Researchers surprised by gender stereotypes in ChatGPT

March 5 2024, by Anne Kirsten Frederiksen

As a student at DTU, Sara Sterlie has analysed ChatGPT and revealed that the online service is extremely stereotypical when it comes to gender roles. Credit: Frida Gregersen

A DTU student has analyzed ChatGPT and revealed that the online service is extremely stereotypical when it comes to gender roles. The analysis is the first step toward providing AI developers with a tool for testing against all types of discriminatory bias.

It caused quite a stir when ChatGPT launched in 2022, giving anyone with internet access the opportunity to use artificial intelligence to create texts and answer questions. Not least because ChatGPT "behaves" like a human and provides answers that a colleague or friend could have written.

In her studies at DTU, Sara Sterlie has focused on artificial intelligence and quickly became interested in investigating bias in ChatGPT in relation to gender stereotypes. It may sound simple, but in order to do so, she had to develop a method for carrying out relevant experiments first.

"When Sara approached me with her ideas for her project, I was immediately interested and agreed to be her supervisor. I already work with bias in artificial intelligence, but I haven't previously worked with language models like ChatGPT," says Professor Aasa Feragen, who primarily works with bias in artificial intelligence used for medical image processing.

Adapting a method for ChatGPT

As her starting point, Sara Sterlie chose Non-Discrimination Criteria—a recognized method for analyzing bias in another type of artificial intelligence model that classifies material, e.g., for assessing medical images. It is easy to train that model to know the difference between X-rays showing healthy or diseased lungs, for example. It is then possible to measure whether the classification model presents too many incorrect answers, e.g., depending on whether the image is of a man or a woman.

"ChatGPT is different in that it doesn't provide predictable answers that fit neatly into categories. Moreover, when we ask the model a question there is not always an inherently true answer, as in classification tasks. The methods usually used to measure bias are therefore not directly applicable to models like ChatGPT. I wanted a solid foundation for my

investigation chose to develop methods by reinterpreting the non-discrimination criteria," explains Sara Sterlie.

Sara Sterlie chose to develop methods simplified to only consider male and female gender. She defined statistical requirements, and designed experimental questions or prompts, as they are called, with a focus on gender-bias, which she tested on GPT models.

"The first were a series of structured prompts, focusing on occupational stereotypes. I wanted to test what gender the model associated with different jobs. My first experiment asks the model to specify the job of a person given a male or female name," says Sara Sterlie.

Biased answers

The result was a far more stereotypical distribution in relation to gender than we have in today's society, with women mainly assigned job titles such as graphic designer, fashion designer, or nurse and men assigned job titles such as software engineer, architect, and executive.

Another type of prompt dealt with typical job functions in professions that work closely together, e.g., doctors and nurses or pilots and flight attendants. The results of the hundreds of experiments Sara Sterlie conducted for each prompt showed that ChatGPT has a hard time associating male pronouns with nurses and an even harder time letting female pronouns handle a pilot's duties of getting a plane ready for landing.

Furthermore, Sara Sterlie also conducted experiments with unstructured prompts, asking ChatGPT to describe the hobbies of a number of high school students with boy and girl names respectively. Sterlie then analyzed the answers, examining how often a word or phrase appeared in a text, among other things. It became abundantly clear that among the

400 responses, there was an unusually large number of female students who engaged in volunteer work with animals, whereas the male students were particularly interested in technology and science.

"All my experiments unanimously showed that ChatGPT exhibits a clear gender bias, both when asked in structured and unstructured ways," says Sara Sterlie.

Researchers surprised

Sara Sterlie and her two supervisors, Aasa Feragen and Nina Weng, who also work with medical image processing, had partially predicted the outcome of the experiments.

"We expected some gender bias, as ChatGPT is trained on material from the internet that to some extent reflects the gender stereotypes we've known for many years. But I was very surprised to see the extent of the bias, particularly in relation to the link between gender and job types. It's way off compared to the distribution in modern society," says Nina Weng.

Sara Sterlie and her supervisors are currently working on completing a scientific article about their findings.

"As far as I know, we're the first to have done this type of analysis. Our long-term goal as researchers is to develop methods and tools that can be used by the developers behind language models like ChatGPT to prevent bias in terms of gender, race, nationality, etc. We're not there yet, but Sara's experiments are the first step," says Nina Weng.

Aasa Feragen adds that she expects that Sara Sterlie's methods will start a global discussion about how to avoid bias in artificial intelligence.

Ensuring fairness

Being interested in uncovering and avoiding bias in ChatGPT and similar generative artificial intelligence models is not criticizing the new technology, say Sara Sterlie and Nina Weng, who both use ChatGPT, for example, to summarize the main points in a text. Rather, they are passionate about ensuring fairness in the texts or images generated by the language model and other generative artificial intelligence models.

"The increasing use of artificial intelligence to create texts or images will affect our perception of the world around us. AI's like ChatGPT are trained on large amounts of data and delivers responses which resembles the patterns in its training data. This means if you don't fit into the norms of the average person in terms of sexuality, family type, or personal preferences, which are typically dominating in the training data, you typically won't be represented in articles etc. produced by these AI models," explains Sara Sterlie.

The researchers want to create the basis for fairness to ensure that artificial intelligence will not exclude representation of other groups in the future.

Provided by Technical University of Denmark

Citation: Researchers surprised by gender stereotypes in ChatGPT (2024, March 5) retrieved 21 June 2024 from https://techxplore.com/news/2024-03-gender-stereotypes-chatgpt.html