23052 ChatGPT maker OpenAI releases 'not fully reliable' tool to detect AI generated content

Guardian Website 01 Feb 23

OpenAI, the research laboratory behind AI program ChatGPT, has released a classifier tool designed to detect whether text has been written by any artificial intelligence, not just ChatGPT, but warns this classifier tool is not completely reliable – yet.
Open AI researchers said that while it was "impossible to reliably detect all AI-written text", good classifiers could pick up signs that text was written by AI. The tool could be useful in cases where AI was used for "academic dishonesty" and when AI chatbots were positioned as humans, they said.

But they admitted the classifier "is not fully reliable" and only correctly identified 26% of AI-written English texts. It also incorrectly labelled human-written texts as probably written by AI tools 9% of the time.

"Our classifier's reliability typically improves as the length of the input text increases. Compared to our previously released classifier, this new classifier is significantly more reliable on text from more recent AI systems."

Since ChatGPT was opened up to public access, it has sparked a wave of concern among educational institutions across the world that it could lead to cheating in exams or assessments.

Lecturers in the UK are being urged to review the way in which their courses were assessed, while some universities have banned the technology entirely and returned to pen-and-paper exams to stop students using AI.
One lecturer at Australia's Deakin university said around one in five of the assessments she was marking over the Australian summer period had used AI assistance.
A number of science journals have also banned the use of ChatGPT in text for papers.

OpenAI said the classifier tool had several limitations, including its unreliability on text below 1,000 characters, as well as the misidentification of some human-written text as AI-written. The researchers also said it should only be used for English text, as it performs "significantly worse" in other languages, and is unreliable on checking code.

"It should not be used as a primary decision-making tool, but instead as a complement to other methods of determining the source of a piece of text," OpenAI said.

OpenAI has now called upon educational institutions to share their experiences with the use of ChatGPT in classrooms.

While most have responded to AI with bans, some have embraced the AI wave. The three main universities in South Australia last month updated their policies to state AI like ChatGPT is allowed to be used so long as it is disclosed.

403 words