Assessing the Efficacy of Sapling AI Content Detector as an Effective Tool for Detecting AI-generated Text from ChatGPT

Madison Farmer, Dr. Adrienne Brundage, Allison Findley, Kaylee Garcia, Macy Strain, Adam Pawson

Department of Entomology, Texas A&M University

Abstract: Through the recent rise of Large Language Models (LLMs) such as ChatGPT, academic institutions are now facing the largest source of academic dishonesty in history. New tools for the detection of model-synthesized text have emerged, promising educators security from these new forms of plagiarism, though, despite their popularity, studies have illustrated the unreliability of these tools, and some institutions have even banned the use of these detection tools due to their unreliability. This experiment aims to confirm the current state of AI-detection tools by applying AI-generated text samples, human-written text samples, and hybrid samples (those containing both AI- and human-written text) on a trial basis. These samples were applied to the Sapling AI Detector and analyzed based on their AI text detection percentile score. It was determined that, while Sapling could identify AI-generated samples with 100% accuracy, it struggled to accurately identify human-written text, with 90% resulting in a false positive. Furthermore, there was no significant difference between the human-written samples' AI detection and their hybridized counterparts when analyzing the hybrid samples.

Keywords: Artificial Intelligence, Detection Tools, Large Language Models, ChatGPT

Artificial intelligence has gained rapid and controversial proficiency through machine learning algorithms and natural language processing of large datasets to complete difficult tasks (Huynh-The et al. 2023, Weber-Wulff et al. 2023). ChatGPT, developed by OpenAI, is the first public large language model (LLM), a type of AI program designed to generate text and understanding language (Weber-Wulff et al., 2023). Large language models have been used by students at various levels to supplement research or academic writing. While there has been debate regarding ethical uses of AI in academia, most institutions agree that students using AI-generated text as their own work is considered plagiarism (Abdelaal et al. 2019). Since the digital age, academic institutions have been chasing after the innovations that offer accessible means of academic dishonesty to students. Resources like Turnitin, an online plagiarism-detection, have allowed academic institutions to compare papers to previously submitted works and detect writing similarities; after the release of ChatGPT, many of these resources have adapted to include AI detection as part of their services, and many new programs have centered their services on AI detection.

Like large language models, AI-detection tools use machine learning algorithms (Dalalah 2023). Machine learning algorithms use reinforcement learning to establish patterns between ideal outputs; in supervised learning, an algorithm is trained on data with known, established results, which helps the model reach ideal answers through pattern recognition; in unsupervised learning, an algorithm uses unlabeled data, which the AI model uses to make predictions in an uncertain environment. Through this process, AI-detection tools (including the tool in Turnitin) are constantly trained through customer input and, in theory, evolve with ChatGPT to improve their detection skills. Recent studies indicate that these AI-detection tools may lack the proper skills to detect AI-generated writing accurately, however, and these tools are widely criticized, with one study noting that even the smallest change of text can greatly alter the detection percentile of AI-generated text (Chaka, 2023). Another study determined that none of the AI-detection tools were able to detect AI-generated text with higher than 80% accuracy (Weber-Wuff, 2023). Additionally, a related study criticized automated plagiarism detectors, including text-matching software, stating that false positive results were yielded at an alarming rate (Foltýnek et al., 2020). Many organizations are criticizing the premature use of AI-detection tools that are not ready to stand the test of time and are currently doing more harm than good.

This experiment aims to contribute to the research that exists regarding AI-detection tools and their efficacy by examining a singular AI-detection tool: Sapling AI. Sapling was given samples of AI-synthesized text, human-written text, and human-written samples that had been altered with AI (known as a "hybrid" text within this study) to compare the accuracy of the AI detection results, which were analyzed for statistical significance.

Based on the general trends of AI-detection tools, it was theorized that Sapling AI would not accurately detect AI-generated text in a given sample. The null hypothesis states that the AI-detection tool cannot accurately differentiate human-written and AI-generated writing, whereas the alternative hypothesis states that the AI-detection tool can accurately differentiate human-written and AI-generated writing.

