University of Regina

CENTRE FOR TEACHING & LEARNING

AI-Writing Detection Tools: What Faculty Need to Know

The emergence of generative artificial intelligence (AI) tools, such as OpenAI's ChatGPT and Anthropic's Claude, has introduced new challenges in verifying student work as their own. In response to these challenges, dozens of AI-writing detection tools have recently become available. While these tools are marketed as a potential solution to this emerging problem, they are better understood as supplementary resources with accuracy and reliability that require cautious, critical consideration. To better understand the appropriate and ethical use of these tools, it is helpful to understand and acknowledge both their specific strengths and limitations.

Strengths of AI-Writing Detection Tools

- **Starting Point for Academic Misconduct Concerns:** AI-writing detection tools can provide an initial indicator of AI use in student submissions, serving as a potential starting point for further investigation.

- **Identifying Patterns:** Detection tools can provide highlighted sections of text that resemble AI-generated patterns, prompting faculty to take a closer look. These tools can also flag portions of the text that may have been written initially by a generative AI tool but has since been moderately edited by a human.

- **Promoting Awareness:** The known utilization of AI-writing detection tools may encourage students to use AI responsibly as part of a learning process rather than as a shortcut to completing assignments.

Limitations of AI-Writing Detection Tools

- **False Positives:** AI-writing detection tools can misidentify complex language, advanced vocabulary, or non-native English patterns as AI-generated content.

- **Lack of Verification:** Unlike plagiarism detection, which references specific sources, AI detection relies on probabilistic patterns, offering no direct way for faculty to verify the results.

- **Confidence Risks:** Treating detection scores as conclusive evidence may lead to unfair accusations and potential harm to students.

Given their strengths and limitations, faculty should employ these tools as helpful aids and should never consider their outputs as definitive proof of academic misconduct. While University of Regina faculty may employ AI-writing detection in assessing potential cases of academic misconduct, results must be evaluated critically using both additional supporting evidence and professional judgment. Currently, the only institutionally assessed, supported, and approved AI-writing detection tool at the University of Regina is Turnitin.

An Overview of Turnitin

Turnitin is an online tool adopted by many educational institutions worldwide to assist instructors in identifying potential plagiarism in student writing. By leveraging its extensive database of academic content, Turnitin cross-references student submissions against scholarly articles, online resources, and previously submitted student papers. This process supports academic integrity by identifying unoriginal content and promoting proper citation practices. ***In response to the release of ChatGPT in 2022, the platform expanded its capabilities to include AI-generated content detection.***

Implementation at the University of Regina

Turnitin is the only institutionally licensed, supported, and approved tool for the detection of plagiarism and AI-writing at the University of Regina. Many faculty and instructors are already familiar with Turnitin, available via UR Courses, where the tool:

- Integrates into courses through the "Add an Activity or Resource" option
- Automatically evaluates student papers for originality upon submission
- Generates similarity reports that highlight potentially plagiarized material, provides links to likely sources, and calculates percentage scores
- Allows students to view their similarity reports, make revisions, and resubmit work (if permitted by the instructor)

Turnitin's AI-Writing Detection Feature

Turnitin assesses submissions for AI-generated content by analyzing text patterns indicative of AI writing. The results appear as an AI writing indicator in Turnitin's similarity report, which can display one of four outcomes:

- **AI Detected:** Indicates a percentage score (between 20% and 100%) representing the amount of AI-generated content identified. In this case, the detection report will be further broken down to indicate what percentage of the text is likely AI-generated and what percentage is likely AI-paraphrased (i.e., text that was AI-generated and then modified by an AI paraphrasing tool).

- **Low Percentage:** Indicates a percentage score between 0-19%, where false positives are more likely. No additional information is provided.

- **Inconclusive data:** Indicates that the AI writing detector cannot process the submission. This may be for one of two reasons:
  - The writing was submitted to Turnitin before the AI detector feature was released.
  - The writing does not meet the submission guidelines related to file size, file type, and minimum/maximum word count.

- **Error:** Indicates that the submission has not been processed by Turnitin. In this case, the submission should be resubmitted at a later time.

Understanding Turnitin's AI Detection Score

1. What Turnitin's AI Detection Score Actually Measures

- Turnitin's AI detection score represents the possibility that parts of a text were generated by artificial intelligence based on probabilistic algorithms that compare language patterns and structures commonly found in AI-generated content against the submitted text.
- This score does not provide a definitive measure of AI authorship. Instead, it indicates how closely the writing resembles patterns typical of AI without fully accounting for individual writing styles, language proficiency, or the nuances of non-standard academic English.

2. Why AI Detection Scores Cannot Stand Alone in Academic Misconduct Cases

- AI detection tools, including Turnitin's, have inherent limitations due to their reliance on machine learning algorithms. These tools are not fully capable of accurately distinguishing between human and AI-generated writing in all cases. Over-reliance on AI detection scores can result in misinterpretations and, in some cases, unwarranted accusations of academic dishonesty.
- For example, a high AI detection score might result from a student's use of advanced vocabulary or formulaic language patterns rather than actual AI-generated content. Students employing specific writing strategies or those with varying proficiency in academic English may unintentionally create text that resembles AI-produced material.
- It is important to understand that, unlike plagiarism detectors, which can reference a source to confirm copied content, AI detectors cannot link to a specific source since none exists in cases of AI-generated writing. This means instructors cannot independently verify the detector's assessment, nor can students review and challenge the results. As a result, there is no clear way to determine how accurately or effectively the AI detector functions.
- Given these limitations, AI detection scores must be treated as supplementary indicators rather than conclusive evidence. Instructors are encouraged to consider the score as one factor among many, using it in conjunction with other evidence to determine if misconduct is likely to have occurred.

Implications for Students

1. The Potential Impacts of AI Detection Tools on Student Work

- **Encouraging More Responsible AI Use:** Knowing that Turnitin can detect AI-generated content may encourage some students to use AI tools more responsibly. If allowed by the instructor, they might focus on using AI for initial brainstorming, idea generation, or research rather than producing entire sections of their assignments. In this way, students may learn to integrate AI supportively rather than dependently, aligning with academic integrity.

- **Encouraging Less Responsible AI Use:** On the other hand, students who wish to avoid detection might seek ways to circumvent detection. There are various AI paraphrasing tools that students might use to alter the AI-generated content so that it appears more "human-like" or passes as original work, thus decreasing the AI detection score. In this way, students might seek methods to reduce detection scores using external tools, which may compromise academic integrity.

2. The Ethical Problems of False Positives

- **False Positives:** A significant ethical and logistical concern with AI detection scores is the potential for false positives. Inaccuracies in detection can lead to cases where genuine student work is flagged as AI-generated. While Turnitin actively attempts to reduce the risk of false positives through its algorithm and similarity score reporting, false positives are still inevitable.

- **Impact on Students:** False positives can unfairly harm students, leading to undue stress, financial loss, potential disciplinary actions, and damaged academic reputations. Moreover, the risk of false positives is particularly high for non-native English speakers and those whose writing style may not align with conventional academic norms. This means that students from diverse linguistic and cultural backgrounds may be disproportionately impacted, which raises concerns about equity and fairness.

3. Privacy Issues Surrounding AI Detection

- **Data Privacy Concerns:** Turnitin is the only institutionally approved tool for AI detection, as it adheres to the university's data security and privacy standards; the university does not support or condone the use of any other third-party tools intended to detect AI. This practice helps to ensure that student intellectual property is handled responsibly and is not used as training data for AI detection tools.