Laissez-Faire Harms: Algorithmic Biases in Generative Language Models (Extended Abstract)

Evan Shieh[1], **Faye-Marie Vassel**[2], **Cassidy R. Sugimoto**[3], **Thema Monroe-White**[4]
[1]Young Data Scientists League
[2]Stanford University
[3]School of Public Policy, Georgia Institute of Technology
[4]Schar School of Policy and Government & School of Computing, George Mason University
evan.shieh@youngdatascientists.org, tmonroew@gmu.edu

Extended version — https://arxiv.org/pdf/2404.07475
Datasets — https://doi.org/10.7910/DVN/WF8PJD

The widespread deployment of generative language models (LMs) raises concerns about societal harms. Despite this, studies of bias in LMs, including attempted self-audits by LM developers, have thus far been conducted in limited contexts (Shieh and Monroe-White 2025). To address this gap, this study examines representational harms in synthetic texts produced by leading language models in response to open-ended creative writing prompts based in the United States.

We conduct our investigation on 500,000 synthetic texts generated by five publicly available generative language models: ChatGPT 3.5 and ChatGPT 4 (developed by Open AI), Llama 2 (Meta), PaLM 2 (Google), and Claude 2.0 (Anthropic). We base our selection of models on both the sizable amount of funding wielded by these companies and their investors (on the order of tens of billions in USD), as well as the prominent policy roles that each company has played on the federal level (Grffith and Metz 2023). At the time of data collection (from August 16th to November 7th, 2023), the selected models were state-of-the-art for each company.

Creative writing prompts reflect three domains of life set in the United States: classroom interactions ("Learning"), the workplace ("Labor"), and interpersonal relationships ("Love"). Informed by intersectionality, we considered the role of power embedded in language by creating one power-neutral scenario and one power-laden scenario for each prompt. For example, power-neutral Learning prompts consist of a single student excelling in an academic subject, whereas the power-laden prompts consist of one star student helping a struggling student in an academic subject.

We then analyze the resulting model responses for textual cues shown to exacerbate socio-psychological harms for minoritized individuals by race, gender, and sexual orientation. To do this at scale, we fine-tuned a coreference resolution model (gpt3.5-turbo) to perform automated extraction of characters' gender references and names at high precision. To evaluate our model, we hand-label inferred gender (based on gender references) and name on an evaluation set of 4,600 uniformly down-sampled story generations from all five models (.0063 95CI). Fine-tuning our model on a non-overlapping set of 150 training examples yields precision above 98% for both gender references and names. Recall reaches 97% for gender references and above 99% for names. Following prior studies, we infer racial signals from first names using fractionalized counting over the Florida Voter Registration Dataset (Sood 2022), which consists of 27 million named voters and self-reported racial identities.

We find that when LMs are used for story writing, they generate texts that reinforce discrimination against minoritized groups by race, gender, and sexual orientation. Using mixed-methods analyses, we identify three specific harms: omission, subordination, and stereotyping.

Stories produced by language models simultaneously under-represent minoritized individuals as main characters while over-representing them as subordinated characters. Diverse consumers, if they are to be represented at all, disproportionately see themselves portrayed by language models as "struggling students" (as opposed to "star students"), "patients" or "defendants" (as opposed to "doctors" or "lawyers"), and a friend or romantic partner who is more likely to borrow money or do the chores for someone else.

The magnitude of bias far exceeds the level of "real-world" inequities. Underrepresentation of non-dominant identities in power-neutral stories exceeds national demographics in the US by up to two orders of magnitude. Meanwhile, non-dominant character identities are up to thousands of times more likely to appear as subordinated than empowered. For example, Claude casts the name "Juan" as a struggling student 1,380 times, yet only once as a star student.

We find that these harms impact every non-dominant group we studied (in the US context). These include individuals with intersectional Asian, Black, Indigenous, Latine, NH/PI, MENA, Female, Non-binary, and Queer identities. Language models propagate a plethora of stereotypes that are known to inflict psychological harm and negative self-perception, including the "glass / bamboo ceiling", "perpetual foreigner", "noble savage", "white savior", and others.

Acknowledgments

We thank Diego Kozlowski, Stella Chen, Rahul Gupta-Iwasaki, Gerald Higginbotham, Bryan Brown, Stephanie Melville, Jay Kim, Dakota Murray, James Evans, Zarek Drozda, Ashley Ding, Princewill Okoroafor, and Hideo Mabuchi for helpful inputs and discussion.

References

Grffith, E.; and Metz, C. 2023. A New Area of A.I. Booms, Even Amid the Tech Gloom. https://www.nytimes.com/2023/01/07/technology/generative-ai-chatgpt-investments.html. Accessed: 2025-08-23.

Shieh, E.; and Monroe-White, T. 2025. Teaching Parrots to See Red: Self-Audits of Generative Language Models Overlook Sociotechnical Harms. In *Proceedings of the AAAI Symposium Series*, volume 6, 333–340.

Sood, G. 2022. Florida Voter Registration Data (2017 and 2022). doi:10.7910/DVN/UBIG3F. Accessed: 2025-08-23.