Coming out to a chatbot? Researchers explore the limitations of mental health chatbots in LGBTQ+ communities

May 13 2024

Credit: Pixabay/CC0 Public Domain

Today, there are dozens of large language model (LLM) chatbots aimed at mental health care—addressing everything from loneliness among seniors to anxiety and depression in teens.

But the efficacy of these apps is unclear. Even more unclear is how well these apps work in supporting specific, marginalized groups like LGBTQ+ communities.

A team of researchers from the Harvard John A. Paulson School of Engineering and Applied Sciences, Emory University, Vanderbilt University and the University of California Irvine, found that while large language models can offer fast, on-demand support, they frequently fail to grasp the specific challenges that many members of the LGBTQ+ community face.

That failure could lead the chatbot to give at best unhelpful and at worst dangerous advice.

The paper is being presented this week at the ACM (Association of Computing Machinery) conference on [Human Factors in Computing System](#) in Honolulu, Hawai'i.

The researchers interviewed 31 participants—18 identifying as LGBTQ+ and 13 as non-LGBTQ+—about their usage of LLM-based chatbots for mental health support and how the chatbots supported their individual needs.

On one hand, many participants reported that the chatbots offered a sense of solidarity and a safe space to explore and express their identities. Some used the chatbots for practice coming out to friends and family, others to practice asking someone out for the first time.

But many of the participants also noted the programs' shortfalls.

One participant wrote, "I don't think I remember any time that it gave me a solution. It will just be like empathetic. Or maybe, if I would tell it that I'm upset with someone being homophobic. It will suggest, maybe

talking to that person. But most of the time it just be like, "I'm sorry that happened to you.'"

"The boilerplate nature of the chatbots' responses highlights their failure to recognize the complex and nuanced LGBTQ+ identities and experiences, making the chatbots' suggestions feel emotionally disengaged," said Zilin Ma, a Ph.D. student at SEAS and co-first author of the paper.

Because these chatbots tend to be sycophantic, said Ma, they're actually very bad at simulating hostility, which makes them ill-suited to practice potentially fraught conversations like coming out.

They also gave some participants staggeringly bad advice—telling one person to quit their job after experiencing workplace homophobia, without considering their financial or personal consequences.

Ma, who is in the lab of Krzysztof Gajos, the Gordon McKay Professor of Computer Science, stressed that while there are ways to improve these programs, it is not a panacea.

"There are ways we could improve these limitations by fine tuning the LLMs for contexts relevant to LGBTQ+ users or implementing context-sensitive guardrails or regularly updating feedback loops, but we wonder if this tendency to implement technology at every aspect of social problem is the right approach," said Ma.

"We can optimize all these LLMs all we want but there are aspects of LGBTQ+ mental health that cannot be solved with LLM chatbots—such as discrimination, bullying, the stress of coming out or the lack of representation. For that, we need a holistic support system for LGBTQ+ people."

One area where LLM chatbots could be useful is in the training of human counselors or online community moderators.

"Rather than having teens in crisis talk to the chatbot directly, you could use the chatbot to train counselors," said Ma. "Then you have a real human to talk to, but it empowers the counselors with technology, which is a socio-technical solution which I think works well in this case."

"Research in public health suggests that interventions that directly target the affected individuals—like the chatbots for improving individual well-being—risk leaving the most vulnerable people behind," said Gajos. "It is harder but potentially more impactful to change the communities themselves through training counselors or online community moderators."

The research was co-authored by Yiyang Mei, Yinru Long, Zhaoyuan "Nick" Su and Gajos.

Provided by Harvard John A. Paulson School of Engineering and Applied Sciences