Too many em dashes? Weird words like ‘delves’? Spotting text written by ChatGPT is still more art than science

Faculty Contributor
Dr. Roger J. Kreuz

People are now routinely using chatbots to write computer code, summarize articles and books, or solicit advice. But these chatbots are also employed to quickly generate text from scratch, with some users passing off the words as their own. This has, not surprisingly, created headaches for teachers tasked with evaluating their students' written work. It's also created issues for people seeking advice on forums like Reddit, or consulting product reviews before making a purchase. Over the past few years, researchers have been exploring whether it's even possible to distinguish human writing from artificial intelligence-generated text. But the best strategies to distinguish between the two may come from the chatbots themselves.

Too good to be human?

Several recent studies have highlighted just how difficult it is to determine whether text was generated by a human or a chatbot. Research participants recruited for a 2021 online study, for example, were unable to distinguish between human- and ChatGPT-generated stories, news articles and recipes. Language experts fare no better. In a 2023 study, editorial board members for top linguistics journals were unable to determine which article abstracts had been written by humans and which were generated by ChatGPT. And a 2024 study found that 94% of undergraduate exams written by ChatGPT went undetected by graders at a British university.

Clearly, humans aren't very good at this.

A commonly held belief is that rare or unusual words can serve as "tells" regarding authorship, just as a poker player might somehow give away that they hold a winning hand. Researchers have, in fact, documented a dramatic increase in relatively uncommon words, such as "delves" or "crucial," in articles published in scientific journals over the past couple of years. This suggests that unusual terms could serve as tells that generative AI has been used. It also implies that some researchers are actively using bots to write or edit parts of their submissions to academic journals. Whether this practice reflects wrongdoing is up for debate.

In another study, researchers asked people about characteristics they associate with chatbot-generated text. Many participants pointed to the excessive use of em dashes – an elongated dash used to set off text or serve as a break in thought – as one marker of computer-generated output. But even in this study, the participants' rate of AI detection was only marginally better than chance.

Given such poor performance, why do so many people believe that em dashes are a clear tell for chatbots? Perhaps it's because this form of punctuation is primarily employed by experienced writers. In other words, people may believe that writing that is "too good" must be artificially generated. But if people can't intuitively tell the difference, perhaps there are other methods for determining human versus artificial authorship.

Stylometry to the rescue?

Some answers may be found in the field of stylometry, in which researchers employ statistical methods to detect variations in the writing styles of authors. I'm a cognitive scientist who authored a book on the history of stylometric techniques. In it, I document how researchers developed methods to establish authorship in contested cases, or to determine who may have written anonymous texts. One tool for determining authorship was proposed by the Australian scholar John Burrows. He developed Burrows' Delta, a computerized technique that examines the relative frequency of common words, as opposed to rare ones, that appear in different texts. It may seem counterintuitive to think that someone's use of words like "the," "and" or "to" can determine authorship, but the technique has been impressively effec- tive.

Burrows' Delta, for example, was used to establish that Ruth Plumly Thompson, L. Frank Baum's successor, was the author of a disputed book in the "Wizard of Oz" series. It was also used to determine that love letters attributed to Confederate Gen. George Pickett were actually the inventions of his widow, LaSalle Corbell Pickett. A major drawback of Burrows' Delta and similar techniques is that they require a fairly large amount of text to reliably distinguish between authors. A 2016 study found that at least 1,000 words from each author may be required. A relatively short student essay, therefore, wouldn't provide enough input for a statistical technique to work its attribution magic.

More recent work has made use of what are known as BERT language models, which are trained on large amounts of human- and chatbot-generated text. The models learn the patterns that are common in each type of writing, and they can be much more discriminating than people: The best ones are between 80% and 98% accurate.

A stylometric technique called Burrow's Delta was used to identify LaSalle Corbell Pickett as the author of love letters attributed to her deceased husband, Confederate Gen. George Pickett. Encyclopedia Virginia.

However, these machine-learning models are "black boxes" – that is, we don't really know which features of texts are responsible for their impressive abilities. Researchers are actively trying to find ways to make sense of them, but for now, it isn't clear whether the models are detecting specific, reliable signals that humans can look for on their own.

A moving target

Another challenge for identifying bot-generated text is that the models themselves are constantly changing – sometimes in major ways. Early in 2025, for example, users began to express concerns that ChatGPT had become overly obsequious, with mundane queries deemed "amazing" or "fantastic." OpenAI addressed the issue by rolling back some changes it had made. Of course, the writing style of a human author may change over time as well, but it typically does so more gradually. At some point, I wondered what the bots had to say for themselves. I asked ChatGPT-4o: "How can I tell if some prose was generated by ChatGPT? Does it have any 'tells,' such as characteristic word choice or punctuation?" The bot admitted that distinguishing human from nonhuman prose "can be tricky." Nevertheless, it did provide me with a 10-item list, replete with examples.

These included the use of hedges – words like "often" and "generally" – as well as redundancy, an overreliance on lists and a "polished, neutral tone." It did mention "predictable vocabulary," which included certain adjectives such as "significant" and "notable," along with academic terms like "implication" and "complexity." However, though it noted that these features of chatbot-generated text are common, it concluded that "none are definitive on their own."

Chatbots are known to hallucinate, or make factual errors. But when it comes to talking about themselves, they appear to be surprisingly perceptive.