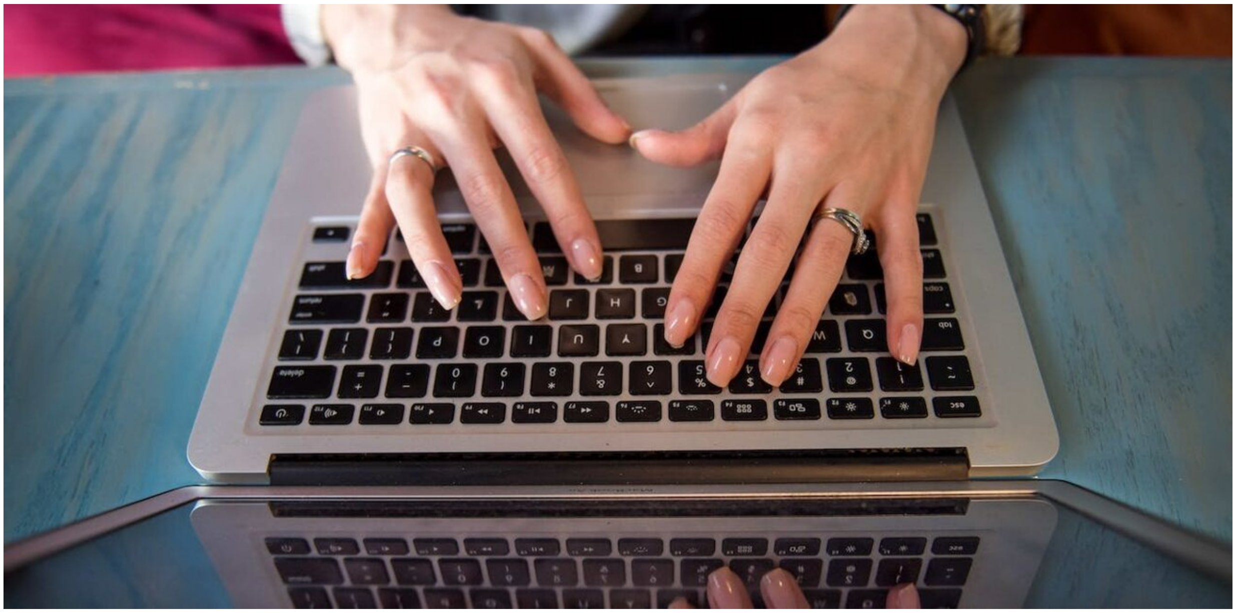
TechXplore

We pitted ChatGPT against tools for detecting AI-written text, and the results are troubling

February 20 2023, by Armin Alimardani and Emma A. Jane



Credit: Melanie Deziel / Unsplash

As the "chatbot wars" rage in Silicon Valley, the growing proliferation of artificial intelligence (AI) tools specifically designed to generate human-like text has left many baffled.

Educators in particular are scrambling to adjust to the availability of software that can produce a moderately competent essay on any topic at a moment's notice. Should we go back to pen-and-paper assessments? Increasing exam supervision? Ban the use of AI entirely?

All these and more have been proposed. However, none of these less-than-ideal measures would be needed if educators could reliably distinguish AI-generated and human-written text.

We dug into several proposed methods and tools for recognizing AI-generated text. None of them are foolproof, all of them are vulnerable to workarounds, and it's unlikely they will ever be as reliable as we'd like.

Perhaps you're wondering why the world's leading AI companies can't reliably distinguish the products of their own machines from the work of humans. The reason is ridiculously simple: the corporate mission in today's high-stakes AI arms is to train "natural language processor" (NLP) AIs to produce outputs that are as similar to human writing as possible. Indeed, public demands for an easy means to spot such AIs in the wild might seem paradoxical, like we're missing the whole point of the program.

A mediocre effort

OpenAI—the creator of ChatGPT—launched a "classifier for indicating AI-written text" in late January.

