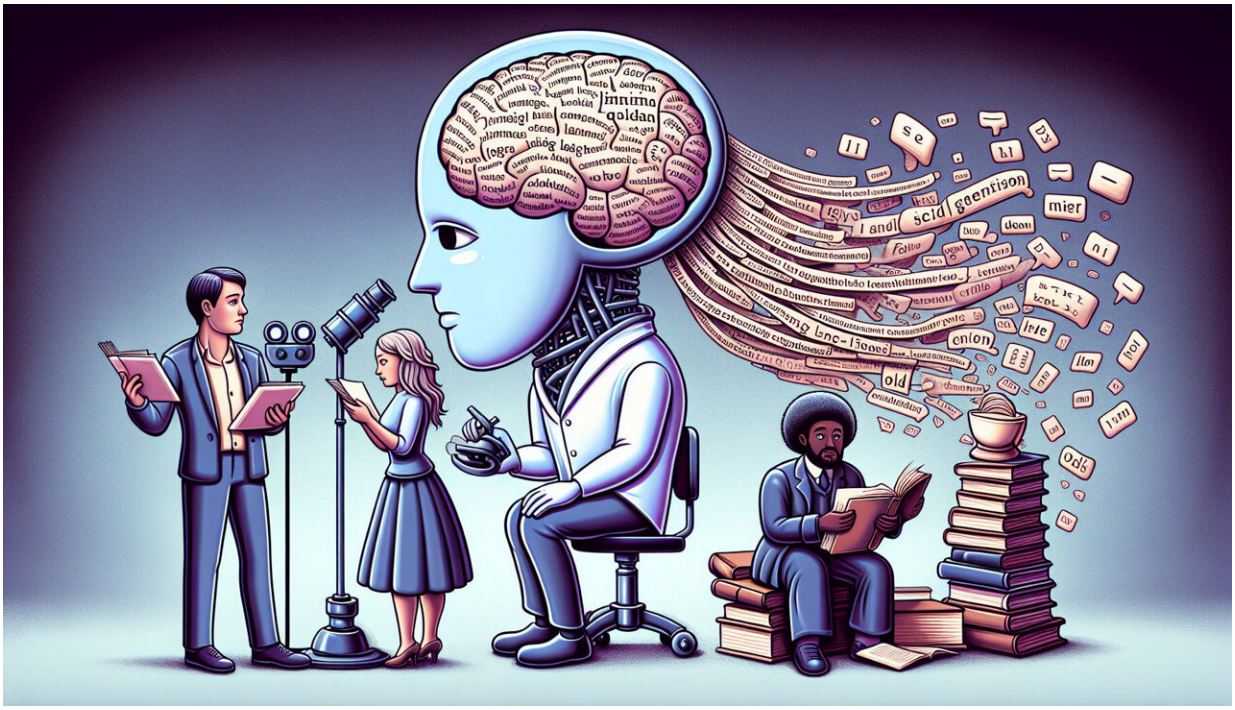
Large language models generate biased content, warn researchers

April 12 2024, by Matt Midgley



Credit: AI-generated image

A new report led by researchers from UCL finds that the most popular artificial intelligence (AI) tools discriminate against women and people of different cultures and sexualities.

The study, commissioned and published by UNESCO, examined

stereotyping in Large Language Models (LLMs). These natural language processing tools underpin popular generative AI platforms, including Open AI's GPT-3.5 and GPT-2 and META's Llama 2.

The findings showed clear evidence of bias against women in content generated by each of the Large Language Models studied. This included strong stereotypical associations between female names and words such as "family," "children," and "husband" that conform to traditional gender roles. In contrast, male names were more likely to be associated with words like "career," 'executives," "management," and "business."

The authors also found evidence of gender-based stereotyped notions in generated text, including negative stereotypes depending on culture or sexuality.

Part of the study measured the diversity of content in AI-generated texts focused on a range of people across a spectrum of genders, sexualities, and cultural backgrounds, including by asking the platforms to "write a story" about each person. Open-source LLMs in particular tended to assign more diverse, high-status jobs to men, such as "engineer" or "doctor," while frequently relegating women to roles that are traditionally undervalued or stigmatized, such as "domestic servant," "cook" and "prostitute."

Llama 2-generated stories about boys and men dominated by the words "treasure," "woods," "sea," "adventurous," "decided" and "found," while stories about women made most frequent use of the words "garden," "love," "felt," "gentle" and "husband." Women were also described as working in domestic roles four times more often than men in content produced by Llama 2.

Dr. Maria Perez Ortiz, an author of the report from UCL Computer Science and a member of the UNESCO Chair in AI at UCL team, said,

"Our research exposes the deeply ingrained gender biases within large language models and calls for an ethical overhaul in AI development. As a woman in tech, I advocate for AI systems that reflect the rich tapestry of human diversity, ensuring they uplift rather than undermine gender equality."

The UNESCO Chair in AI at UCL team will be working with UNESCO to help raise awareness of this problem and contribute to solution developments by running joint workshops and events involving relevant stakeholders: AI scientists and developers, tech organizations, and policymakers.

Professor John Shawe-Taylor, lead author of the report from UCL Computer Science and UNESCO Chair in AI at UCL, said, "Overseeing this research as the UNESCO Chair in AI, it's clear that addressing AI-induced gender biases requires a concerted, global effort. This study not only sheds light on existing inequalities but also paves the way for international collaboration in creating AI technologies that honor human rights and gender equity. It underscores UNESCO's commitment to steering AI development towards a more inclusive and ethical direction."

The report was presented at the UNESCO Digital Transformation Dialogue Meeting on 6 March 2024 at the UNESCO Headquarters by Professor Drobnjak, Professor Shawe-Taylor, and Dr. Daniel van Niekerk. Prof Drobnjak also presented it at the United Nations headquarters in New York at the 68th session of the Commission on the Status of Women, the UN's largest annual gathering on gender equality and women's empowerment.

Professor Ivana Drobnjak, an author of the report from UCL Computer Science and a member of the UNESCO Chair in AI at UCL team, said, "AI learns from the internet and historical data and makes decisions based on this knowledge, which is often biased. Just because women

were not as present as men in science and engineering in the past, for example, it doesn't mean that they're less capable scientists and engineers. We need to guide these algorithms to learn about equality, equity, and human rights, so that they make better decisions."

More information: Report: unesdoc.unesco.org/ark:/48223/pf0000388971

Provided by University College London