Is AI-generated content actually detectable?

May 31 2023, by Georgia Jiang

Credit: Pixabay/CC0 Public Domain

In recent years, artificial intelligence (AI) has made tremendous strides thanks to advances in machine learning and growing pools of data to learn from. Large language models (LLMs) and their derivatives, such as OpenAI's ChatGPT and Google's BERT, can now generate material that is increasingly similar to content created by humans. As a result, LLMs

have become popular tools for creating high-quality, relevant and coherent text for a range of purposes, from composing social media posts to drafting academic papers.

Despite the wide variety of potential applications, LLMs face increasing scrutiny. Critics, especially educators and original content creators, view LLMs as a means for plagiarism, cheating, deception and manipulative social engineering.

In response to these concerns, researchers have developed novel methods to help distinguish between human-made content and machine-generated texts. The hope is that the ability to identify automated content will limit LLM abuse and its consequences.

But University of Maryland computer scientists are working to answer an important question: can these detectors accurately identify AI-generated content?

The short answer: No—at least, not now

"Current detectors of AI aren't reliable in practical scenarios," said Soheil Feizi, an assistant professor of computer science at UMD. "There are a lot of shortcomings that limit how effective they are at detecting. For example, we can use a paraphraser and the accuracy of even the best detector we have drops from 100% to the randomness of a coin flip. If we simply paraphrase something that was generated by an LLM, we can often outwit a range of detecting techniques."

In a recent paper, Feizi described two types of errors that impact an AI text detector's reliability: type I (when human text is detected as AI-generated) and type II (when AI-generated text is simply not detected).

"Using a paraphraser, which is now a fairly common tool available

online, can cause the second type of error," explained Feizi, who also holds a joint appointment in the University of Maryland Institute for Advanced Computer Studies. "There was also a recent example of the first type of error that went viral. Someone used AI detection software on the U.S. Constitution and it was flagged as AI-generated, which is obviously very wrong."

According to Feizi, such mistakes made by AI detectors can be extremely damaging and often impossible to dispute when authorities like educators and publishers accuse students and other content creators of using AI. When and if such accusations are proven false, the companies and individuals responsible for developing the faulty AI detectors could also suffer reputational loss.

In addition, even LLMs protected by watermarking schemes remain vulnerable against spoofing attacks where adversarial humans can infer hidden watermarks and add them to non-AI text so that it's detected to be AI-generated. Reputations and intellectual property may be irreversibly tainted after faulty results—a major reason why Feizi calls for caution when it comes to relying solely on AI detectors to authenticate human-created content.

"Let's say you're given a random sentence," Feizi said. "Theoretically, you can never reliably say that this sentence was written by a human or some kind of AI because the distribution between the two types of content is so close to each other. It's especially true when you think about how sophisticated LLMs and LLM-attackers like paraphrasers or spoofing are becoming."

"The line between what's considered human and artificial becomes even thinner because of all these variables," he added. "There is an upper bound on our detectors that fundamentally limits them, so it's very unlikely that we'll be able to develop detectors that will reliably identify

AI-generated content."

Another view: More data could lead to better detection

UMD Assistant Professor of Computer Science Furong Huang has a more optimistic outlook on the future of AI detection.

Although she agrees with her colleague Feizi that current detectors are imperfect, Huang believes that it is possible to point out artificially generated content—as long as there are enough examples of what constitutes human-created content available. In other words, when it comes to AI analysis, more is better.

"LLMs are trained on massive amounts of text. The more information we feed to them, the better and more human-like their outputs," explained Huang, who also holds a joint appointment in the University of Maryland Institute for Advanced Computer Studies. "If we do the same with detectors—that is, provide them more samples to learn from—then the detectors will also grow more sophisticated. They'll be better at spotting AI-generated text."

Huang's recent paper on this topic examined the possibility of designing superior AI detectors, as well as determining how much data would be required to improve its detection capabilities.

"Mathematically speaking, we'll always be able to collect more data and samples for detectors to learn from," said UMD computer science Ph.D. student Souradip Chakraborty, who is a co-author of the paper. "For example, there are numerous bots on social media platforms like Twitter. If we collect more bots and the data they have, we'll be better at discerning what's spam and what's human text on the platform."

Huang's team suggests that detectors should take a more holistic approach and look at bigger samples to try to identify this AI-generated "spam."

"Instead of focusing on a single phrase or sentence for detection, we suggest using entire paragraphs or documents," added Amrit Singh Bedi, a research scientist at the Maryland Robotics Center who is also a co-author of Huang's paper. "Multiple sentence analysis would increase accuracy in AI detection because there is more for the system to learn from than just an individual sentence."

Huang's group also believes that the innate diversity within the human population makes it difficult for LLMs to create content that mimics human-produced text. Distinctly human characteristics such as certain grammatical patterns and word choices could help identify text that was written by a person rather than a machine.

"It'll be like a constant arms race between generative AI and detectors," Huang said. "But we hope that this dynamic relationship actually improves how we approach creating both the generative LLMs and their detectors in the first place."

What's next for AI and AI detection

Although Feizi and Huang have differing opinions on the future of LLM detection, they do share several important conclusions that they hope the public will consider moving forward.

"One thing's for sure—banning LLMs and apps like ChatGPT is not the answer," Feizi said. "We have to accept that these tools now exist and that they're here to stay. There's so much potential in them for fields like education, for example, and we should properly integrate these tools into systems where they can do good."

Feizi suggests in his research that security methods used to counter generative LLMs, including detectors, don't need to be 100% foolproof—they just need to be more difficult for attackers to break, starting with closing the loopholes that researchers already know about. Huang agrees.

"We can't just give up if the detector makes one mistake in one instance," Huang said. "There has to be an active effort to protect the public from the consequences of LLM abuse, particularly members of our society who identify as minorities and are already encountering social biases in their lives."

Both researchers also believe that multimodality (the use of text in conjunction with images, videos and other forms of media) will also be key to improved AI detection in the future. Feizi cites the use of secondary verification tools already in practice, such as authenticating phone numbers linked to social media accounts or observing behavioral patterns in content submissions, as additional safeguards to prevent false AI detection and bias.

"We want to encourage open and honest discussion about ethical and trustworthy applications of generative LLMs," Feizi said. "There are so many ways we can use these AI tools to improve our society, especially for student learning or preventing the spread of misinformation."

As AI-generated texts become more pervasive, researchers like Feizi and Huang recognize that it's important to develop more proactive stances in how the public approaches LLMs and similar forms of AI.

"We have to start from the top," Huang said. "Stakeholders need to start having a discussion about these LLMs and talk to policymakers about setting ground rules through regulation. There needs to be oversight on how LLMs progress while researchers like us develop better detectors,

watermarks or other approaches to handling AI abuse."

Both papers are published on the *arXiv* preprint server.

More information: Vinu Sankar Sadasivan et al, Can AI-Generated Text be Reliably Detected?, *arXiv* (2023). DOI: 10.48550/arxiv.2303.11156

Souradip Chakraborty et al, On the Possibilities of AI-Generated Text Detection, *arXiv* (2023). DOI: 10.48550/arxiv.2304.04736

Provided by University of Maryland