Methods

Generating the Samples

For this experiment, three categories of samples were produced: AI-generated and human-written. A third hybrid category was created using the human-written category and ChatGPT.

ChatGPT was first accessed to generate the AI-written samples. Then, the prompt "generate 20 random essay prompts for college students" was inputed into the program. Once the prompts were created, they were saved into a separate file for reference. Then, to generate the samples, each essay prompt was inserted into the prompt "generate a 1000 word essay with the following prompt: '*x prompt*.'" Each AI-

generated essay was then saved into its own document and labeled "AI samples" with a number between one and twenty. It is important to note that although the prompt was specific about the length of the output, not every sample met the criteria. Upon noticing that some documents appeared significantly longer than others, it was discovered that ChatGPT had begun disregarding this request, even if it was requested to fix this error for the same prompt. It was determined that the variance in length of these prompts would not significantly affect the experiment and that a range of 500-1000 words would be sufficient.

For the collection of the human-written samples, a nonrandom sample of colleagues, family members, and friends were gathered and requested to submit high-school or collegiate-level writing samples that were written prior to November 30, 2022. This specific threshold date was requested because this was the date that ChatGPT, the AI model used for this experiment, was released for public use. Twenty different writing samples were obtained from various Texas, Pennsylvania, and Arkansas students. Each sample was saved into its own document and labeled "Human Samples" with a number between one and twenty. While it is necessary to acknowledge the small sample size and the fact that this sample was nonrandom, the human-written writing samples varied greatly in writing format and style, with the sample set including scientific papers, creative writing assignments, personal reflections, and persuasive essays. The human-written samples also varied in length, similar to the AI-generated samples.

In order to generate the hybrid sample set, careful analysis was conducted on the human-written samples. Then, the human-written samples were separated by length; one group contained samples with five paragraphs or less, and the other group contained samples with more than five paragraphs. Then, one paragraph was removed from each sample. The second paragraph was removed from samples with five paragraphs or less, and the third was removed from samples with more than five paragraphs. The removed paragraph was then inserted into the following prompt: "reword this paragraph: '*x paragraph*'" and imputed into ChatGPT. This new AI-generated paragraph was then re-inserted into its respective sample in the same place where the original paragraph was removed. Each sample was then labeled "Hybrid Sample" with a number between one and twenty.

Treatment

Due to budget constraints, a cost-free AI-detection tool was chosen, although paid programs were noted to be more accurate than free programs. Sapling AI Content Detector was selected, as it reflected the best efficacy estimate for the given budget. Another contributing factor to the Sapling AI Content Detector selection was that it used a percentile scale to determine the extent of AI fabrication present in each sample.

The AI-generated samples were tested first as the control for the physical testing of the samples. Then, the human-written samples were tested, followed by the hybrid samples. Each sample was tested for a total of 10 trials

in the same fashion. Each sample was copied and pasted into the AI detection medium, and the detection percentile was recorded. This process was repeated for each sample ten times by deleting the inputted text from the AI detection text box and re-inserting the sample text into the text box.

Statistical Analysis

An ANOVA was performed for the dataset of both the Human and Hybrid sample sets. Because each AI-generated sample was used as a control and, therefore was detected to be 100% AI-generated text, it was impossible to conduct an ANOVA on this dataset (as this dataset has zero variance).

Results

The results of this experiment deliver controversial evidence detailing the efficacy of AI-detection tools when focusing on collegiate-level writing. When Sapling AI was tasked with analyzing each subsample of texts, the AI-generated sample group was accurately identified 100% of the time, with a sample group was zero. In contrast, the human-written sample group was frequently misidentified as AI writing, with a total mean of 25.8% AI text detection (Figure 1). The standard deviation of the human-written sample group was 33.03%, ranging larger than the average mean of the sample group itself. The hybrid samples continued to follow the expected trend and displayed an increase in AI-detected text with a total mean of 44.5% (Figure 1). The standard deviation for the hybrid sample was 29.98%, which is within the range of the total mean calculation for the sample group but is still considered a significant standard deviation.

