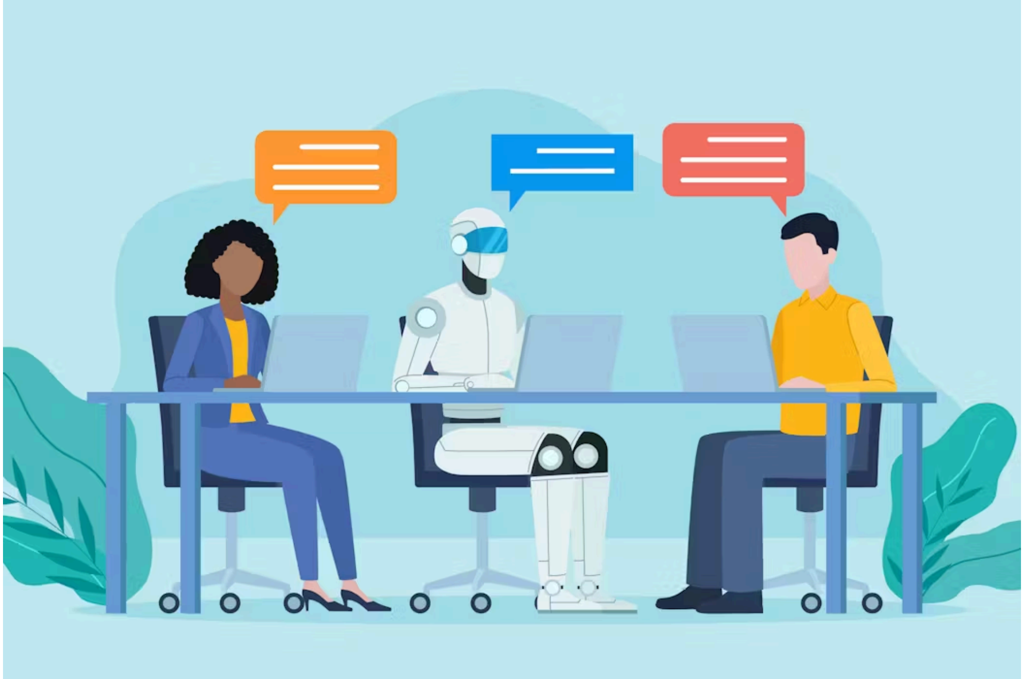
Why it’s so hard to tell if a piece of text was written by AI – even for AI

Ambuj Tewari, Professor of Statistics, University of Michigan

Published: December 18, 2025 8:33am EDT



Large language models have become extremely good at mimicking human writing.

Robert Wicher/iStock via Getty Images

People and institutions are grappling with the consequences of AI-written text. Teachers want to know whether students’ work reflects their own understanding; consumers want to know whether an advertisement was written by a human or a machine.

Writing rules to govern the use of AI-generated content is relatively easy. Enforcing them depends on something much harder: reliably detecting whether a piece of text was generated by artificial intelligence.

Some studies have investigated whether humans can detect AI-generated text. For example, people who themselves use AI writing tools heavily have been shown to accurately detect AI-written text. A panel of human evaluators can even outperform automated tools in a controlled setting. However, such expertise is not widespread, and individual judgment can be inconsistent. Institutions that need consistency at a large scale therefore turn to automated AI text detectors.

The problem of AI text detection

The basic workflow behind AI text detection is easy to describe. Start with a piece of text whose origin you want to determine. Then apply a detection tool, often an AI system itself, that analyzes the text and produces a score, usually expressed as a probability, indicating how likely the text is to have been AI-generated. Use the score to inform downstream decisions, such as whether to impose a penalty for violating a rule.

This simple description, however, hides a great deal of complexity. It glosses over a number of background assumptions that need to be made explicit. Do you know which AI tools might have plausibly been used to generate the text? What kind of access do you have to these tools? Can you run them yourself, or inspect their inner workings? How much text do you have? Do you have a single text or a collection of writings gathered over time? What AI detection tools can and cannot tell you depends critically on the answers to questions like these.

There is one additional detail that is especially important: Did the AI system that generated the text deliberately embed markers to make later detection easier?

These indicators are known as watermarks. Watermarked text looks like ordinary text, but the markers are embedded in subtle ways that do not reveal themselves to casual inspection. Someone with the right key can later check for the presence of these markers and verify that the text came from a watermarked AI-generated source. This approach, however, relies on cooperation from AI vendors and is not always available.

How AI text detection tools work

One obvious approach is to use AI itself to detect AI-written text. The idea is straightforward. Start by collecting a large corpus, meaning collection of writing, of examples labeled as human-written or AI-generated, then train a model to distinguish between the two. In effect, AI text detection is treated as a standard classification problem, similar in spirit to spam filtering. Once trained, the detector examines new text and predicts whether it more closely resembles the AI-generated examples or the human-written ones it has seen before.

The learned-detector approach can work even if you know little about which AI tools might have generated the text. The main requirement is that the training corpus be diverse enough to include outputs from a wide range of AI systems.

But if you do have access to the AI tools you are concerned about, a different approach becomes possible. This second strategy does not rely on collecting large labeled datasets or training a separate detector. Instead, it looks for statistical signals in the text, often in relation to how specific AI models generate language, to assess whether the text is likely to be AI-generated. For example, some methods examine the probability that an AI model assigns to a piece of text. If the model assigns an unusually high probability to the exact sequence of words, this can be a signal that the text was, in fact, generated by that model.

Finally, in the case of text that is generated by an AI system that embeds a watermark, the problem shifts from detection to verification. Using a secret key provided by the AI vendor, a verification tool can assess whether the text is consistent with having been generated by a watermarked system. This approach relies on information that is not available from the text alone, rather than on inferences drawn from the text itself.

Limitations of detection tools

Each family of tools comes with its own limitations, making it difficult to declare a clear winner. Learning-based detectors, for example, are sensitive to how closely new text resembles the data they were trained on. Their accuracy drops when the text differs substantially from the training corpus, which can quickly become outdated as new AI models are released. Continually curating fresh data and retraining detectors is costly, and detectors inevitably lag behind the systems they are meant to identify.

Statistical tests face a different set of constraints. Many rely on assumptions about how specific AI models generate text, or on access to those models' probability distributions. When models are proprietary, frequently updated or simply unknown, these assumptions break down. As a result, methods that work well in controlled settings can become unreliable or inapplicable in the real world.

Watermarking shifts the problem from detection to verification, but it introduces its own dependencies. It relies on cooperation from AI vendors and applies only to text generated with watermarking enabled.

More broadly, AI text detection is part of an escalating arms race. Detection tools must be publicly available to be useful, but that same transparency enables evasion. As AI text generators grow more capable and evasion techniques more sophisticated, detectors are unlikely to gain a lasting upper hand.

Hard reality

The problem of AI text detection is simple to state but hard to solve reliably. Institutions with rules governing the use of AI-written text cannot rely on detection tools alone for enforcement.

As society adapts to generative AI, we are likely to refine norms around acceptable use of AI-generated text and improve detection techniques. But ultimately, we’ll have to learn to live with the fact that such tools will never be perfect.

Ambuj Tewari receives funding from NSF and NIH.