The classifier was trained on external AIs as well as the company's own text-generating engines. In theory, this means it should be able to flag essays generated by BLOOM AI or similar, not just those created by ChatGPT.

We give this classifier a C– grade at best. OpenAI admits it accurately identifies only 26% of AI-generated text (true positive) while incorrectly labeling human prose as AI-generated 9% of the time (false positive).

OpenAI has not shared its research on the rate at which AI-generated text is incorrectly labeled as human-generated text (false negative).

A promising contender

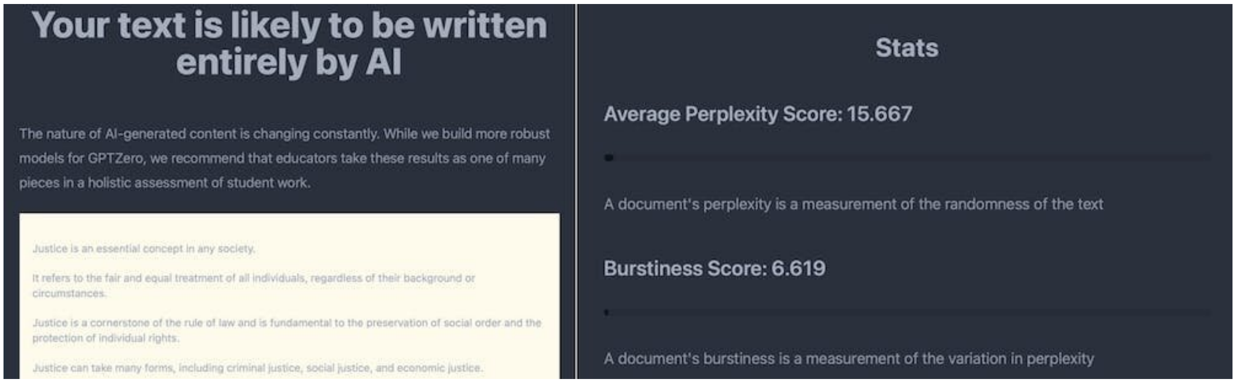
A more promising contender is a classifier created by a Princeton University student during his Christmas break.

Edward Tian, a computer science major minoring in journalism, released the first version of GPTZero in January.

This app identifies AI authorship based on two factors: perplexity and burstiness. Perplexity measures how complex a text is, while burstiness compares the variation between sentences. The lower the values for these two factors, the more likely it is that a text was produced by an AI.

We pitted this modest David against the goliath of ChatGPT.

First, we prompted ChatGPT to generate a short essay about justice. Next, we copied the article—unchanged—into GPTZero. Tian's tool correctly determined that the text was likely to have been written entirely by an AI because its average perplexity and burstiness scores were very low.



Your text is likely to be written entirely by AI

The nature of AI-generated content is changing constantly. While we build more robust models for GPTZero, we recommend that educators take these results as one of many pieces in a holistic assessment of student work.

Justice is an essential concept in any society.

It refers to the fair and equal treatment of all individuals, regardless of their background or circumstances.

Justice is a cornerstone of the rule of law and is fundamental to the preservation of social order and the protection of individual rights.

Justice can take many forms, including criminal justice, social justice, and economic justice.

