Cianna Grama
Professor Moher
DAT 201: Thematic Inquiry
17 February 2025

Examining The Implications Of Ethical And Social Bias In LLMs

As large language models (LLMs) have become increasingly used for decision-making in various sectors, understanding the bias in these models is of increasing importance. LLMs are types of AI programs that can recognize and generate text and can do so by being trained on huge sets of data (Cloudflare). The data LLMs are trained on have biases, so the models reflect and perpetuate those biases, and in some cases amplify them because they are written to respond confidently. The issue of bias in these models is heightened by the increasing usage of these models in all sectors, specifically in industry. According to REF, 77% of companies use or explore AI in their business, and 83% claim that AI is a top priority in their plans (Prestianni). Given the growing reliance on LLMs in decision-making, it is critical to address the biases these models inherit from their training data and tackle this issue.

One aspect of bias in LLMs is perpetuated through text generation. “Dataset and Metrics for Measuring Biases in Open-Ended Language Generation” by Dhamala, et.al examines bias in LLM text generation. This study compared Wikipedia text to text generated by LLM from neutral prompts and assigned a level of bias using VALER (Valence Aware Dictionary and Sentiment Reasoner). Bias was assessed in the following 5 domains: profession, gender, race, religion, and political ideology. The study found that the “majority of these models exhibit a larger social bias than human-written Wikipedia text across all domains” (Dhamala et, al). Bias existed in all models, with the strongest in GPT-2, CTRL-OPN, and CTRL-THT. Some specific findings were: bias was found against African Americans, Islam and Atheism-related prompts produced more negative content than Christianity, fascism prompts generated surprisingly positive outcomes in some cases, and gender stereotypes existed in professions. The results of this study emphasize the need to address, make known to users, and reduce biases in LLM reasoning. These biases are particularly concerning given the increasing reliance on LLMs in industries where decisions that impact people’s lives, such as hiring, are informed by these tools.

Another aspect of bias in LLMs is ethical bias. “The moral machine experiment on large language models” by Kazuhiro Takemoto studied the ethical decision-making of LLM models

ChatGPT, PaLM2, and Llama2. The study generated scenarios exploring preferences on whose lives to save based on 6 attributes: species (humans over pets), social value, gender, age, fitness, and utilitarianism (choosing over one group and another larger group). The results were analyzed by calculating the average marginal component effect (AMCE) for each attribute. The results found some consistent trends within the models, such as saving human lives over pet lives, which aligns with human preference. Another consistent trend was sparing less fit people over fit ones, which is inconsistent with human preference according to the study. The different models had nuanced differences among themselves. For example, GPT models fell more in line with human responses compared to PaLM2 and Llama2. Overall, the models had variations in their responses and were not consistently in line with human ethics. Divulgence from human ethics highlights the importance of developing this area of LLMs, especially in cases where AI may be required to decide between one life or another. One example of this is with self-driving cars. As they develop and grow on the market, external circumstances may require them to actively decide on whose life to preserve.

Both of the studies reinforce the idea that LLMs are not neutral tools. LLMs carry the biases of their training data, which is made even stronger by their programmed emphasis on confidence in their responses. The convergence of social and ethical biases in LLMs emphasizes the need for more balanced training data to reduce biases, more transparency in training data, and public awareness of the biases these tools hold. The biases of these models have the potential to shape the decisions of their users, which can potentially harm those whom the models are biased against. Therefore, it is essential that these models are improved and that regulations are put in place to ensure that the models in use are rid of harmful biases.

When determining how to fix the bias in these models, we must examine and alter the data that these models are being trained on, as the training data has the majority share of the biases in these models. Future research should explore methods for de-biasing training data or creating synthetic datasets that better reflect diverse and equitable perspectives. Additionally, users need transparency about the data LLMs are trained on to enable better evaluation and adjustments. However, it is important to note that training data is human-generated. Humans are inherently biased, and virtually everything they produce reflects this bias. So, how can we train LLMs not to have a bias if the only training data available is biased? It seems as though the answer to removing bias in LLMs is removing bias in the human data that trains them. As

impossible as that may seem, we must determine how to remove these biases in LLMs as humans grow increasingly dependent on them.

References

Cloudflare. What Is an LLM (Large Language Model)? , www.cloudflare.com/learning/ai/what-is-large-language-model/. Accessed 18 Feb. 2025.

Dhamala, Jwala, et al. “BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation.” Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Mar. 2021, pp. 862–872, https://doi.org/10.1145/3442188.3445924.

Prestianni, Timothy. “131 AI Statistics and Trends for (2024).” National University, 14 Jan. 2025, www.nu.edu/blog/ai-statistics-trends/#:~:text=According%20to%20research%20completed%20by,priority%20in%20their%20business%20plans.

Takemoto, Kazuhiro. “The Moral Machine Experiment on large language models.” Royal Society Open Science, vol. 11, no. 2, 7 Feb. 2024, https://doi.org/10.1098/rsos.231393.