The overall results gathered were consistent with the hypothesis regarding each sample group: the detection tool used was able to identify AI-generated text accurately but often gave false positives, claiming that human-written text was generated by AI as well. Sapling AI also struggled to accurately identify AI-generated text when it was embedded in human writing, as shown in the hybrid sample (Figure 4), as well as the wide standard deviation of the total mean (Figure 1).

Regarding the specific data gathered in the sample set of AI-generated text, the lack of variability between trial results was worth noting. Each sample yielded the same result through all ten trials, delivering a standard deviation of zero and demonstrating a significant efficiency in Sapling AI's ability to detect unaccompanied AI-generated text, (Figure 2).

The human-written trials painted a different picture (Figure 3); while AI-generated text trials were consistent with no variability, the human-written text trials varied greatly between samples, indicating a struggle for the detection tool to identify human-written writing accurately. Several of the human-written samples received next-to-zero AI detection percentiles (samples 7, 8, and 17 all received a percentile of 0.1%), but only two samples accurately received an AI detection percentage of 0% (samples 1 and 4). Many samples received high levels of AI detection, all of which were inaccurate, as each sample was written before the release of ChatGPT, the first public large-language model, on November 30, 2022.

Notable AI misidentifications include Sample 9, with a detection percentile of 100%; Sample 19, with a detection percentile of 97.6%; and Sample 20, with a detection percentile of 74.9%. Results for human-written text varied greatly, however trials of the same samples remained consistent, yielding the same result for each sample each time and resulting in an individual standard deviation of 0%. This shows that while the findings of Sapling AI's analyses were inaccurate, they were consistent. Overall, the outcome of this sample set leans toward a low but present detection rate with a few dramatic outliers.

The hybrid sample group (which consisted of the same samples from the human-written sample group with one AI-regenerated paragraph inserted into the text) yielded results consistent with the hypothesis. The general trend of this data was that most of the hybrid samples yielded a higher detection percentile than their human-written counterparts (Figure 4, Figure 5), though there were a few outliers. While most samples followed the predicted trend, some samples depicted the opposite (Sample 19), and six out of the twenty samples depicted no change at all (Figure 5). This variance illustrates the idea that AI-detection tools struggle to identify human text and differentiate what was written by a machine and what was written by a person when the two writing styles are intertwined.

Statistical Analyses Results

An ANOVA, or an analysis of variance test, was performed on each dataset to test for significance. (It was not possible to conduct an ANOVA on the AI-generated dataset, as the variance of each sample was zero.) The ANOVA conducted on the human-written and hybrid samples produced a P-value of 0.068785 and an F-ratio of 3.507911 (Table 1). For the alternative hypothesis to be supported, the P-value must be lower than 0.05. Additionally, for the F-ratio to suggest that the data gathered is not a result of chance, the F-ratio must be a value close to 1. As neither of these criteria were met, it was concluded that this data supports the null hypothesis and declares that there is no statistical difference between the AI detection percentages of the two different sample groups. Therefore, Sapling AI is unable to consistently identify the inserted AI text when intertwined with human-written text and cannot be a reliable tool to differentiate between AI and human-written writing.

