Study finds it's impossible to reliably detect AI-generated text

Alarm bells are sounding off in College Park, Maryland.

According to a recent study, researchers doubt it will ever be possible to reliably detect AI-generated text. Along with four computer science doctoral students, Professor Soheil Feizi's study asks, "Can AI-generated text be reliably detected?" Their answer, unfortunately, is no.

Given that large language models (LLMs) can be used to plagiarize, conduct convincing social engineering attacks, and spread misinformation at scale, this study is definitely a cause for concern.

Current AI-generated text detection tools leave a lot to be desired

OpenAI's AI-generated text detector is terribly inaccurate. In fact, OpenAI admits that it is not reliable, as the tool only correctly identifies 26% of AI-written text (true positives). Additionally, 9% of the time, it mislabels human-written text as being AI-written (false positives). Another popular tool on the market, GPTZero, essentially measures the randomness of a given piece of text. According to GPTZero's FAQ page, that tool is capable of identifying human-created text 99% of the time and AI-generated text 85% of the time, although some may take umbrage with this assertion.

Can AI-generated text be reliably detected?

Through empirical analysis, the University of Maryland scholars looked at several popular AI text-detection models on the market and found them to be unreliable. Examining watermarking schemes, zero-shot classifiers, and neural network-based detectors, they found that a paraphrasing attack can help adversaries evade AI detection. They write, "We show that a paraphrasing attack, where a lightweight neural network-based paraphraser is applied to the output text of the AI-generative model, can evade various types of detectors."

Additionally, they claim that watermarking-based detectors can be easily spoofed, making it seem as if human-made text is watermarked. Such adversarial spoofing

attacks could ruin the credibility of watermarking LLM development companies.

Feizi believes that social media account verification may be a good way to fight the spread of disinformation. In an email to *The Register*, Feizi explains,

"***I think we need to learn to live with the fact that we may never be able to reliably say if a text is written by a human or an AI.*** *Instead, potentially we can verify the 'source' of the text via other information. For example, many social platforms are starting to widely verify accounts. This can make the spread of misinformation generated by AI more difficult.*"

It's almost too obvious to point out, but relying on social media platforms—Meta, Twitter, and others—to handle account verification in a way that controls misinformation and benefits society, is a tough ask. As a quick example, it remains to be seen whether a blue check mark next to a Twitter handle denotes a reputable account.

GPT detectors are biased against non-native English writers

When it comes to detecting AI-generated content, it's still early days; however, another recent study suggests that these text detection tools are biased against non-native English speakers. A Stanford study found that these tools commonly misclassified non-native English speakers' writing as being AI-generated.

The Stanford researchers looked at seven widely-used GPT detectors: Originality.ai; Quil.org; Sapling; OpenAI (GPT-2); Crossplag; GPTZero, and ZeroGPT. When comparing the detection tools' performance against native English speakers' writing and non-native English speakers' writing, the scholars found that GPT detectors unfairly penalize non-native speakers. Finding a real bias against non-native English writers, the Stanford scholars caution against relying on such detection tools in academic settings.

As LLMs like ChatGPT improve, it will likely become increasingly difficult to differentiate between human-written content and AI-generated text. However, it's important to note that strides have been made on the AI-generated image and video detection front.

It's easier to watermark images and video

Unlike AI-generated text, which the Maryland scholars claim will be nearly impossible to authenticate, synthetic images and video are easier to identify. By watermarking an image or video at inception, it's possible to establish that media's content provenance.

The author of the video or image, as well as the location and metadata, can be cryptographically signed, timestamped, and stored on a blockchain. Earlier this month, San Diego-based startup Truepic collaborated with Revel.ai and Nina Schick to create what they're calling the "world's first digitally transparent deepfake video."

Entitled "Mirror of Reflection," this deepfake is sensationalist, but effective. Schick rhetorically asks, "What if our reality is changing? What if we can no longer rely on our senses to determine the authenticity of what we see and hear?"

If the scholars from University of Maryland are to be believed, such a day isn't too far away.

About ManageEngine

ManageEngine crafts comprehensive IT management software with a focus on making your job easier. Our 120+ award-winning products and free tools cover everything your IT needs. From network and device management to security and service desk software, we're bringing IT together for an integrated, overarching approach to optimize your IT.

For more information, visit **manageengine.com**

ManageEngine Insights is the thought-leadership and knowledge sharing platform of ManageEngine. As the go-to destination for tech enthusiasts, we offer tailored content specifically crafted to keep you in the know about the ever-evolving world of technology.

LinkedIn | **Twitter** | **Facebook** | **Instagram**