Stats

Average Perplexity Score: 15.667

A document's perplexity is a measurement of the randomness of the text

Burstiness Score: 6.619

A document's burstiness is a measurement of the variation in perplexity

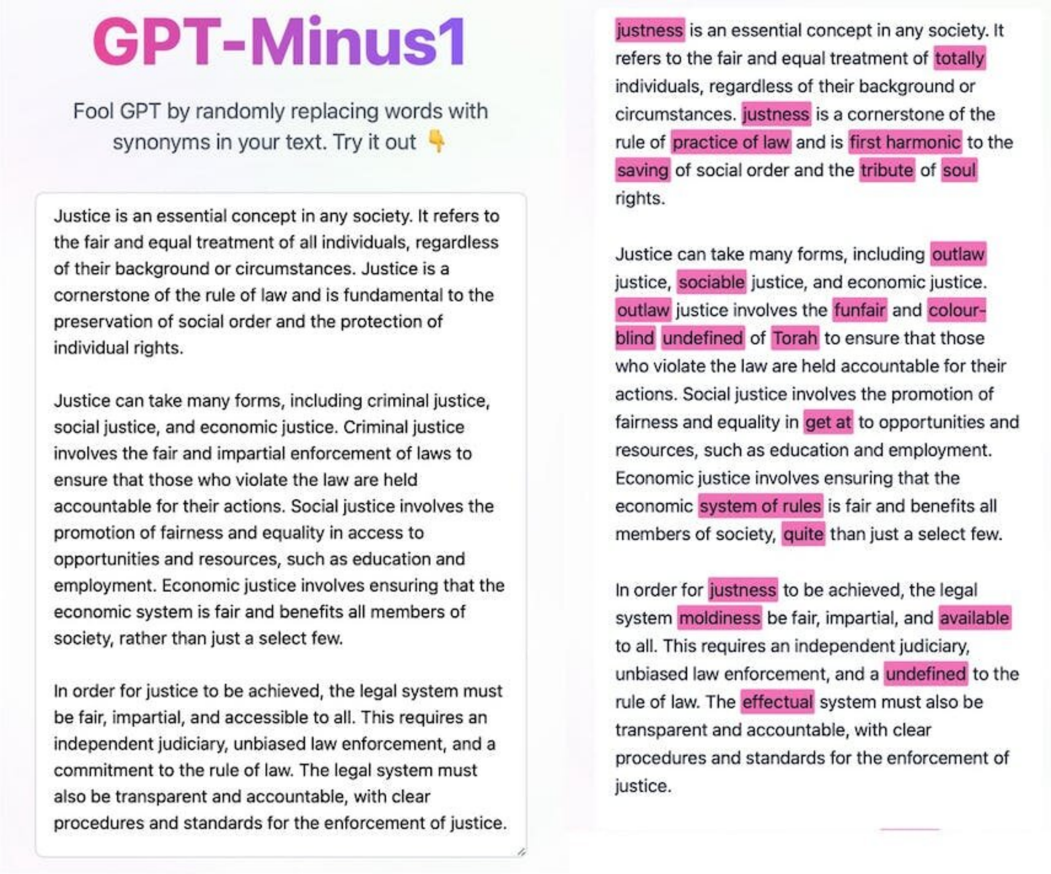
GPTZero measures the complexity and variety within a text to determine whether it is likely to have been produced by AI. Credit: GTPZero

Fooling the classifiers

An easy way to mislead AI classifiers is simply to replace a few words with synonyms. Websites offering tools that paraphrase AI-generated text for this purpose are already cropping up all over the internet.

Many of these tools display their own set of AI giveaways, such as peppering human prose with "tortured phrases" (for example, using "counterfeit consciousness" instead of "AI").

To test GPTZero further, we copied ChatGPT's justice essay into GPT-Minus1—a website offering to "scramble" ChatGPT text with synonyms. The image on the left depicts the original essay. The image on the right shows GPT-Minus1's changes. It altered about 14% of the text.



GPT-Minus1

Fool GPT by randomly replacing words with synonyms in your text. Try it out 👇

Justice is an essential concept in any society. It refers to the fair and equal treatment of all individuals, regardless of their background or circumstances. Justice is a cornerstone of the rule of law and is fundamental to the preservation of social order and the protection of individual rights.

Justice can take many forms, including criminal justice, social justice, and economic justice. Criminal justice involves the fair and impartial enforcement of laws to ensure that those who violate the law are held accountable for their actions. Social justice involves the promotion of fairness and equality in access to opportunities and resources, such as education and employment. Economic justice involves ensuring that the economic system is fair and benefits all members of society, rather than just a select few.