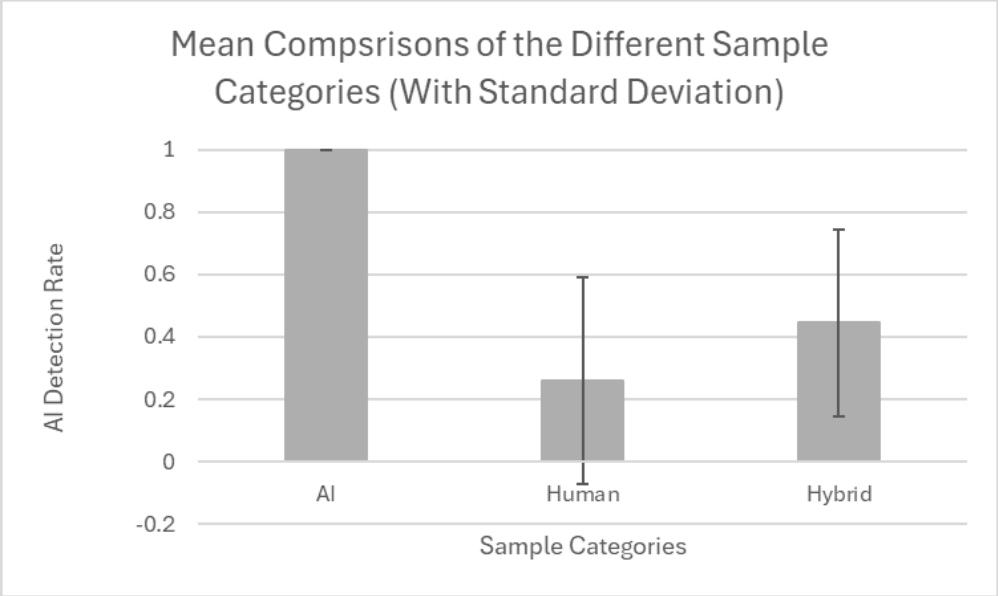

Figure 1. *Mean Comparisons of the Different Sample Categories. Illustrates the total means of all three sample groups: AI-generated, Human-written, and the Hybrid group. Error bars represent the total standard deviations of each respective category. The standard deviation of the AI-generated category was zero.*

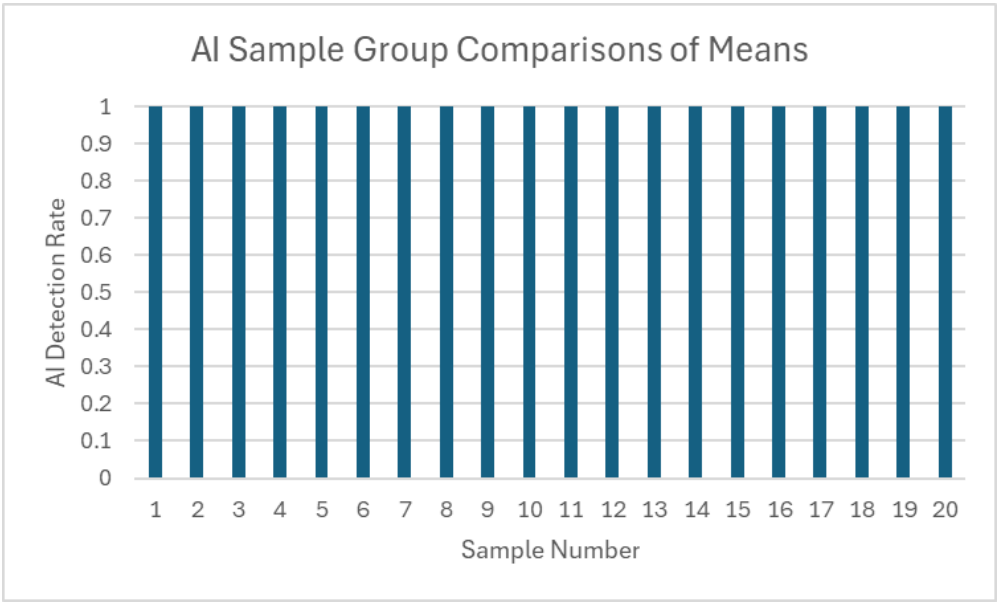

Figure 2. *AI Sample Group Comparison of Means. Visually represents the lack of variance of the AI sample group. This sample group was unable to be assessed via ANOVA due to this lack of variance.*

