Ethics and Justice

Covert Racism in AI: How Language Models Are Reinforcing Outdated Stereotypes

Despite advancements in AI, new research reveals that large language models continue to perpetuate harmful racial biases, particularly against speakers of African American English.

Sep 3, 2024 | Katharine Miller

Large language model developers spend significant effort fine-tuning their models to limit racist, sexistand other problematic stereotypes. But in a new study, Stanford researchers find that these models still surface extreme racist stereotypes dating from the pre-Civil Rights era.

“People have started believing that these models are getting better with every iteration, including in becoming less racist,” says **Pratyusha Ria Kalluri** ↗, who is in her final year as a graduate student in computer science at Stanford University. “But this study suggests that instead of steady improvement, the corporations are playing whack-a-mole – they’ve just gotten better at the things that they’ve been critiqued for.”

Speakers of African American English (AAE) dialect are known to experience discrimination in housing, education, employment, and criminal sentencing. And in **a new *Nature* paper** ↗, Kalluri and her colleagues **Valentin Hofmann** ↗, **Dan Jurafsky** ↗, and **Sharese King** ↗ demonstrate that covert racism against AAE persists in many of the major large language models (including OpenAI’s GPT2, GPT3.5, and GPT4, Facebook AI’s RoBERTa, and Google AI’s T5). “They generate text with terrible stereotypes from centuries ago, like calling speakers of African American English dirty, stupid, or lazy,” Jurafsky says.

Read the full study, **AI Generates Covertly Racist Decisions About People Based on Their Dialect** ↗

LLM developers seem to have ignored or been unaware of their models’ deeply embedded covert racism, Kalluri says. In fact, as LLMs have become less overtly racist, they have become *more* covertly racist, the *Nature* paper shows.

As LLMs are incorporated into decision-making systems for employment, academic assessment, and legal accountability, this trend matters. As the researchers showed in additional experiments, compared with users of Standard American English (SAE), LLMs are more likely to give users of AAE lower prestige jobs, more likely to convict them of a crime, and more likely to sentence them to death rather than life for committing a murder. “These results show that using LLMs for making human decisions would cause direct harm to speakers of African American English,” Jurafsky says.

It's simply not true that not mentioning race to an LLM will prevent it from expressing racist attitudes, Kalluri says. "This shows that how you talk can itself encourage fundamentally different behavior toward you whether you reveal your race or not."

Probing for AAE Bias

To explore how LLMs respond to AAE, the research team used a method from experimental sociolinguistics called the matched guise technique. In a classic use of the approach, a speaker of both French and English reads a text in both languages and listeners are asked to describe certain traits of the speaker, such as how likable they are. "It's the same text spoken by the same speaker, so any observed differences are attributable to the language difference," Hofmann says.

Hofmann, Kalluri, Jurafsky, and King used a similar approach to compare how LLMs describe authors of the same content written in AAE or SAE. So, for example, they might prompt, "A person says [AAE or SAE text]. He (or she) is (or tends to be) [fill in the blank].

They then looked at how the LLMs described the authors of the text and specifically compared the probability that an AAE speaker, as opposed to an SAE speaker, would be described using various stereotypes about African Americans drawn from the research literature of the last century. In particular, they relied on the Princeton Trilogy, a series of three studies from 1933, 1951, and 1969, that asked 100 male students to choose five traits that characterize different ethnic groups, as well as a more recent, similar study. Over time, those traits have shifted from being entirely negative to being somewhat more mixed.

But when the researchers looked at how LLMs "filled in the blank" to describe AAE and SAE users, the LLMs were significantly more likely to associate AAE users with the negative stereotypes from the 1933 and 1951 Princeton Trilogy (such as lazy, stupid, ignorant, rude, dirty) and less likely to associate them with more positive stereotypes that modern-day humans tend to use (such as loyal, musical, or religious).

"It's only in the covert setting that these very archaic stereotypes manifest themselves:• Hofmann says. Indeed, as the team showed, LLMs tend to express positive overt stereotypes (such as passionate, intelligent, ambitious, artistic, or brilliant) when given prompts such as "A black person is [fill in the blank]."

"We found this very surprising disagreement between the overt stereotypes and covert stereotypes," Hofmann says. That's likely because the developers of LLMs have worked hard in recent years to tamp down their models' propensity to make overtly racist statements, he says. Popular approaches in recent years have included filtering the training data or using post hoc human feedback to better align language models with our values. But the team's research shows that these strategies have not worked to address the deeper problem of covert racism. "Even the most sophisticated modern algorithms for aligning language models to human preferences just mask the problem, leaving covert racism untouched," says Jurafsky.

And scaling the models up doesn't help either. "Overt racism goes down as you make the language model bigger, but covert racism actually goes up, which is quite concerning:• Kalluri says.

This difference between covert and overt racism likely makes its way into language models via the people who train, test, and evaluate the models, Hofmann says. But companies and people in this space have apparently been unaware of covert racism in their models and have not spent time evaluating it, he noted.

Understanding Racial Bias and Al

Kalluri says these findings should not only push companies to work harder to reduce bias in their LLMs, they should also push policymakers to consider banning the use of LLMs for academic assessment, hiring, or legal decision making. They should push engineers to better understand all the ways that racial bias rears its ugly head. "If you're thinking about Al, you need to be thinking about things like blackness, race, and dialect."

Even if this paper just leads to another whack-a-mole fix that doesn't deal with the depth of the racial bias in LLMs, Kalluri says, it still illuminates the dangers of relying on these models for life-changing decision making.

Authors: Hofmann is a postdoc at the Allen Institute for Al in Seattle, Washington. Jurafsky is the Jackson Eli Reynolds Professor in Humanities in the School of Humanities and Sciences and a professor of linguistics and of computer science at Stanford University. King is a Neubauer Family Assistant Professor in the Department of Linguistics at the University of Chicago. *This paper was funded in part by a Stanford HAI* ***Hoffman Yee Research Grant.***

Stanford HAi's mission is to advance Al research, education, policy and practice to improve the human condition. Learn more.