In order for justice to be achieved, the legal system must be fair, impartial, and accessible to all. This requires an independent judiciary, unbiased law enforcement, and a commitment to the rule of law. The legal system must also be transparent and accountable, with clear procedures and standards for the enforcement of justice.

justness is an essential concept in any society. It refers to the fair and equal treatment of totally individuals, regardless of their background or circumstances. justness is a cornerstone of the rule of practice of law and is first harmonic to the saving of social order and the tribute of soul rights.

Justice can take many forms, including outlaw justice, sociable justice, and economic justice. outlaw justice involves the funfair and colour-blind undefined of Torah to ensure that those who violate the law are held accountable for their actions. Social justice involves the promotion of fairness and equality in get at to opportunities and resources, such as education and employment. Economic justice involves ensuring that the economic system of rules is fair and benefits all members of society, quite than just a select few.

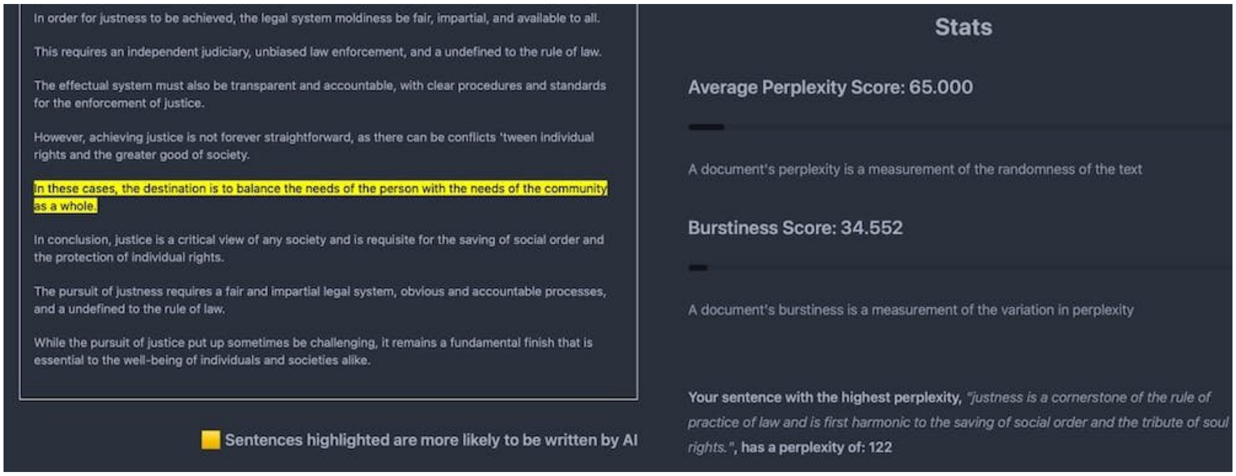
In order for justness to be achieved, the legal system moldiness be fair, impartial, and available to all. This requires an independent judiciary, unbiased law enforcement, and a undefined to the rule of law. The effectual system must also be transparent and accountable, with clear procedures and standards for the enforcement of justice.

GPT-Minus1 makes small changes to text to make it look less AI-generated. Credit: GPT-Minus1

We then copied the GPT-Minus1 version of the justice essay back into GPTZero. Its verdict?

"Your text is most likely human written but there are some sentences with low perplexities."

It highlighted just one sentence it thought had a high chance of having been written by an AI (see image below on left) along with a report on the essay's overall perplexity and burstiness scores which were much higher (see image below on the right).



In order for justness to be achieved, the legal system moldiness be fair, impartial, and available to all.

This requires an independent judiciary, unbiased law enforcement, and a undefined to the rule of law.

The effectual system must also be transparent and accountable, with clear procedures and standards for the enforcement of justice.

However, achieving justice is not forever straightforward, as there can be conflicts 'tween individual rights and the greater good of society.

In these cases, the destination is to balance the needs of the person with the needs of the community as a whole.