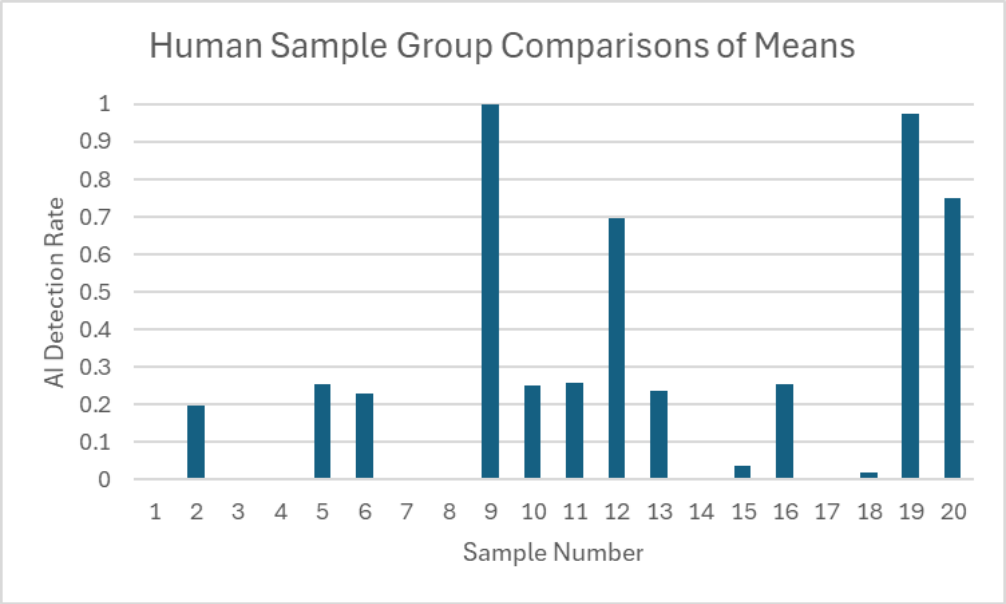

Figure 3. *Human Sample Group Comparisons of Means. Depicts the contrast of AI detection percentages among human-written samples. AI detection in this sample set ranges from 100% (depicted as '1') to 0% (depicted as '0').*

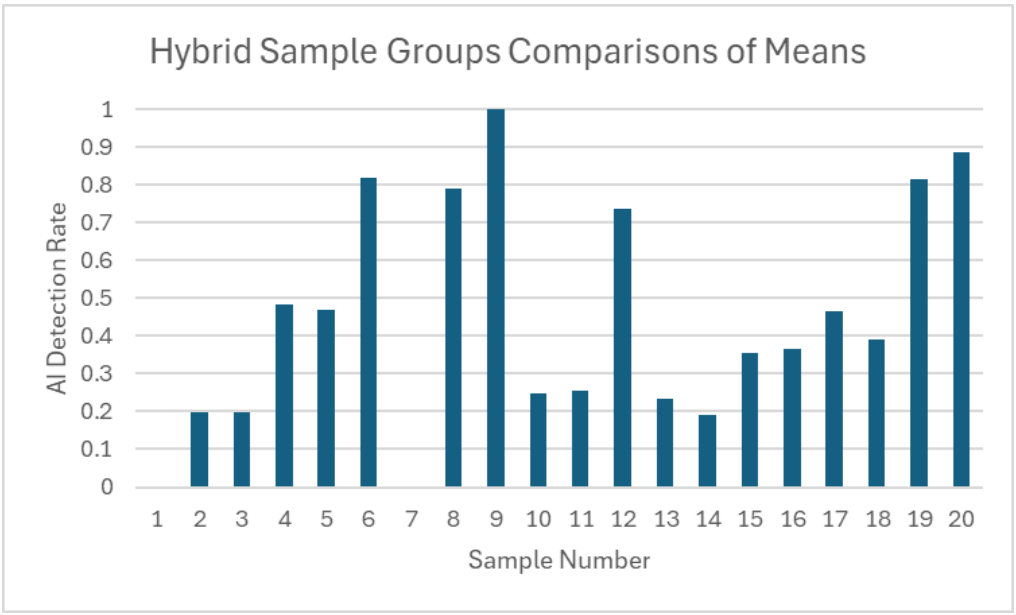

Figure 4. *Hybrid Sample Groups Comparisons of Means. Depicts the contrast of AI detection percentages among hybrid samples. AI detection in this sample set ranges from 100% (depicted as '1') to 0% (depicted as '0').*

