# Efficacy of Detecting AI Plagiarism in Higher Education

**Henry Collier**
Marshall University,Trenton, NJ USA

hcollier@tesu.edu

**Abstract:** Whether the text has been directly copied from a source and presented as the author's work, or if the author poorly paraphrased their source, the result is plagiarism. Plagiarism stems from a desire to get a good grade. It is the result of being sloppy, not having confidence in one's ability to paraphrase correctly, a lack of understanding of what needs to be cited, thinking they are supposed to reproduce what the experts have said rather than synthesize the expert's opinions, they panic as the deadline approaches and of course they are lazy. Plagiarism is not new to academia and detecting it has been made easier with plagiarism-detecting tools and simply conducting effective searches on the Internet. However, new plagiarism methods are being developed as technology grows, which are significantly harder to detect. Students now use artificial intelligence to write their academic papers. Detecting if AI was used in producing an academic paper is difficult because the paper has been properly paraphrased and/or cited. New tools like GPTZero and ZeroGPT have been developed to address this issue, but these tools are untested, and academic faculty don't want to pin a plagiarism charge on a student based on an untested tool. This study looks to address this issue and determine if either or both tools effectively determine whether AI has written something.

**Keywords:** Plagiarism, Artificial Intelligence, ZeroGPT, GPTZero

## 1. Introduction

Plagiarism within higher education is something that educators have struggled with for decades. Historically, in order to detect plagiarism, the teacher needed to have first-hand knowledge of the sources or go through each source painstakingly to see if plagiarism existed or not. This was a time-consuming endeavor, one that most educators felt was necessary to ensure student success. The Internet brought speed and efficiency to this process by allowing teachers to input suspected plagiarism examples into a search engine, and then within moments, sources that were potentially used will be returned, and suspected text will be highlighted. Using the Internet increases the efficiency of detecting plagiarism by reducing the time a teacher needs to commit to plagiarism detection.

Along with solving problems, technology can also create new ones. As Artificial Intelligence (AI) continues to advance, educators are starting to see more and more papers being produced that are highly likely to be written by AI and not by the student (King, 2023) (Francke & Alexander, 2019) (Hutson, 2024). Based on the simplest definition of plagiarism, having AI write a student's paper qualifies as plagiarism because the work is not the student's original work (Khalil & Er. 2023). When checked, through either traditional methods or an Internet search engine, papers that AI has written rarely return results demonstrating plagiarism (Dien, 2023). This is because AI does a great job of paraphrasing the data it is using to write the paper and, therefore, does not provide direct results (Mahajan, et al., 2023)(Kleebayoon & Wiwanitkit, 2023). With the increased availability of AI tools like ChatGPT, Microsoft Copilot, CopyAI, Jasper AI, and even Grammarly, more and more student papers appear to be heavily influenced by AI (Mahajan, et al., 2023).

Traditionally, one of the ways that a professor would suspect plagiarism is because the suspect paper would not be reflective of the author's previously submitted writing samples (Alzahrani, et al., 2012). The suspect paper will frequently appear to have been written by someone with significantly more education than the author (Alzahrani, et al., 2012). It further must be noted, that as classroom sizes grow, it will become more difficulte for faculty members truly get to know their student's writing style and detecting suspected plagiarism will become more difficult. Example 1 below is an AI-written sentence based on the human-written sentence in example 2(Leki & Carson, 2012).

- Example 1

*A person's social media habits and online presence are critical factors that significantly enhance the success of social engineers.*

- Example 2

*A person's social media habits and their web presence is a significant factor in social engineers being successful.*

Example 1 is clear and precise while avoiding redundancy and uses terms like 'critical factors' and 'enhance', which are direct and effective when conveying the topic. While example 2 is slightly less clear, the term 'social

engineers being successful' is vague. Example 1 is stronger and better written, and if the writer submitted this as their assignment after submitting samples more in line with example 2, it would be evident that there is a high likelihood of plagiarism.

