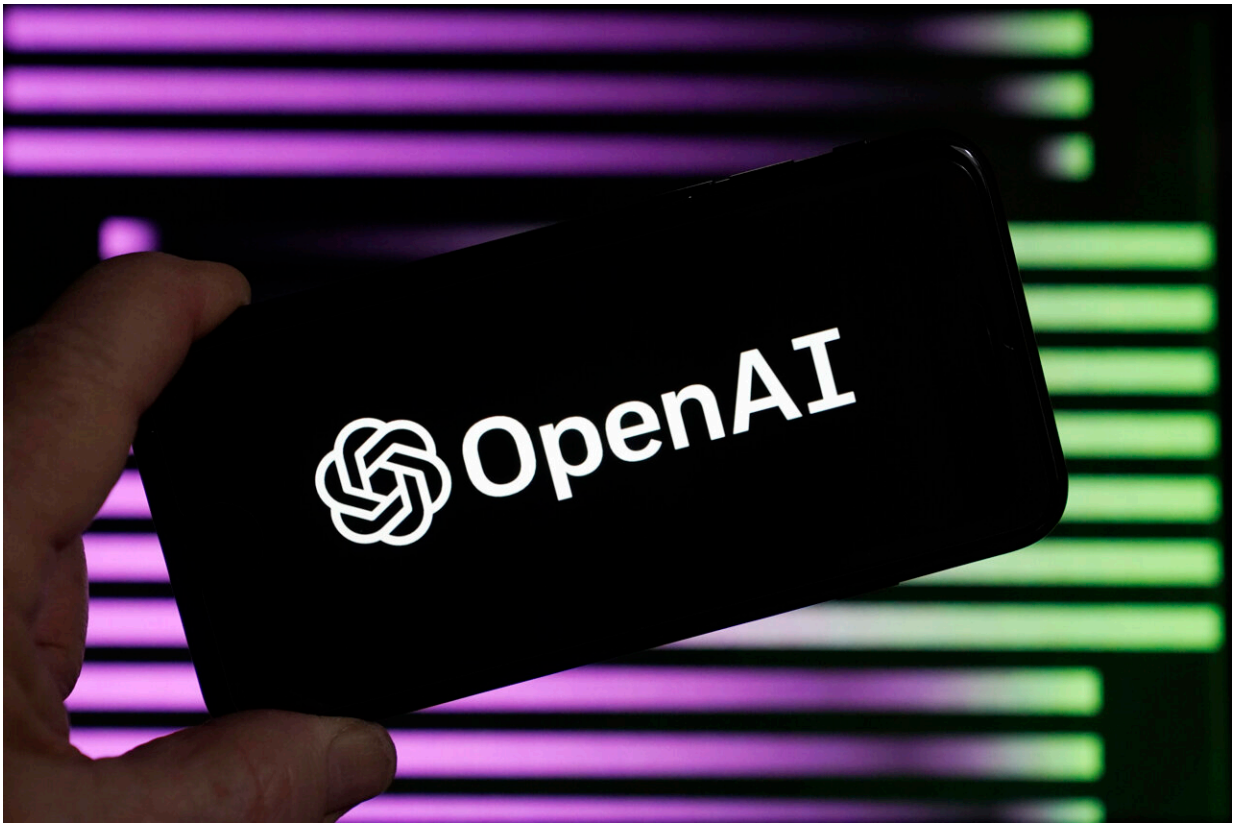
Cheaters beware: ChatGPT maker releases AI detection tool

January 31 2023, by MATT O'BRIEN and JOCELYN GECKER



The logo for OpenAI, the maker of ChatGPT, appears on a mobile phone, in New York, Tuesday, Jan. 31, 2023. OpenAI is launching a new tool in an effort to curb its reputation as a freewheeling cheating machine with a new tool Tuesday that can help teachers detect if a student or artificial intelligence wrote that homework. Credit: AP Photo/Richard Drew

The maker of ChatGPT is trying to curb its reputation as a freewheeling cheating machine with a new tool that can help teachers detect if a student or artificial intelligence wrote that homework.

