M FLINT OFFICE OF ONLINE & DIGITAL EDUCATION

Author: Nick Gaspar ngaspar@umich.edu

The Unreliability of AI-generated Writing Detection

AI detection tools, designed to identify AI-generated content, currently face reliability challenges. These tools often struggle with distinguishing between AI-generated written text and high-quality student work, potentially leading to false positives that can unfairly penalize students. Additionally, AI technology is rapidly evolving, and as it becomes more sophisticated, the tools designed to detect its output must also advance, but frequently lag behind. This discrepancy can result in both false negatives, where AI-generated writing goes undetected, and false positives, where genuine student work is mistakenly flagged. For these reasons, relying solely on AI-generated detection tools without human oversight can compromise academic integrity and fairness.

Limitations of Current AI Detection Tools

Turnitin's own description of its AI-generated writing detection capability underscores the tool's limitations and reinforces why it should not be used as the sole basis for disciplinary actions in academic settings. The company acknowledges that its model may misidentify both human and AI-generated text. This reliance risks unfair judgments on student work, either by missing AI-generated content or by mistakenly accusing students of misconduct.
For more details, see Turnitin’s AI writing detection update and AI writing resources.

OpenAI provides us with further insights into the limitations of this technology. Current AI detectors fail to reliably distinguish between AI-generated and human-written content, as evidenced by instances where even iconic human-authored texts were mislabeled as AI produced. This unreliability is compounded by the risk of disproportionately affecting students who may have a formulaic writing style or are non-native English speakers. OpenAI's research suggests that even with advancements, these tools could still be circumvented by minor edits to AI-generated content.
For more information, see OpenAI’s full guidance.

Research on AI Detection

Andrew Myers' article at Stanford highlights critical issues with AI detectors, especially regarding their performance in identifying non-native English speakers' writing as AI-generated. According to research conducted at Stanford, these detectors show a stark bias: while they perform nearly perfectly on essays by U.S.-born eighth graders, they incorrectly flag a significant majority (61.22%) of TOEFL essays, written by non-native English students, as AI-generated. This bias is even more pronounced as all detectors identified 18 out of 91 TOEFL student essays (about 19%) as AI-generated, and 89 out of 91 essays (97%) were flagged by at least one.

James Zou, a professor at Stanford, explains that these detectors often rely on the 'perplexity' metric, which measures the sophistication of writing, a measure where non-native speakers might naturally lag due to differences in lexical richness, diversity, and syntactic complexity. This reliance results in a disproportionate and unfair disadvantage to non-native speakers, raising significant ethical and fairness concerns. Experts like Zou highlight how easily these detectors can be manipulated, a broader review of AI text detection reveals additional systemic challenges.

Read the article:
https://hai.stanford.edu/news/ai-detectors-biased-against-non-native-english-writers

The survey "A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions" raises important questions about the reliability of AI-generated content detection tools. It highlights the inherent complexities in detecting text generated by large language models, particularly the challenge of accurately distinguishing between human and machine-generated text. Although there have been advancements in detection methods, such as watermarking and neural-based detectors, significant issues remain. These include difficulties with out-of-distribution scenarios, such as cross-domain, cross-lingual, and cross-LLM challenges; susceptibility to evasion through techniques like paraphrasing; and shortcomings in current evaluation frameworks.

The limitations highlight the unreliability of AI detection tools as the sole measure in educational settings. The survey also points out that these tools might struggle with real-world data and may not effectively handle new instances of LLM-generated text. This can result in both false positives and negatives, which could unjustly affect student evaluations and academic integrity assessments.

Read the paper: https://arxiv.org/pdf/2310.14724

While the outlined challenges with AI detection tools are significant, proponents argue that these tools are essential for preserving academic integrity by detecting AI-generated submissions efficiently. The intentions behind AI detection tools are to uphold academic standards, the current state of these technologies suggests the need for a more balanced approach that involves human oversight and educational strategies rather than reliance solely on automated systems.

Assessments in the AI Era

After exploring the complexities and limitations of AI detection tools, it's clear that adapting our assessment methods is essential to maintain academic integrity in the age of GenAI. If you're seeking practical guidance on designing assessments that truly reflect student learning in this new era, please visit our dedicated website. Here, you'll find comprehensive resources across three key areas: Discussion Board Design, Assignment Design, and Test Design. Each section offers innovative strategies and real-world examples from faculty who have successfully navigated the challenges of integrating AI tools into their evaluation practices.

What Other Institutions Have to Say

Massachusetts Institute of Technology
https://mitsloanedtech.mit.edu/ai/teach/ai-detectors-dont-work/
University of Central Florida
https://fctl.ucf.edu/technology/artificial-intelligence/
Vanderbilt University
https://www.vanderbilt.edu/brightspace/2023/08/16/guidance-on-ai-detection-and-why-were-disabling-turnitins-ai-detector/