Medical Xpress

Study assesses GPT-4's potential to perpetuate racial, gender biases in clinical decision making

December 18 2023

Credit: Pixabay/CC0 Public Domain

Large language models (LLMs) like ChatGPT and GPT-4 have the potential to assist in clinical practice to automate administrative tasks, draft clinical notes, communicate with patients, and even support clinical decision making. However, preliminary studies suggest the models can encode and perpetuate social biases that could adversely affect historically marginalized groups.

A new study by investigators from Brigham and Women's Hospital has evaluated the tendency of GPT-4 to encode and exhibit racial and gender biases in four clinical decision support roles. Their results are published in *The Lancet Digital Health*.

"While most of the focus is on using LLMs for documentation or administrative tasks, there is also excitement about the potential to use LLMs to support clinical decision making," said corresponding author Emily Alsentzer, Ph.D., a postdoctoral researcher in the Division of General Internal Medicine at Brigham and Women's Hospital. "We wanted to systematically assess whether GPT-4 encodes racial and gender biases that impact its ability to support clinical decision making."

Alsentzer and colleagues tested four applications of GPT-4 using the Azure OpenAI platform. First, they prompted GPT-4 to generate patient vignettes that can be used in medical education. Next, they tested GPT-4's ability to correctly develop a differential diagnosis and treatment plan for 19 different patient cases from an *NEJM* Healer, a medical education tool that presents challenging clinical cases to medical trainees.

Finally, they assessed how GPT-4 makes inferences about a patient's clinical presentation using eight case vignettes that were originally generated to measure implicit bias. For each application, the authors assessed whether GPT-4's outputs were biased by race or gender.

For the medical education task, the researchers constructed ten prompts that required GPT-4 to generate a patient presentation for a supplied diagnosis. They ran each prompt 100 times and found that GPT-4 exaggerated known differences in disease prevalence by demographic group.

"One striking example is when GPT-4 is prompted to generate a vignette for a patient with sarcoidosis: GPT-4 describes a Black woman 81% of the time," Alsentzer explains. "While sarcoidosis is more prevalent in Black patients and in women, it's not 81% of all patients."

Next, when GPT-4 was prompted to develop a list of 10 possible diagnoses for the *NEJM* Healer cases, changing the gender or race/ethnicity of the patient significantly affected its ability to prioritize the correct top diagnosis in 37% of cases.

"In some cases, GPT-4's decision making reflects known gender and racial biases in the literature," Alsentzer said. "In the case of pulmonary embolism, the model ranked panic attack/anxiety as a more likely diagnosis for women than men. It also ranked sexually transmitted diseases, such as acute HIV and syphilis, as more likely for patients from racial minority backgrounds compared to white patients."

When asked to evaluate subjective patient traits such as honesty, understanding, and pain tolerance, GPT-4 produced significantly different responses by race, ethnicity, and gender for 23% of the questions. For example, GPT-4 was significantly more likely to rate Black male patients as abusing the opioid Percocet than Asian, Black, Hispanic, and white female patients when the answers should have been identical for all the simulated patient cases.

Limitations of the current study include testing GPT-4's responses using a limited number of simulated prompts and analyzing model performance using only a few traditional categories of demographic identities. Future work should investigate biases using clinical notes from the electronic health record.

"While LLM-based tools are currently being deployed with a clinician in the loop to verify the model's outputs, it is very challenging for clinicians to detect systemic biases when viewing individual patient cases," Alsentzer said. "It is critical that we perform bias evaluations for each intended use of LLMs, just as we do for other machine learning models in the medical domain. Our work can help start a conversation about GPT-4's potential to propagate bias in clinical decision support applications."

Additional BWH authors include Jorge A Rodriguez, David W Bates, and Raja-Elie E Abdulnour. Additional authors include Travis Zack, Eric Lehman, Mirac Suzgun, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, and Atul J Butte.

More information: Travis Zack et al, Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study, *The Lancet Digital Health* (2023). DOI: 10.1016/S2589-7500(23)00225-X

Provided by Brigham and Women's Hospital

Citation: Study assesses GPT-4's potential to perpetuate racial, gender biases in clinical decision making (2023, December 18) retrieved 27 April 2024 from https://medicalxpress.com/news/2023-12-gpt-potential-perpetuate-racial-gender.html