In conclusion, justice is a critical view of any society and is requisite for the saving of social order and the protection of individual rights.

The pursuit of justness requires a fair and impartial legal system, obvious and accountable processes, and a undefined to the rule of law.

While the pursuit of justice put up sometimes be challenging, it remains a fundamental finish that is essential to the well-being of individuals and societies alike.

🟨 Sentences highlighted are more likely to be written by AI

Stats

Average Perplexity Score: 65.000

A document's perplexity is a measurement of the randomness of the text

Burstiness Score: 34.552

A document's burstiness is a measurement of the variation in perplexity

Your sentence with the highest perplexity, *"justness is a cornerstone of the rule of practice of law and is first harmonic to the saving of social order and the tribute of soul rights."*, **has a perplexity of: 122**

Running an AI-generated text through an AI-fooling tool makes it seem 'more human'. Credit: GPTZero

Tools such as Tian's show great promise, but they aren't perfect and are also vulnerable to workarounds. For instance, a recently released YouTube tutorial explains how to prompt ChatGPT to produce text with high degrees of—you guessed it—perplexity and burstiness.

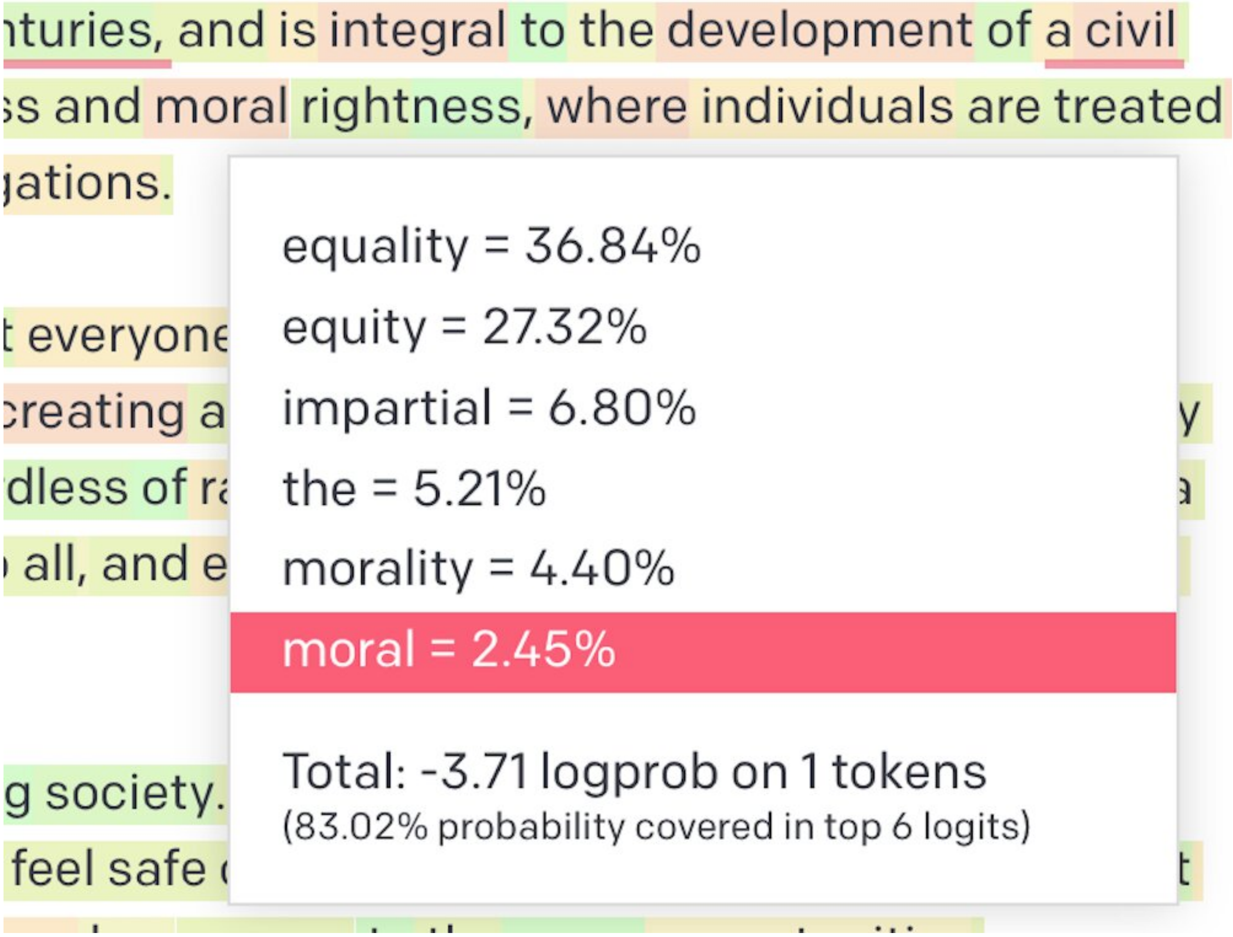
Watermarking

Another proposal is for AI-written text to contain a "watermark" that is invisible to human readers but can be picked up by software.

Natural language models work on a word-by-word basis. They select which word to generate based on statistical probability.

However, they do not always choose words with the highest probability of appearing together. Instead, from a list of probable words, they select one randomly (though words with higher probability scores are more likely to be selected).

This explains why users get a different output each time they generate text using the same prompt.



turies, and is integral to the development of a civil

s and moral rightness, where individuals are treated

ations.

t everyone

creating a

dless of ra

all, and e

g society.

feel safe

equality = 36.84%

equity = 27.32%

impartial = 6.80%

the = 5.21%

morality = 4.40%

moral = 2.45%

Total: -3.71 logprob on 1 tokens

(83.02% probability covered in top 6 logits)

