# Turnitin’s AI writing detectors wrongly flag 4% of human work, sparking student backlash



The increasing reliance on AI writing detection tools in educational settings has sparked significant controversy, highlighting the precarious balance between innovation and fairness. Recently, Turnitin, a widely used academic integrity service, acknowledged a troubling aspect of its AI detection capabilities. According to their assessments, the software mistakenly identifies human-written content as AI-generated approximately 4% of the time, a statistic that raises alarms among students and educators alike.

This false positive rate is particularly concerning given the rise in AI-assisted learning, with many students unaware that their submissions might be scrutinised under such a lens. Turnitin's Chief Product Officer, Annie Chechitelli, noted that while the document-level false positive rate can be as low as 1% for texts containing a significant amount of AI writing, the sentence-level discrepancies can reach around 4%. This issue becomes more pronounced in mixed documents where both human and AI-generated content coexist, leading to complications in detection and assessment.

The implications of these findings are particularly dire for students. A recent petition led by Kelsey Auman at the University at Buffalo illustrates the anxiety surrounding AI detection tools. Auman, close to graduating from her master’s programme, recounted an alarming experience where her assignments were flagged by Turnitin, resulting in potential graduation delays for her and her classmates. As she pointed out, “It never occurred to me that I should keep evidence to prove my honesty when submitting my work.” The petition, which has garnered over 1,000 signatures, reflects a broader movement among students advocating for the suspension of AI detection technologies due to fears of erroneous accusations.

Further complicating the situation is the broader concern regarding the potential for unfair targeting of non-native English speakers. Research indicates that AI detection tools are particularly prone to flagging submissions from these students, with a notable study revealing that writing from non-native speakers was labelled as AI-generated 61% of the time. This disparity raises serious questions about the equitable application of these tools within diverse student populations.

In response to rising scrutiny, some institutions have opted to disable Turnitin's AI detection features entirely. Vanderbilt University, for example, ceased to use the tool after determining that the risk of false positives and the lack of transparency regarding the algorithm's operation outweighed its potential benefits. The decision underlines a significant shift in the landscape of academic integrity, where the reliability of AI evaluation mechanisms is now under intense review.

The question remains: can educational institutions balance the necessity of upholding academic integrity with the risks associated with increasingly complex AI detection systems? As conversations surrounding these technologies evolve, it becomes imperative that educators remain vigilant and foster open dialogues with students. According to Chechitelli, educators should utilise AI detection scores as a starting point for meaningful discussions rather than definitive conclusions about academic misconduct.

As the landscape of education adapts to technological advancements, the interplay between AI writing tools and academic integrity will undoubtedly require ongoing reflection and adjustment to ensure that fairness and justice are at the forefront of education practices.

### Reference Map

1. Paragraphs 1, 3, 4
2. Paragraphs 2, 6
3. Paragraphs 5, 7
4. Paragraphs 5, 6
5. Paragraph 6
6. Paragraphs 5, 6
7. Paragraph 7

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://www.nytimes.com/2025/05/17/style/ai-chatgpt-turnitin-students-cheating.html> - Please view link - unable to able to access data
2. <https://www.turnitin.com/blog/understanding-the-false-positive-rate-for-sentences-of-our-ai-writing-detection-capability> - In this blog post, Turnitin's Chief Product Officer, Annie Chechitelli, discusses the false positive rates of their AI writing detection tool. She explains that the document-level false positive rate is less than 1% for documents with 20% or more AI writing. However, the sentence-level false positive rate is around 4%, meaning there's a 4% chance that a sentence flagged as AI-written might actually be human-written. This is more common in documents containing a mix of human and AI-written content, especially near transitions between the two.
3. <https://www.insidehighered.com/news/quick-takes/2023/06/01/turnitins-ai-detector-higher-expected-false-positives> - This article reports on Turnitin's admission that their AI writing detection tool has a higher false positive rate than initially claimed. Originally, Turnitin promoted the tool as having a less than 1% false positive rate. However, real-world usage has shown discrepancies, with the company now acknowledging a higher incidence of false positives, particularly when less than 20% of a document is identified as AI-written. The article also notes that Turnitin plans to add an asterisk to scores under 20% to indicate reduced reliability.
4. <https://www.thejournal.com/articles/2023/05/30/turnitin-shares-stats-from-ai-detection-of-38m-submissions-tweaks-detector-feature.aspx/> - In this article, Turnitin shares statistics from the first six weeks of their AI writing detection feature, processing 38.5 million submissions. The findings revealed that 3.5% of submissions contained more than 80% AI-written text, and just under 10% had at least 20% AI-written content. In response to educator feedback, Turnitin introduced updates to the AI detection feature, including adding an asterisk to scores under 20% to indicate less reliability and increasing the minimum word count for submissions to 300 to improve accuracy.
5. <https://www.vanderbilt.edu/brightspace/2023/08/16/guidance-on-ai-detection-and-why-were-disabling-turnitins-ai-detector/> - Vanderbilt University announced the disabling of Turnitin's AI detection tool after several months of testing and feedback. The decision was based on concerns about the tool's reliability and the risk of false positives. The university highlighted that Turnitin's AI detection tool had a 1% false positive rate, which could lead to incorrect labeling of human-written papers as AI-generated. Additionally, there were concerns about the tool's transparency and the lack of detailed information on how it determines AI-generated content.
6. <https://www.highereddive.com/news/turnitin-false-positives-AI-detector/652356/> - This article discusses Turnitin's acknowledgment of higher-than-expected false positives in their AI writing detection tool. The company reported a higher incidence of false positives when less than 20% of AI writing is detected in a document. To address these concerns, Turnitin plans to display an asterisk when the detector spots under 20% of AI writing, indicating that the score is less reliable. The article also notes that Turnitin is increasing the minimum word count for submissions to 300 words to improve accuracy.
7. <https://www.themarkup.org/machine-learning/2023/08/14/ai-detection-tools-falsely-accuse-international-students-of-cheating> - This article highlights instances where AI detection tools, including Turnitin, have falsely accused international students of cheating. The author discusses how these tools are more likely to flag writing by non-native English speakers as AI-generated due to predictable word choices and simpler sentence structures. The article also references a study by Stanford computer scientists that found AI detectors flagged writing by non-native speakers as AI-generated 61% of the time, compared to rarely for native English speakers.