The new AI Text Classifier launched Tuesday by OpenAI follows a weeks-long discussion at schools and colleges over fears that ChatGPT's ability to write just about anything on command could fuel academic dishonesty and hinder learning.

OpenAI cautions that its new tool —like others already available—is not foolproof. The method for detecting AI-written text "is imperfect and it will be wrong sometimes," said Jan Leike, head of OpenAI's alignment team tasked to make its systems safer.

"Because of that, it shouldn't be solely relied upon when making decisions," Leike said.

Teenagers and college students were among the millions of people who began experimenting with ChatGPT after it launched Nov. 30 as a free application on OpenAI's website. And while many found ways to use it creatively and harmlessly, the ease with which it could answer take-home test questions and assist with other assignments sparked a panic among some educators.

By the time schools opened for the new year, New York City, Los Angeles and other big public school districts began to block its use in classrooms and on school devices.

The Seattle Public Schools district initially blocked ChatGPT on all school devices in December but then opened access to educators who want to use it as a teaching tool, said Tim Robinson, the district spokesman.

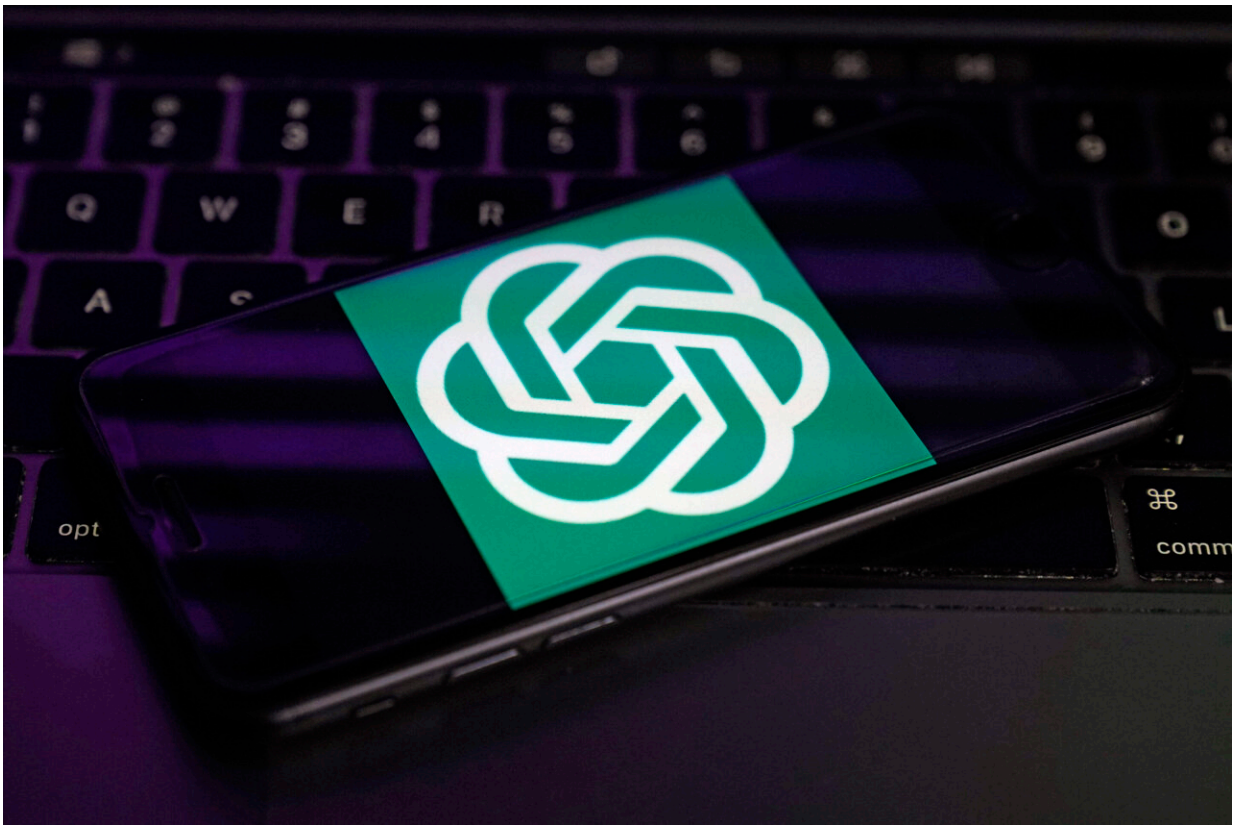
"We can't afford to ignore it," Robinson said.

