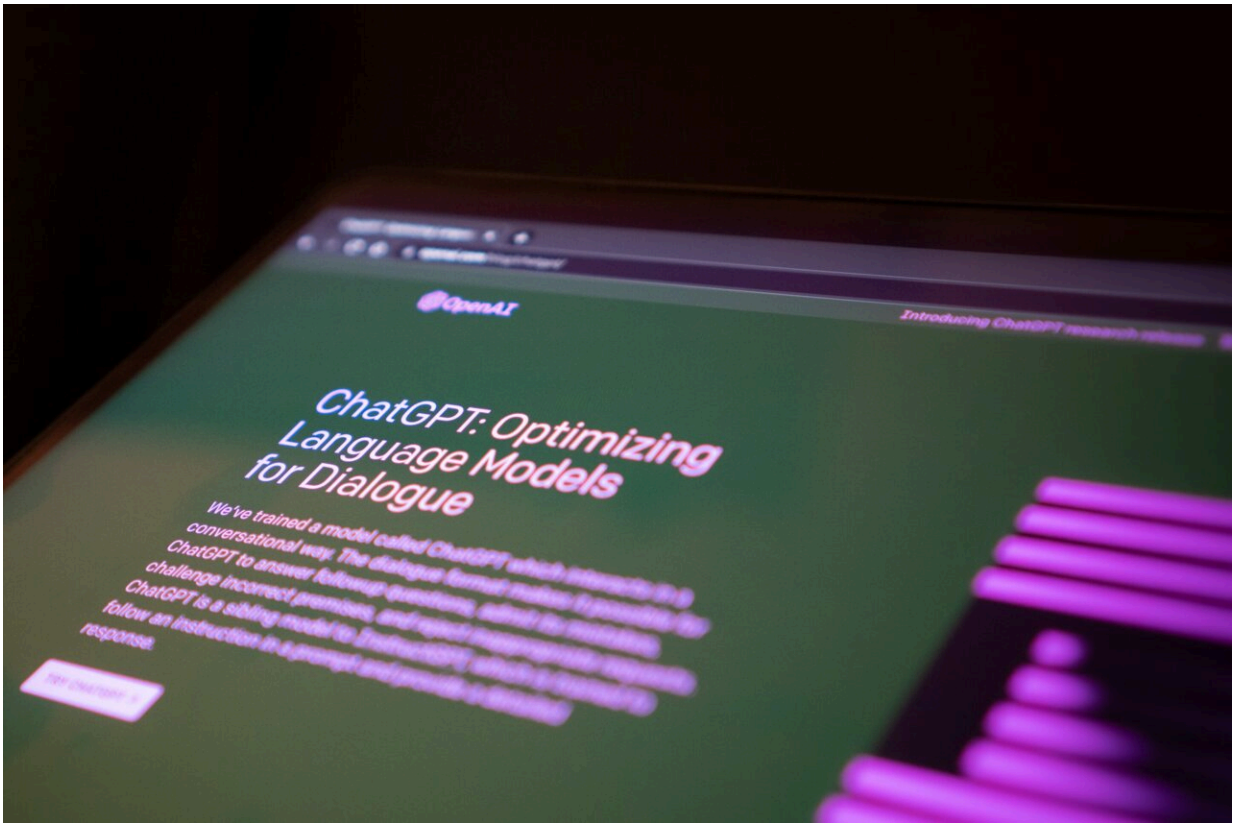
OpenAI launches new tool to deter cheating on its own platform

February 14 2023, by Elissa Miolene



Credit: Unsplash/CC0 Public Domain

ChatGPT—an artificial intelligence tool that can write essays, poems and emails on any subject with the click of a cursor—sent shockwaves throughout the education world when it was introduced late last year.

Now, its creators have built a new program that can help catch students who use the AI bot to cheat.

But instead of quelling teachers' fears, the new detection tool has been somewhat of a letdown within the technology and education worlds. Created by San Francisco-based company OpenAI, the platform identifies AI-written text accurately only a quarter of the time—and gives a false positive for nearly 1 in 10 submissions. Even a different detection tool created by a Princeton college student works slightly better.

Some experts worry that the company's detection program could lead to wrongly accusing students of plagiarism.

"I'm surprised OpenAI would release this tool with its current performance level," said Victor Lee, the faculty lead for AI + Education at the Stanford Accelerator for Learning. "If it's taken up too quickly, it could be really risky and harmful to students."

After the chatbot's release late last year, school districts across the country reacted with alarm. With the ability to churn out content in seconds, ChatGPT was seen by many as the ultimate cheat code for students—and, some believed, the nail in the coffin for original writing. Students began turning in AI-written assignments; companies began integrating ChatGPT into their copywriting protocols. And within weeks, its use was banned in schools across New York City, Los Angeles and Seattle.

Some of the world's largest tech companies scrambled to catch up. Last week, Microsoft revealed it was integrating ChatGPT's technology into its Bing search engine. Google followed with its own version, called Bard.

"It's going to change everything," said Jake Carr, an English teacher in Chico, earlier this month.

By developing a tool to distinguish between human-written and AI-written text, the creators of ChatGPT, San Francisco-based OpenAI, hope to deter the bot's involvement with automated misinformation campaigns, cheating in schools and other sorts of AI-infused deception.

"We're making this classifier publicly available to get feedback on whether imperfect tools like this one are useful," OpenAI said in a recent blog post. "Our work on the detection of AI-generated text will continue, and we hope to share improved methods in the future."

But the company's text classifier is late to the game. In early January, Princeton student Edward Tian built GPTZero, a tool with the tagline "humans deserve the truth."

When comparing the same two ChatGPT-written essays on OpenAI's text classifier and Tian's GPTZero—three paragraphs on the dangers of plagiarism—the text classifier said the text was "possibly AI-generated," while GPTZero said it was "likely to be written entirely by AI."

On top of that, the "AI text classifier" is not particularly reliable. According to the company, the tool identified computer-written text in 26% of their evaluations, while incorrectly labeling human-written text as AI 9% of the time. Detection is particularly poor with text shorter than 1,000 characters. Still, OpenAI said its text classifier is more accurate than other tools trying to analyze text.

Lee said he appreciates the company's openness about its classifier's limitations, but he still has concerns.

"I do worry that many people will hear that there is a tool that will detect

things without looking at the fine print about its performance."

MediaNews Group, Inc.

Distributed by Tribune Content Agency, LLC.