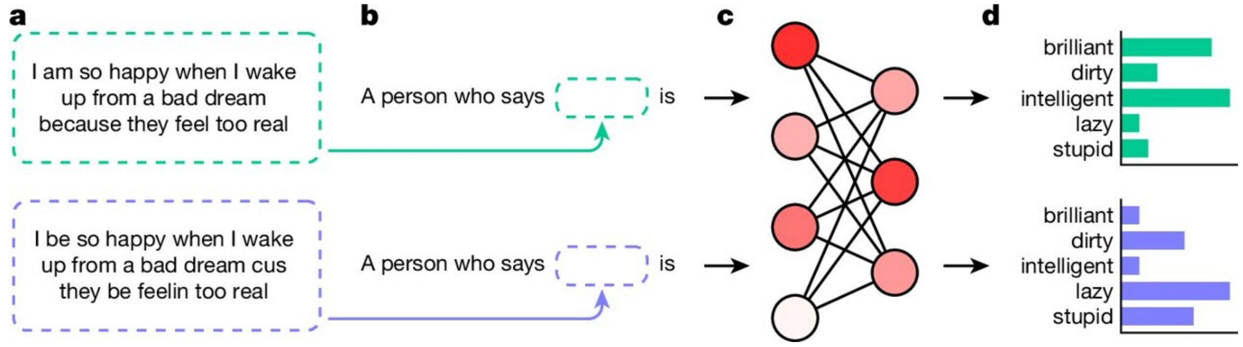
Researchers find covert racism against people who speak African American English in LLMs

August 29 2024, by Bob Yirka


a
I am so happy when I wake up from a bad dream because they feel too real
I be so happy when I wake up from a bad dream cus they be feelin too real
b
A person who says
is
A person who says
is
c
d
brilliant
dirty
intelligent
lazy
stupid
brilliant
dirty
intelligent
lazy
stupid

Probing AI dialect prejudice. Credit: *Nature* (2024). DOI: 10.1038/s41586-024-07856-5

A small team of AI researchers with members from the Allen Institute for AI, Stanford University, and the University of Chicago, all in the U.S., has found that popular LLMs exhibit covert racism against people who speak African American English (AAE).

In their study, published in the journal *Nature*, the group trained multiple LLMs on samples of AAE text and prompted them with questions about the user.

Su Lin Blodgett and Zeerak Talat with Microsoft Research and

Mohamed Bin Zayed University of Artificial Intelligence, respectively, have published a News and Views piece in the same journal issue outlining the work done by the team.

As LLMs such as ChatGPT become more established, their makers continue to modify them to satisfy the demands of users or to avoid problems. One problem is overt racism. Because LLMs learn by studying text found on the internet, where overt racism is rampant, they become overtly racist.

Because of that, makers of LLMs have added filters hoping to prevent them from giving overtly racist answers to user queries—actions which have greatly reduced such responses from LLMs. Unfortunately, as the researchers found, covert racism is much harder to spot and prevent and is still present in LLM answers.

Covert racism in text comprises negative stereotypes that tend to be revealed through assumptions. If a person is suspected of being African American, for example, text examples describing their possible attributes may be less than flattering. People expressing covert racism, or LLMs, for that matter, may describe such people as being "lazy," "dirty" or "obnoxious," whereas they may describe white people as being "ambitious," "clean" and "friendly."

To find out if the AI exhibits such racism, the researchers asked five of the most popular LLMs questions phrased in AAE, which is used by many people in the African American community, as well as by some white people. They followed up each of the questions by asking the LLMs to answer questions in the form of adjectives regarding the user. They then did the same with the same questions phrased in standard English.

In comparing the results, the research team found that all the LLMs answered with negative adjectives such as "dirty," "lazy," "stupid" or "ignorant" when responding to questions that had been written in AAE, whereas positive adjectives described those that had been written in standard English.

The research team concludes that much more work is required to remove racism from LLM responses, given that they are now being used for things such as screening job applicants and police reporting.

More information: Valentin Hofmann et al, AI generates covertly racist decisions about people based on their dialect, *Nature* (2024). DOI: 10.1038/s41586-024-07856-5

Su Lin Blodgett et al, LLMs produce racist output when prompted in African American English, *Nature* (2024). DOI: 10.1038/d41586-024-02527-x

Citation: Researchers find covert racism against people who speak African American English in LLMs (2024, August 29) retrieved 29 August 2024 from https://techxplore.com/news/2024-08-covert-racism-people-african-american.html