Mental Health Stigma across Diverse Genders in Large Language Models - Abstract

Lucille Njoo[1,*,†], Lee Janzen-Morel[1,*,†], Inna Wanyin Lin[1] and Yulia Tsvetkov[1]

[1]*Paul G. Allen School of Computer Science, University of Washington, 185 E Stevens Way NE, Seattle, WA 98195. United States.*

Abstract

Mental health stigma manifests differently for different genders, often being more associated with women and overlooked with men. Prior work in NLP has shown that gendered mental health stigmas are captured in large language models (LLMs). However, in the last year, LLMs have changed drastically: newer, generative models not only require different methods for measuring bias, but they also have become widely popular in society, interacting with millions of users and increasing the stakes of perpetuating gendered mental health stereotypes. In this paper, we examine gendered mental health stigma in GPT3.5-Turbo, the model that powers OpenAI's popular ChatGPT. Building off of prior work, we conduct both quantitative and qualitative analyses to measure GPT3.5-Turbo's bias between binary genders, as well as to explore its behavior around non-binary genders, in conversations about mental health. We find that, though GPT3.5-Turbo refrains from explicitly assuming gender, it still contains implicit gender biases when asked to complete sentences about mental health, consistently preferring female names over male names. Additionally, though GPT3.5-Turbo shows awareness of the nuances of non-binary people's experiences, it often over-fixates on non-binary gender identities in free-response prompts. Our preliminary results demonstrate that while modern generative LLMs contain safeguards against blatant gender biases and have progressed in their inclusiveness of non-binary identities, they still implicitly encode gendered mental health stigma, and thus risk perpetuating harmful stereotypes in mental health contexts.

Keywords

NLP, large language models, bias, fairness, gender, mental health, stigma, intersectionality

Machine Learning for Cognitive and Mental Health Workshop (ML4CMH), AAAI 2024, Vancouver, BC, Canada

*Corresponding authors.

†Authors contributed equally.

lnjoo@cs.washington.edu (L. Njoo); ljanzen@cs.washington.edu (L. Janzen-Morel)