Tackling Bias in Large Language Models

Vithya Yogarajan and Gillian Dobbie

School of Computer Science, University of Auckland, New Zealand
{vithya.yogarajan, g.dobbie}@auckland.ac.nz

1 Abstract

Large language models (LLMs) are powerful decision-making tools widely adopted in healthcare, finance, and transportation. Embracing the opportunities and innovations of LLMs is inevitable. However, LLMs inherit stereotypes, misrepresentations, discrimination, and societies' biases from various sources resulting in concerns about equality, diversity, and fairness.

The tutorial provides an overview of bias in LLMs—what it is, how it is detected and measured, and methods for mitigating bias. It incorporates real-world examples from New Zealand, where Māori are the indigenous population and underrepresented. After describing bias and its sources in the LLM development pipelines, the tutorial delves into current methods for detecting bias and the evaluation metrics recently introduced for bias measurement. It covers the state of the art in mitigating bias in LLMs. Since the area is in its infancy, the tutorial concludes with many open research questions. The examples provide participants the opportunity to delve into the methods that are introduced through hands-on exercises.

2 Description and Outline

The launch of OpenAI's ChatGPT in November 2022 [12] is potentially the most significant milestone in the advances of large language models (LLMs). It is reported that ChatGPT gained over 100 million users within the first two months of release. The underlying technology of such LLMs is the key to innovations, and there are examples of LLMs exhibiting remarkable capabilities across high-stakes decision applications in healthcare, criminal justice and finance[18, 5, 11]. However, introducing and using LLMs comes with biases and disparities, resulting in concerns about equity [13, 9, 16, 10]. For example, [6] found 83% of the occupation prompts generated text using GPT-3 with male identifiers, and [1] show GPT-3's output has a higher violent bias against Muslims than other religious groups.

Bias can be defined as the disparate treatment or outcomes between social groups that arise from historical and structural power imbalances [2, 4, 7], and can be related to gender, social status, race, language, disability, and more. This can incorporate representational harms such as misrepresentation, stereotyping, disparate system performance, and direct and indirect discrimination [2, 4, 7]. LLMs inherit stereotypes and misrepresentations of societies from the training data [3, 17], and can also amplify these biases [8, 1]. In addition to the training data, sources of bias can arise from various stages of the machine-learning pipeline, including data collection, algorithm design, and user interactions.

We begin the tutorial by providing an overview of LLMs, defining and explaining bias, and discussing the ongoing research in this space. Throughout the tutorial, we provide examples of scenarios for Aotearoa New Zealand (NZ) [16, 15, 14]. We then provide an overview of bias metrics and bias benchmark datasets, with hands-on examples, followed by debiasing techniques. The effectiveness of mitigating bias in LLMs will depend on the debiasing techniques used. It can be measured by considering the relative change in the bias of LLMs before and after applying the method [14]. An outline of the tutorial with estimated durations is provided below:

1. Introduction to LLMs and bias [Duration: 15 mins]

    (a) Overview of Large Language Models (LLMs)

    (b) Overview and definition of bias

      - Bias definition and its types
      - Sources of bias in LLM development pipelines
      - Downstream task and impact

    (c) Current research trends

    (d) Interactive/hands-on Examples (NZ)

2. Bias detection

    (a) Bias metrics [Duration: 25 mins]

      i. Overview of bias metrics

        - Holistic Evaluation of Language Models (HELM) bias score
        - Toxicity
        - Regard score
        - Honest score

      ii. Interactive/hands-on Examples (NZ)

    (b) Benchmark Datasets [Duration: 15 mins]

    BREAK [Duration: 10 mins]

3. Bias mitigation [Duration: 40 mins]

    (a) Overview of mitigation techniques

      - Data-related
      - Prompt-based

- In-training
- Intra-processing
- Guardrails

(b) Interactive/hands-on Examples

4. Open Research Avenues and closing remarks [Duration: 10 mins]

3 Target Audience

The target audiences include peer researchers, students, policymakers, and industry practitioners who work with or plan to use LLMs for research or applications. This tutorial will help them understand the techniques for detecting and mitigating bias in LLMs. We will also provide examples and applications with code walk-throughs for a hands-on experience of these techniques.

4 Previous tutorials

This is the first time we are presenting this tutorial.

5 Links to video recordings

Here's a couple of recent presentations:

1. Keynote at New Zealand Software Engineering Conference 2023 `https://www.youtube.com/watch?v=mREkhz6_HRs`

2. Data to Machine Learning, Royal Society NZ 2023 `https://www.youtube.com/watch?v=t8DvIjGg9-M`

6 Organisers

Vithya Yogarajan is a Research Fellow at the School of Computer Science at the University of Auckland. Vithya's primary research interest lies at the intersection of Artificial Intelligence (AI) and mitigating bias, with a special focus on applications of AI in the health sector.

Gillian Dobbie is a Professor at the School of Computer Science at the University of Auckland. She leads the school's flagship Ethical Computing project. Gillian's work ranges from theory to application, from proving algorithms' correctness to experimental computer science. She has published more than 130 peer-reviewed research papers and served as a reviewer for numerous highly regarded conferences and journals.

Vithya and Gillian have collaborated on mitigating bias and developing bias benchmark datasets. They have published several papers together in this emerging research area.

References

[1] Abid, A., Farooqi, M., Zou, J.: Persistent anti-muslim bias in large language models. In: AAAI/ACM Conference on AI, Ethics, and Society. pp. 298–306 (2021)

[2] Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning: Limitations and Opportunities. MIT Press (2023)

[3] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: ACM FAccT. pp. 610–623 (2021)

[4] Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of "bias" in NLP. In: ACL. pp. 5454–5476 (2020)

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

[6] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., et al.: Language models are few-shot learners. NeurIPS **33**, 1877–1901 (2020)

[7] Crawford, K.: The trouble with bias, Keynote at NeurIPS (2017)

[8] Crutchley, M.: Book review: Race after technology: Abolitionist tools for the New Jim Code (2021)

[9] Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., et al.: Racial disparities in automated speech recognition. National Academy of Sciences **117**(14), 7684–7689 (2020)

[10] Liang, P.P., Wu, C., Morency, L.P., Salakhutdinov, R.: Towards understanding and mitigating social biases in language models. In: ICML. pp. 6565–6576. PMLR (2021)

[11] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)

[12] Chatgpt: Optimizing language models for dialogue (2022)

[13] Thiago, D.O., Marcelo, A.D., Gomes, A.: Fighting hate speech, silencing drag queens? AI in content moderation and risks to LGBTQ voices online. Sexuality & culture **25**(2), 700–732 (2021)

[14] Yogarajan, V., Dobbie, G., Keegan, T.T., Neuwirth, R.J.: Tackling bias in pre-trained language models: Current trends and under-represented societies. arXiv preprint arXiv:2312.01509 (2023)

[15] Yogarajan, V., Dobbie, G., Leitch, S., Reith, D.: Developing a fair AI-based healthcare framework with feedback loop. In: KDH-IJCAI (2023)

[16] Yogarajan, V., Dobbie, G., Pistotti, T., Bensemann, J., Knowles, K.: Challenges in annotating datasets to quantify bias in under-represented society. In: EthAIcs-IJCAI. pp. 1–15 (2023)

[17] Yogarajan, V., Dobbie, G., Gouk, H.: Effectiveness of debiasing techniques: An indigenous qualitative analysis. In: ICLR TinyPapers. pp. 1–5 (2023)

[18] Yogarajan, V., Montiel, J., Smith, T., Pfahringer, B.: Transformers for multi-label classification of medical text: an empirical comparison. In: AIME. pp. 114–123. Springer (2021)