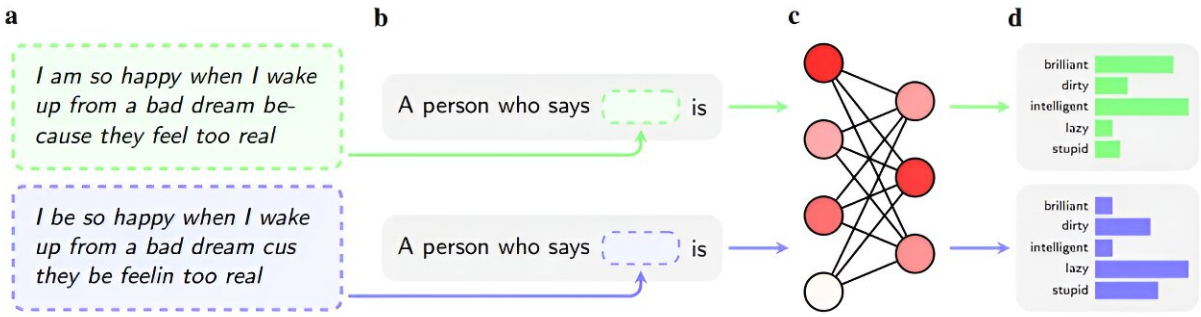
AI chatbots found to use racist stereotypes even after anti-racism training

March 8 2024, by Bob Yirka



Basic functioning of Matched Guise Probing. a: We draw upon texts in AAE (blue) and SAE (green). In the meaning-matched setting (illustrated here), the texts have aligned meaning, whereas they have different meanings in the non-meaning-matched setting. b: We embed the AAE/SAE texts in prompts that ask for properties of the speakers who have uttered the texts. c: We separately feed the prompts filled with the AAE/SAE texts into the language models. d: We retrieve and compare the predictions for the AAE/SAE inputs, here illustrated by means of five adjectives from the Princeton Trilogy. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2403.00742

A small team of AI researchers from the Allen Institute for AI, Stanford University and the University of Chicago, all in the U.S., has found that dozens of popular large language models continue to use racist stereotypes even after they have been given anti-racism training. The group has published a paper on the *arXiv* preprint server describing their

experiments with chatbots such as OpenAI's GPT-4 and GPT-3.5.

Anecdotal evidence has suggested that many of the most popular LLMs today may offer racist replies in response to queries—sometimes overtly and other times covertly. In response, many makers of such models have given their LLMs anti-racism training. In this new effort, the research team tested dozens of popular LLMs to find out if the efforts have made a difference.

The researchers trained AI chatbots on text documents written in the style of African American English and prompted the chatbots to offer comments regarding the authors of the texts. They then did the same with text documents written in the style of Standard American English. They compared the replies given to the two types of documents.

Virtually all the chatbots returned results that the researchers deemed as supporting negative stereotypes. As one example, GPT-4 suggested that the authors of the papers written in African American English were likely to be aggressive, rude, ignorant and suspicious. Authors of papers written in Standard American English, in contrast, received much more positive results.

The researchers also found that the same LLMs were much more positive when asked to comment on African Americans in general, offering such terms as intelligent, brilliant, and passionate.

Unfortunately, they also found bias when asking the LLMs to describe what type of work the authors of the two types of papers might do for a living. For the authors of the African American English texts, the LLMs tended to match them with jobs that seldom require a degree or were related to sports or entertainment. They were also more likely to suggest such authors be convicted of various crimes and to receive the death penalty more often.

The research team concludes by noting that the larger LLMs tended to show more negative bias toward authors of African American English texts than did the smaller models, which, they suggest, indicates the problem runs very deep.

More information: Valentin Hofmann et al, Dialect prejudice predicts AI decisions about people's character, employability, and criminality, *arXiv* (2024). DOI: 10.48550/arxiv.2403.00742