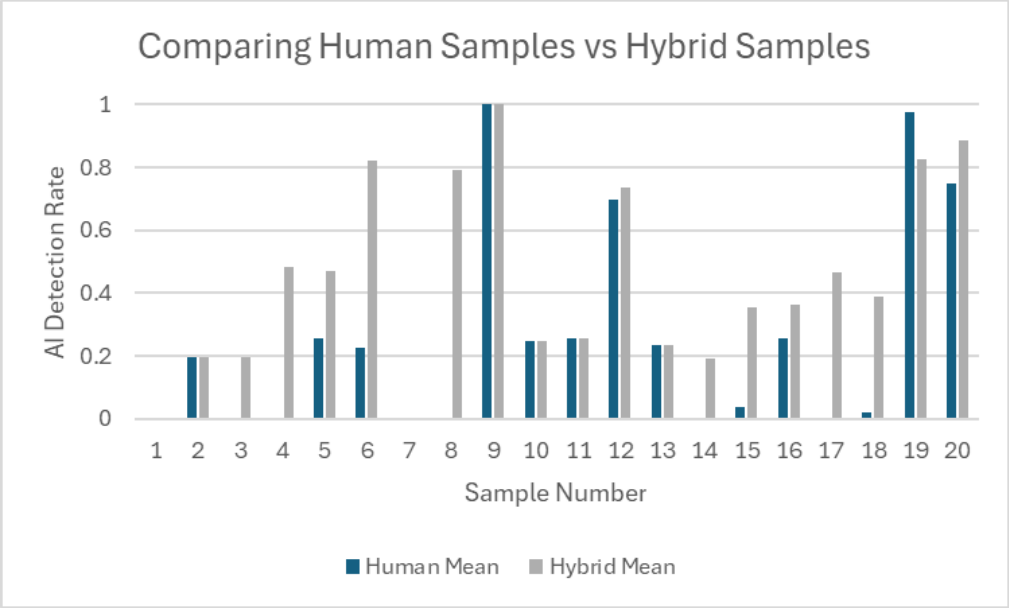

Figure 5. *Comparing Human Samples vs Hybrid Samples. It depicts a side-by-side comparison of each human-written sample and its hybridized counterpart. Again, the AI detection scale ranges from 100% (depicted as '1') to 0% (depicted as '0').*

Table 1. *ANOVA Human vs. Hybrid*

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.349129	1	0.349129	3.507911	0.068785	4.098172
Within Groups	3.781998	38	0.099526			
Total	4.131127	39				

Depicts the results of the ANOVA run between sample groups. Important data to note is the F-ratio, which is 3.507911, and the P-value, which is 0.068785. P-values over 0.05 are considered to be statistically insignificant.

Discussion

This study illuminated multiple concerns present throughout the academic community about the reliability of using AI-detection tools as an anti-cheating deterrent. It demonstrated that while the detection tool used for this study was efficient at detecting AI-generated writing on its own, it struggled to accurately identify human-written writing

or hybrid texts. The findings repeatedly demonstrated that this detection tool was likely to misidentify human-written text as AI-generated text but rarely misidentified AI-generated text as human writing.

While the detection of AI-generated text was exceptionally accurate, the detection of human-written text trials yielded different results. These findings revealed that Sapling AI consistently misidentified specific words and phrases from nearly every trial set. In fact, only two human-written samples out of twenty were correctly identified as having 0% AI-generated text. The variability in this sample group contrasts the consistency in the AI-generated text sample group, indicating that Sapling AI is much quicker to identify AI-generated writing than it is to identify human-written writing. The hybrid writing sample also showed discrepancies in the detection tool's ability to identify human-written writing. The data gathered from this sample set was variable and statistically insignificant, meaning that there was no reliable variation among the data and that the dataset gathered was likely generated by chance.

The results of the ANOVA test were determined to be statistically insignificant, thus supporting the null hypothesis that the Sapling AI-detection tool is not an effective tool to differentiate human-written text from AI-generated text due to the high likelihood of false positives (Dalalah 2023).

