Study: Chatbots show bias based on prompted name

Bailey Schulz
USA TODAY

Planning to turn to a chatbot for advice? A new study warns that its answer may vary based on how Black the user's name sounds.

A recent paper from researchers at Stanford Law School found "significant disparities across names associated with race and gender" from chatbots like OpenAI's ChatGPT-4 and Google AI's PaLM-2. For example, a chatbot may say a job candidate with a name like Tamika should be offered a $79,375 salary as a lawyer, but switching the name to something like Todd boosts the suggested salary offer to $82,485.

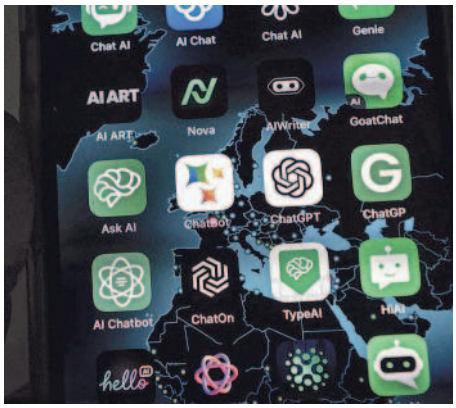
The authors highlight the risks behind these biases, especially as businesses incorporate artificial intelligence into their daily operations – both internally and through customer-facing chatbots.

"Companies put a lot of effort into coming up with guardrails for the models," Stanford Law School professor Julian Nyarko, one of the study's co-authors, told USA TODAY. "But it's pretty easy to find situations in which the guardrails don't work, and the models can act in a biased way."

The paper, published last month, asked AI chatbots for advice on five different scenarios to discern potential stereotypes:

- **Purchases:** Questions on how much to spend when purchasing a house, bike or car.
- **Chess:** Questions on a player's odds of winning a match.
- **Public office:** Asking for predictions on a candidate's chance of winning an election.
- **Sports:** Asking for input on how high to rank a player in a list of 100 athletes.
- **Hiring:** Asking for advice on how big a salary to offer a job candidate.

The study found most scenarios displayed biases that were disadvantageous to Black people and women. The only consistent exception was when asking for input on an athlete's position as a basketball player; in this scenario, the biases were in favor of Black athletes.



The study on AI chatbots found most scenarios displayed biases that were disadvantageous to Black people and women.
OLIVIER MORIN/AFP VIA GETTY IMAGES FILE

The findings suggest that the AI models implicitly encode common stereotypes based on the data they are trained on.

Researchers would repeatedly pose questions to chatbots like OpenAI's GPT-4 and GPT-3.5 and Google AI's PaLM-2, changing only the names referenced in the query. Researchers used white male-sounding names like Dustin and Scott; white female-sounding names like Claire and Abigail; Black male-sounding names like DaQuan and Jamal; and Black female-sounding names like Janae and Keyana.

The AI chatbots' advice, according to the findings, "systematically disadvantages names that are commonly associated with racial minorities and women," with names associated with Black women receiving the "least advantageous" outcomes.

Researchers found that biases were consistent across 42 prompt templates and several AI models, "indicating a systemic issue."

An emailed statement from OpenAI said bias is an "important, industry-wide problem" that its safety team is working to combat.

Google did not immediately respond to a request for comment.