The district is also discussing possibly expanding the use of ChatGPT into classrooms to let teachers use it to train students to be better critical thinkers and to let students use the application as a "personal tutor" or to help generate new ideas when working on an assignment, Robinson said.

School districts around the country say they are seeing the conversation around ChatGPT evolve quickly.

"The initial reaction was 'OMG, how are we going to stem the tide of all the cheating that will happen with ChatGPT,'" said Devin Page, a technology specialist with the Calvert County Public School District in Maryland. Now there is a growing realization that "this is the future" and blocking it is not the solution, he said.

"I think we would be naïve if we were not aware of the dangers this tool poses, but we also would fail to serve our students if we ban them and us from using it for all its potential power," said Page, who thinks districts like his own will eventually unblock ChatGPT, especially once the company's detection service is in place.



The logo for OpenAI, the maker of ChatGPT, appears on a mobile phone, in New York, Tuesday, Jan. 31, 2023. OpenAI is launching a new tool in an effort to curb its reputation as a freewheeling cheating machine with a new tool Tuesday that can help teachers detect if a student or artificial intelligence wrote that homework. Credit: AP Photo/Richard Drew

OpenAI emphasized the limitations of its detection tool in a blog post Tuesday, but said that in addition to deterring plagiarism, it could help to detect automated disinformation campaigns and other misuse of AI to mimic humans.

The longer a passage of text, the better the tool is at detecting if an AI or human wrote something. Type in any text—a college admissions essay, or a literary analysis of Ralph Ellison's "Invisible Man" —- and the tool

will label it as either "very unlikely, unlikely, unclear if it is, possibly, or likely" AI-generated.

But much like ChatGPT itself, which was trained on a huge trove of digitized books, newspapers and online writings but often confidently spits out falsehoods or nonsense, it's not easy to interpret how it came up with a result.

"We don't fundamentally know what kind of pattern it pays attention to, or how it works internally," Leike said. "There's really not much we could say at this point about how the classifier actually works."

Higher education institutions around the world also have begun debating responsible use of AI technology. Sciences Po, one of France's most prestigious universities, prohibited its use last week and warned that anyone found surreptitiously using ChatGPT and other AI tools to produce written or oral work could be banned from Sciences Po and other institutions.

In response to the backlash, OpenAI said it has been working for several weeks to craft new guidelines to help educators.

"Like many other technologies, it may be that one district decides that it's inappropriate for use in their classrooms," said OpenAI policy researcher Lama Ahmad. "We don't really push them one way or another. We just want to give them the information that they need to be able to make the right decisions for them."

It's an unusually public role for the research-oriented San Francisco startup, now backed by billions of dollars in investment from its partner Microsoft and facing growing interest from the public and governments.

France's digital economy minister Jean-Noël Barrot recently met in

California with OpenAI executives, including CEO Sam Altman, and a week later told an audience at the World Economic Forum in Davos, Switzerland that he was optimistic about the technology. But the government minister—a former professor at the Massachusetts Institute of Technology and the French business school HEC in Paris—said there are also difficult ethical questions that will need to be addressed.

"So if you're in the law faculty, there is room for concern because obviously ChatGPT, among other tools, will be able to deliver exams that are relatively impressive," he said. "If you are in the economics faculty, then you're fine because ChatGPT will have a hard time finding or delivering something that is expected when you are in a graduate-level economics faculty."

He said it will be increasingly important for users to understand the basics of how these systems work so they know what biases might exist.