AI has created a situation where educators need to work harder to determine whether a student's submission is their own work (Dien, 2023) (Hutson, 2024). From an educational perspective, ensuring students are doing the work themselves is important, as it is the best way for them to learn and become effective in their chosen industry. Certainly, educators need to teach them how to use AI as a tool, but they also need to ensure they teach the students how to use it ethically. In order to ensure we are holding students to the standards, there needs to be AI detection tools that provide results with a high efficacy rate. The two most common tools are ZeroGPT and GPTZero. This study aims to determine if either or both of the tools are effective at determining if a writing sample has been written by AI or by a student. The hope is that this tool will give educators an additional resource that can be used to support student success and maintain academic integrity.

## 2. Methodology

This study is a comparative analysis comparing documents that are proven to have been written by humans and AI. The human component is broken into two sections: Native English speakers and English as a second language (ESL). Within higher education, there is concern that students whose first language is not English will have a higher false positive rate when their work is submitted to the two detection tools. This concern is because individuals who have learned English as a second language tend to write more formally, using less slang, and contractions. Resulting in writing that is less casual and more formal.

Furthermore, ESL speaker's use of articles and literal translations differs from that of native English speakers (Leki & Carson, 2012). For example, an ESL speaker might say, "I went to hospital," while a native English speaker would say, "I went to a hospital" or "I went to the hospital." Additionally, someone who is an ESL might say "I am making a photo" while someone who is a native speaker would say "I am taking a photo." These are just a couple of examples of how someone who is a native English speaker will write differently than an ESL speaker. These differences in sentence structure and word choice could lead the AI detection tool to determine that an ESL speaker's work has been written by AI.

This study consists of 150 papers, 50 from each category. The papers associated with a human were selected based on their publication date to ensure they could not have been influenced by AI. All human-written papers were published or written before November 2022, when ChatGPT was first released to the public. This does not count for Apple's Siri, Google Now, and Microsoft Cortana, which were introduced on smartphones between 2011 and 2014. It further does not account for early forms of AI that were not available to the general public but could have been available to researchers. Although there is the possibility that these tools may have influenced writing, at the time they were introduced, they could not write full papers but rather provide a response to a specific question.

Portions of each paper were submitted to the two tools, and the results were recorded and analyzed. ZeroGPT returns a response of how much of the writing was likely written by AI while GPTZero broke the results into three categories—AI written, mixed and human. For this study, only the percentage of AI was considered to ensure the two tools could be compared equitably.

## 3. Results

The results were statistically analyzed using an unpaired two-tailed T-test and a standard deviation assessment. Furthermore, a comparative analysis was completed using whether or not something was 100% AI-written or 100% human-written. Although the goal would be for the tool to 100% of the time identify the source of a writing sample correctly, it must be noted that this is a game of probability, and therefore, other values must be considered. For example, is there a return rate that is not 100% that demonstrates an extremely high probability of being written by AI or not? One of the goals of this study is to identify what the lowest probability value asserts whether or not AI likely wrote something.

**Table 1: Standard deviation-human, English as first language.**

| | | ZeroGPT | GPT Zero |
|---|---|---|---|
| **Standard** | **s =** | **0.269034** | **0.0587367** |
| Variance | $s^2$ = | 0.072379 | 0.00345 |
| Count | n = | 50 | 50 |
| Mean | $\bar{x}$ = | 0.1696 | 0.041 |
| Sum of Squares | SS = | 3.546592 | 0.16905 |

**Table 2: Standard deviation-human, English as second language.**

| | | ZeroGPT | GPTZero |
|---|---|---|---|
| **Standard Deviation** | **s =** | **0.10683288** | **0.19667688** |
| Variance | $s^2$ = | 0.011413265 | 0.038681796 |
| Count | n = | 50 | 50 |
| Mean | $\bar{x}$ = | 0.031 | 0.0628 |
| Sum of Squares | SS = | 0.55925 | 1.895408 |

