishra, and Pushpak Bhattacharyya. 2023. Eyes show the way: Modelling gaze behaviour for hallucination detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11424–11438, Singapore. Association for Computational Linguistics.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Marianna Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee. 2019. Identifying fluently inadequate output in neural and statistical machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 233–243.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li, Pavan Kapanipathi, and Kartik Talamadupula. 2021. Looking beyond sentence-level natural language inference for question answering and text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1336.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained 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)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. Bbq: A hand-built bias benchmark for question answering.

Clément Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo

Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. 2021. Data-questeval: A referenceless metric for data-to-text semantic evaluation. *arXiv preprint arXiv:2104.07555*.

Nihar Sahoo, Himanshu Gupta, and Pushpak Bhattacharyya. 2022. Detecting unintended social bias in toxic language datasets. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 132–143, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nihar Sahoo, Niteesh Mallela, and Pushpak Bhattacharyya. 2023. With prejudice to none: A few-shot, multilingual transfer learning approach to detect social bias in low resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13316–13330, Toronto, Canada. Association for Computational Linguistics.

Nihar Ranja Sahoo, Gyana Prakash Beria, and Pushpak Bhattacharyya. 2024a. IndicCONAN: A multilingual dataset for combating hate speech in indian context. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22313–22321.

Nihar Ranjan Sahoo, Pranamya Prashant Kulkarni, Narjis Asad, Arif Ahmad, Tanu Goyal, Aparna Garimella, and Pushpak Bhattacharyya. 2024b. IndiBias: A benchmark dataset to measure social biases in language models for indian context.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp.

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Questeval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Sandhya Singh, Prapti Roy, Nihar Sahoo, Niteesh Mallela, Himanshu Gupta, Pushpak Bhattacharyya, Milind Savagaonkar, Nidhi, Roshni Ramnani, Anutosh Maitra, and Shubhashis Sengupta. 2022. Hollywood identity bias dataset: A context oriented bias analysis of movie dialogues.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations.

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P Parikh. 2019. Sticking to the facts: Confident decoding for faithful data-to-text generation. *arXiv preprint arXiv:1910.08684*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020a. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Yuhang Wang, Dongyuan Lu, Chao Kong, and Jitao Sang. 2023. Towards alleviating the object bias in prompt tuning-based factual knowledge extraction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4420–4432, Toronto, Canada. Association for Computational Linguistics.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020b. Towards faithful neural table-to-text generation with content-matching constraints. *arXiv preprint arXiv:2005.00969*.

Pengfei Yu and Heng Ji. 2023. Self information update for large language models through mitigating exposure bias. *arXiv preprint arXiv:2305.18582*.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*.