GPT detectors can be biased against non-native English writers

July 10 2023

Credit: Unsplash/CC0 Public Domain

In an opinion paper published July 10 in the journal *Patterns*, researchers show that computer programs commonly used to determine if a text was

written by artificial intelligence tend to falsely label articles written by non-native language speakers as AI-generated. The researchers caution against the use of such AI text detectors for their unreliability, which could have negative impacts on individuals including students and those applying for jobs.

"Our current recommendation is that we should be extremely careful about and maybe try to avoid using these detectors as much as possible," says senior author James Zou, of Stanford University. "It can have significant consequences if these detectors are used to review things like job applications, college entrance essays or high school assignments."

AI tools like OpenAI's ChatGPT chatbot can compose essays, solve science and math problems, and produce computer code. Educators across the U.S. are increasingly concerned about the use of AI in students' work and many of them have started using GPT detectors to screen students' assignments. These detectors are platforms that claim to be able to identify if the text is generated by AI, but their reliability and effectiveness remain untested.

Zou and his team put seven popular GPT detectors to the test. They ran 91 English essays written by non-native English speakers for a widely recognized English proficiency test, called Test of English as a Foreign Language, or TOEFL, through the detectors. These platforms incorrectly labeled more than half of the essays as AI-generated, with one detector flagging nearly 98% of these essays as written by AI. In comparison, the detectors were able to correctly classify more than 90% of essays written by eighth-grade students from the U.S. as human-generated.

Zou explains that the algorithms of these detectors work by evaluating text perplexity, which is how surprising the word choice is in an essay. "If you use common English words, the detectors will give a low perplexity score, meaning my essay is likely to be flagged as AI-

generated. If you use complex and fancier words, then it's more likely to be classified as human written by the algorithms," he says. This is because large language models like ChatGPT are trained to generate text with low perplexity to better simulate how an average human talks, Zou adds.

As a result, simpler word choices adopted by non-native English writers would make them more vulnerable to being tagged as using AI.

The team then put the human-written TOEFL essays into ChatGPT and prompted it to edit the text using more sophisticated language, including substituting simple words with complex vocabulary. The GPT detectors tagged these AI-edited essays as human-written.

"We should be very cautious about using any of these detectors in classroom settings, because there's still a lot of biases, and they're easy to fool with just the minimum amount of prompt design," Zou says. Using GPT detectors could also have implications beyond the education sector. For example, search engines like Google devalue AI-generated content, which may inadvertently silence non-native English writers.

While AI tools can have positive impacts on student learning, GPT detectors should be further enhanced and evaluated before putting into use. Zou says that training these algorithms with more diverse types of writing could be one way to improve these detectors.

More information: James Zou, GPT detectors are biased against non-native English writers, *Patterns* (2023). DOI: 10.1016/j.patter.2023.100779. www.cell.com/patterns/fulltext … 2666-3899(23)00130-7

Provided by Cell Press