# Evaluating the Effectiveness of Turnitin's AI Writing Indicator Model

*Lori Salem, Student Success Center*
*Stephanie Fiore, Center for the Advancement of Teaching*
*Stephen Kelly, Student Success Center*
*Benjamin Brock, Center for the Advancement of Teaching*

## Introduction:

Turnitin recently developed what they call an "AI writing indicator model" that is intended to help instructors determine if a student has submitted work that is AI-generated. The model is integrated into Turnitin's existing plagiarism detection software licensed at Temple, and is therefore convenient as it is already embedded in Canvas [1]. The process for using it is familiar to anyone who has used Turnitin's signature product, the plagiarism detector, and it returns an "AI score" that closely mirrors the "similarity score" used in Turnitin's plagiarism detection tool.

If Turnitin's tool is effective in detecting AI-generated text, it would find a ready audience among faculty members at Temple who are concerned about students' misuse of generative AI tools, as they are anxious to find ways to detect and deter it.

But there are reasons to be concerned. Detecting AI-generated text is not a simple matter, and no detector tool can definitively identify AI-generated text. But the design of Turnitin's AI detector–which closely mirrors its plagiarism detector–may mislead some users into thinking the process is the same. Turnitin's plagiarism detector produces what we're calling a "flag report," in which specific sentences in a student's paper are flagged because they are similar to sentences in other sources. For each of these passages, Turnitin provides a link back to an original source, which the instructor can then verify. The instructor can also see an intuitive relationship between the amount of flagged text, and Turnitin's summative "similarity score." All of this allows the instructor to make an informed decision about whether a student's text is actually plagiarized.

Turnitin's AI detector includes the same elements as the plagiarism detector – the flag report and the summative score. But the flag report does not include links back to an original text, because there is no "original" text to link back to. The summative "AI-generated text" score appears to match the number of words that the tool has flagged as being AI-generated, but since there is no link-back to an original source, instructors have no way to verify if that text is in fact AI-generated. The company has

provided a description of how it evaluates text, but nothing in that description would allow an instructor to make an informed decision about whether the flagged text is actually AI generated or not. Users of the AI detection tool must simply trust the algorithm and the company that produced it.

So, should we trust the tool? We decided to test it ourselves. The Student Success Center and the Center for the Advancement of Teaching designed and conducted a test of Turnitin's AI-generated text detector.[2] Our results suggest that the tool could have value in certain limited situations. But the tool also has problems that weigh significantly against its utility. What follows is a description of our test, a discussion of the results, and our recommendations.

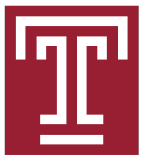
**Method:**

We compiled a set of 120 sample texts, including 30 each of the following:

- Texts entirely written by humans[3]
- Texts entirely generated by AI[4]
- "Disguised" AI-generated texts[5]
- Hybrid texts, produced partly by AI and partly by a human

We included samples of the "disguised" texts because we assume that students have heard (as we have) that it is possible to fool an AI-text detector by making superficial changes to the output produced by ChatGPT using other technologies. If a student was really determined to "cheat" with AI, and get away with it, they would probably consider running their AI-generated texts through a "paraphraser" program. Such programs make superficial changes to the surface-level features of the AI text, allegedly making it harder for AI-text detectors to get a read on the text.

Our inclusion of "hybrid" texts requires a fuller discussion because it gets at the heart of the challenges with using AI-text detectors. A student might create a hybrid text to "fool" an AI-text detector. For example, a student might make small changes to an AI-generated text–adding a sentence or two, changing the text formatting, changing certain words, adding personal observations solely for the purpose of disrupting the AI-text detectors' programming. Or they might create a hybrid text because they want to try out the tool to see what it can do, or simply for the purpose of convenience as a time and effort-saving device.



But a student might also create a hybrid text because it was required for their coursework. Many faculty are interested in teaching students how to use generative AI as a tool for writing, and they create writing assignments that are designed to teach students how to combine their own work with AI-generated work. For example, they may encourage students to use an AI tool to create an outline for a presentation or a

business plan, which the student then drafts. Or they may encourage students to use an AI tool to "edit" their essays after they have finished writing them. They may also have students engage with the tool to encourage evaluation of information the tool provides or other exercises that require critical thinking. Or they may have students use it to role-play with AI playing one of the roles, or have the tool write quiz questions that will allow students to test themselves for gaps in knowledge. Temple's stance toward the use of AI is designed explicitly to permit (and even encourage) faculty to engage in this kind of pedagogical exploration.