While previous research focuses on the accuracy of AI-detection tools regarding AI-generated writing, there are fewer efforts to apply the same theories to comparing human-written text against AI-generated text (Gao et al., 2022). Additionally, many studies aim to approach each detection tool with similar criteria for "accurate" and "inaccurate" findings, even though each AI-detection tool utilizes different programming and analyzing techniques to reach conclusions (Aremu, 2023). For instance, many detection tools (including Sapling AI) use a percentile scale to determine how much of a given sample was likely to have been written by an AI model. Other detection tools, such as OpenAI, use other methods, such as determining a trademarked scale of "likely", "possibly", and "unlikely", as well as other descriptors that do not transcribe easily into quantitative data (Aremu 2023).

Regardless of these differences among studies, the findings of this experiment are consistent with previously published research. The lack of success of the AI-detection tool's ability to differentiate AI writing from human writing in the hybrid sample set has been previously noted, as well as the stark success of the AI-generated sample set's detection rate (Aremu, 2023).

There are several limitations of this study that must be taken into account when processing this new information. First, this study was conducted with a nonrandom sample of less than twenty college students participating in providing samples for the human-written samples. Because of this, it may be likely that the level of writing, the writing style of a particular participant, or a geographical grammar trend may alter the true reliability of the detection tool used once applied to the

population. Another limitation of this study is that the experiment only focused on the abilities of one AI-detection tool (Sapling AI). While Sapling AI is a well-known and widely used detection tool, there are many other competitors that may yield different results. This research may be used as supplementary material to suggest inaccuracies in AI-detection tools as a whole, though it should not be used to concretely disprove the effectiveness of any detection tool outside of Sapling AI. From this point in the study, the protocol may be applied to any AI-detection tool that utilizes a percentile scale to analyze text samples. This study may also provide special insight into the intersection of academia and technology. While it is important that instructors have methods of preventing plagiarism in their classrooms, those tools must be accurate. This research may be built upon allowing researchers to better understand the current faults with AI-detection tools and build reliable tools for this purpose in the future.

Acknowledgments: Special thanks to Dr. Adrienne Brundage for advising the creation of this experiment. Thank you to Blake Hyatt, Lucy Kainer, Lara Amiouny, Sara Farmer, Macy Strain, Allison Findley, Jennifer Itson, and Elizabeth Ventura for providing samples for the human-written sample set.

References

Thien Huynh-The, Quoc-Viet Pham, Xuan-Qui Pham, Thanh Thi Nguyen, Zhu Han, Dong-Seong Kim, Artificial intelligence for the metaverse: A survey, Engineering Applications of Artificial Intelligence, Volume 117, Part A, 2023, 105581, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2022.105581.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S. et al. Testing of detection tools for AI-generated text. Int J Educ Integr 19, 26 (2023). https://doi.org/10.1007/s40979-023-00146-z

ChatGPT. (2024). https://chat.openai.com/ Accessed 11-18 March 2024.

Chaka, Chaka. View of Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools | Journal of Applied Learning and Teaching. (2023). https://journals.sfu.ca/jalt/index.php/jalt/article/view/861/621

Foltýnek T, Dlabolová D, Anohina-Naumeca A, Razı S, Kravjar J, Kamzola L, Guerrero-Dib J, ÇeIlik Ö, Weber-Wulff D (2020) Testing of support tools for plagiarism detection. International Journal of Educational Technology in Higher Education, 17(1): 1–31. doi:10.1186/s41239-020-00192-4.

Gao, C.A., Howard, F.M., Markov, N.S. et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. npj Digit. Med.6,75(2023). https://doi.org/10.1038/s41746-023-00819-6

Aremu, Toluwani, Unlocking Pandora's Box: Unveiling the Elusive Realm of AI Text Detection (June 6, 2023). Available at SSRN: https://ssrn.com/abstract=4470719 or http://dx.doi.org/10.2139/ssrn.4470719