One of OpenAI's natural language model interfaces (Playground) gives users the ability to see the probability of selected words. In the above screenshot (captured on Feb 1, 2023), we can see that the likelihood of the term 'moral' being selected is 2.45%, which is much less than 'equality' with 36.84%. Credit: OpenAI Playground

Put simply, watermarking involves "blacklisting" some of the probable words and permitting the AI to only select words from a "whitelist." Given that a human-written text will likely include words from the "blacklist," this could make it possible to differentiate it from an AI-generated text.

However, watermarking also has limitations. The quality of AI-generated text might be reduced if its vocabulary was constrained. Further, each text generator would likely have a different watermarking system—so text would next to checked against all of them.

Watermarking could also be circumvented by paraphrasing tools, which might insert blacklisted words or rephrase essay questions.

An ongoing arms race

AI-generated text detectors will become increasingly sophisticated. Anti-plagiarism service TurnItIn recently announced a forthcoming AI writing detector with a claimed 97% accuracy.

However, text generators too will grow more sophisticated. Google's ChatGPT competitor, Bard, is in early public testing. OpenAI itself is expected to launch a major update, GPT-4, later this year.

It will never be possible to make AI text identifiers perfect, as even OpenAI acknowledges, and there will always be new ways to mislead them.

As this arms race continues, we may see the rise of "contract paraphrasing": rather than paying someone to write your assignment, you pay someone to rework your AI-generated assignment to get it past the detectors.

There are no easy answers here for educators. Technical fixes may be part of the solution, but so will new ways of teaching and assessment (which may including harnessing the power of AI).

We don't know exactly what this will look like. However, we have spent the past year building prototypes of open-source AI tools for education and research in an effort to help navigate a path between the old and the new—and you can access beta versions at Safe-To-Fail AI.

This article is republished from The Conversation under a Creative Commons license. Read the original article.

Provided by The Conversation

Citation: We pitted ChatGPT against tools for detecting AI-written text, and the results are troubling (2023, February 20) retrieved 10 April 2024 from https://techxplore.com/news/2023-02-pitted-chatgpt-tools-ai-written-text.html