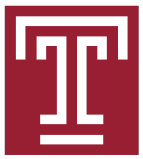
To reflect these various possibilities, we created hybrid samples that broadly fell into these three categories:

- Texts in which the ideas and organization of the document were generated by AI, but the prose was edited (minimally or extensively) by a human
- Texts in which the ideas and organization of the document were generated by a human, but the prose was created, edited, or corrected by AI
- Texts that were a compilation of ideas/sentences generated by AI and ideas/sentences written by a human, in varying ratios

Note that across these categories, there is variation not just in the *amount* of work completed by the human (student), but also in the *type* of work the student completes. This is critical because when they evaluate student work, many faculty differentiate between the quality of the ideas in the text and the quality of the prose. Consider, for example, that essay grading rubrics typically have separate points assigned to ideas and organization, on the one hand, and the "mechanics" of the paper, on the other.

We submitted all 120 samples to Turnitin[6], and we recorded the summative scores that were returned.[7] We then analyzed the summative scores within each category (human, AI, AI-disguised, and hybrid) and across the categories. Finally, we collected the flag reports for a subset of the "hybrid" texts, and we analyzed the degree to which the flagged text correctly corresponded with AI-generated parts of the text.

**Results:**



The accuracy rate of Turnitin's summative scores varied significantly across the categories of texts. Turnitin was most effective in identifying text that was 100% human generated. It was least effective in identifying texts that were "hybrid." Turnitin also had a small but consistent problem with rating texts that do not conform to certain file specifications. Seven of the 120 samples we submitted—nearly 6% of the total

samples—could not be rated because they did not meet Turnitin's file requirements. We counted all such errors as "inaccurate" responses.

- Human-generated texts:

  Turnitin correctly identified 28 of 30 samples in this category, or 93%. One sample was rated incorrectly as 11% AI-generated[8], and another sample was not able to be rated.

- AI-generated texts:

  Turnitin correctly identified 23 of 30 samples in this category as being 100% AI generated, or 77%. Five samples were rated as partially (52-97%) AI-generated. Two samples were not able to be rated.

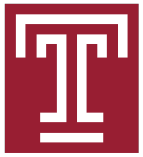
- Disguised AI-generated texts:

  Turnitin correctly identified 19 of 30 of these texts as being 100% AI generated, or 63%. Eight samples were rated as partially (43-95%) AI-generated. Three samples were not able to be rated.

- Hybrid texts:

  Determining the correctness of Turnitin's scores on the hybrid texts proved to be a challenging exercise. Technically, since the texts were all partly human-written and partly AI-generated, a "correct" score could be considered to be any score between 1-99% (in other words, not 0% and not 100%.) By that metric, Turnitin correctly identified 13 of 30 hybrid texts as being neither fully human-written nor fully-AI-generated, or 43%. Of the remaining texts, 6 were identified as 100% AI and 7 were identified as 100% human-written. One text was unable to be rated.

  However, in some cases, a technically incorrect score for a hybrid sample could be seen as a fair representation of how the sample was created. For example, one sample was written by ChatGPT and then very lightly edited by a human writer. Turnitin rated that sample as 100% AI-generated. While technically incorrect, that score is still substantially correct. The same was also true in the



opposite direction, where a student wrote a complete draft, ChatGPT proofread it, and the detector determined 0% AI contribution. This may be technically incorrect, but it feels substantially correct.

If we evaluate the scores based on whether they seem “substantially correct” (an admittedly subjective judgment), Turnitin’s ratings may look even less accurate.

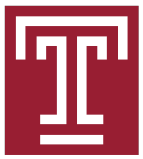
Anywhere from seven to seventeen of the samples received ratings that could be said to be “substantially correct.” That’s an accuracy rating of only 23-57%.Such judgments, of course, require knowledge of how the samples were produced, which was possible within the confines of this controlled experiment but would be impossible for most instructors using the tool to evaluate student submissions.

When we evaluated the “Flags” reports, we found no relationship at all between sections of the paper that were flagged as AI generated, and the parts that actually were generated by AI. Turnitin flagged some sentences as AI generated that were mostly or entirely human-written and vice versa.

**Discussion:**

Our tests reveal that Turnitin’s summative score does a reasonably good job of correctly identifying texts that are entirely or nearly entirely written by humans. The summative score is also somewhat effective in identifying if AI has been used to produce a text, in whole or in part. Of the 90 samples in which AI was used, it correctly identified 77 of them as having >1% AI generated text, an 86% success rate. The fact that the tool is more accurate in identifying human-generated text than AI-generated text is by design. The company realized that users would be unwilling to use a tool that produced significant numbers of false positives, so they “tuned” the tool to give human writers the benefit of the doubt.

These results suggest that Turnitin’s AI detector tool could be useful in situations where the use of AI was strictly prohibited. If the summative score indicated that the paper was partly or wholly AI generated, this would be a pretty good indicator that the student did not abide by the instructor’s policies. The summative score could also be useful in ruling out egregious misuse of AI tools. If Turnitin returns a score of 0% AI generated, these results suggest that an instructor can be reasonably certain that the student wrote all or most of the paper. But note that even in these cases, Turnitin’s summative scores can only serve as “indicators” not definitive findings. An instructor would still need to



talk with the student and consider other contextual information before making “an informed decision.”

The potential positive uses of the tool need to be weighed against some significant drawbacks that emerged in this study. First, by design or not, a 14% error rate in detecting AI generated text is not insignificant. Moreover, Turnitin’s tool is evidently fairly easy to fool. Its ability to identify AI-generated texts that had been “disguised”

(using free, widely-advertised, and easy-to-use tools) was not impressive. This means that if our primary purpose for using this tool is to “catch cheaters,” we’re going to miss a lot of them.[9]

And there are several even larger concerns with this product. The summative scores that Turnitin produced for hybrid texts were markedly inaccurate, and that problem is magnified by the fact that Turnitin’s flag reports are so inaccurate. We saw no relationship between the sentences in the text that were flagged as AI generated and the sentences that actually were AI generated. This means that instructors cannot rely on the Turnitin report to tell them either how much of the paper, or what parts of the paper, were contributed by the student. For most instructors, knowing what the student is responsible for (and in particular, knowing which ideas were generated by the student) is essential to evaluating student learning and for ensuring that students have followed the guidelines for permissible AI use stated in their syllabi.

The problems described mainly affect hybrid texts, but that is a significant concern because hybrid texts are likely to comprise a much higher percentage–perhaps the majority–of the actual papers that get submitted in the Fall semester and beyond. As faculty begin experimenting with using AI in the classroom, they will increasingly design assignments that require students to engage with AI tools *as part of* their learning process. Every one of those assignments will produce a hybrid text. [11]

---

[1] Unlike other AI-detectors, Turnitin is purchased on a university-wide license; faculty do not need to create accounts or pay fees to use it.
[2] Note that our project is narrowly focused on evaluating the efficacy of Turnitin’s AI writing detector. Although we collected and used student papers as part of the project, the papers themselves are not the focus of the project. Therefore, IRB approval was not required.
[3] These papers were all written by Temple undergraduate students before November 2022, when ChatGPT became widely available. Most of the papers were pulled from archives of student work collected for the GenEd program and the Writing-Intensive (WI) course program.

Others were class assignments that students brought to the Writing Center. The papers vary widely in length, topic, course level, and genre.

[4] We created these papers using ChatGPT. We wanted these papers to reflect the kinds of writing that undergraduate students complete for their course work, so our prompts to ChatGPT were based on real assignments created by Temple faculty. Again, many of these assignments were derived from GenEd or WI course syllabi, and they varied widely in terms of length, topic, course level, and genre.

[5] These texts are all "disguised" versions of the same AI-generated samples we created for this test. We ran half of our AI-generated texts through AI Article Spinner and the other half through Paraphraser, two free tools that purport to be able to "fool" AI text detectors by rephrasing AI-generated text.

[6] Ninety of the 120 samples were submitted to Turnitin twice, to test whether the tool provides consistent responses to the same sample. The responses were, in fact, consistent across the two submissions.

[7] Turnitin provides scores on a scale 0% - 100% AI-generated. In other words, a 0% score means Turnitin believes the text was written entirely by a human, and 100% indicates that Turnitin thinks the text was written entirely by AI.

[8] Turnitin notes that low percentages—e.g. 11% AI—are likely to be false positives.

[9] The higher accuracy rates for detecting human written text, compared to AI generated texts is by design. The company was clearly concerned about the possible harm to students who were accused of misusing AI tools, so they "tuned" the tool to give human writers the benefit of the doubt. In many of our samples, Turnitin seemed to slightly over-rate the percentage of the text that was human written.

[10] Among other things, this might involve designing and discussing during class a detailed policy describing how and when AI tools may be used by students; asking students to submit reflective process description summaries with their writing assignments; requiring students to submit complete transcripts of their AI chats along with their final assignments; requiring students to keep copies of every stage of their work (notes, drafts, etc.) in case the instructor asks to see them, etc. Link to the CAT website for more ideas.

[11] We would like to thank Jonah Chambers and Jennifer Zaylea from the Center for the Advancement of Teaching for their work in testing the writing samples.