**Table 3: Standard deviation-AI written.**

| | | ZeroGPT | GPTZero |
|---|---|---|---|
| **Standard Deviation** | **s =** | **0.26140086** | **0.2037454** |
| Variance | $s^2$ = | 0.06833041 | 0.0415122 |
| Count | n = | 50 | 50 |
| Mean | $\bar{x}$ = | 0.895542 | 0.9002 |
| Sum of Squares | SS = | 3.3481901 | 2.034098 |

The standard deviation between ZeroGPT and GPTZero demonstrates the significance of the results of this experiment. All the standard deviation rates are less than 1%, which means the results demonstrate that the tools assess the paper with a certain level of preciseness and equity across the spectrum of papers. However, it is important to note that preciseness and equity do not equate to accuracy but rather consistency. It can be said that these tools evaluate writing samples in a consistent manner.

When the data sets are processed through a two-tailed P-test with a pre-determined significance (alpha) level of .05, some of the results are statistically significant, while others are less significant. Tables 4-6 show the results of the two-tailed P-test conducted on the results of this experiment.

**Table 4: Two-Tailed P-Test, English as a first language.**

| | |
|---|---|
| T-value | 3.30 |
| P-value | 0.001300 |
| df | 98.00 |
| Standard Error of Difference | 0.04 |

By conventional criteria, the results for the English as a first language category are very statistically significant. The extremely low p-value suggests that the results are unlikely to have occurred by chance. When the results for the English as a second language and AI written categories p-values are evaluated against the pre-determined alpha value, they are considered not statistically significant.

**Table 5: Two-Tailed P-Test, English as a second language.**

| | |
|---|---|
| T-value | 1.00 |
| P-value | 0.32 |
| df | 98.00 |
| Standard Error of Difference | 0.032 |

**Table 6: Standard deviation-AI written.**

| | |
|---|---|
| T-value | 0.09940 |
| P-value | 0.92100 |
| df | 98.00000 |
| Standard Error of Difference | 0.047 |

When the results are evaluated from a structural perspective, it is clear that neither tool is 100% effective. GPTZero identified 21 papers that were written by a human whose first language was English as having been fully or partially written by AI. This results in a 42% misidentification rate. On the other hand, ZeroGPT demonstrated that AI had fully or partially written 49 papers resulting in a 98% misidentification rate. When a human wrote the papers, and English was their second language, the results showed that GPTZero identified four papers as having been partially or fully written by AI, which is a less than 8% misidentification rate. While ZeroGPT identified 42 papers as having been partially or fully written by AI, which equates to an 85% misidentification rate.

Regarding the AI-written papers, ZeroGPT flagged 18 papers as having been partially or entirely written by a human, and GPTZero flagged 14 papers. Since the AI-written papers were created by this author using ChatGPT, they are known to be 100% AI-written papers with no human intervention. This means that ZeroGPT had a 36% misidentification rate, while ZeroGPT had a 28% misidentification rate. Table 7 shows the misidentification rates side by side.

**Table 7: Misidentification Rates**

| | **Zero GTP** | **GPTZero** |
|---|---|---|
| Human English First Language | 98% | 42% |
| Human English Second Language | 85% | 8% |
| AI Generated | 36% | 28% |

Since the AI-written papers are confirmed to have been written by AI, it was expected that the tools would have identified 100% of the papers as having been written by AI, with no human intervention. However, it is feasible that the tools might have misidentified a small number as having had human influence. If the results were 1% or less, it could be determined as accurate and reliable. Since the results are both above 25%, this significantly reduces the reliability rate of these tools.

If the percentage of misidentification is similar across all three groups, with a degree of separation that is less than 5%, an educator could possibly use the tool to successfully determine the probability that AI wrote the submitted work. However, even in this case, the professor should not rely solely on the tools to determine whether plagiarism has occurred. It is better to err on the side of caution than to improperly accuse a student of cheating.

## 4. Future Work

This study shows that there is a need for more work on this subject. A future study is being designed that will incorporate a larger pool of papers that were all written by humans before 2000, which should further guarantee that no AI influence could be possible. There might be some difficulty in this next study when it comes to determining if English is a first or second language. Since the results of this study did not show a significant disadvantage to individuals where English is a second language, the next study may not split the categories apart. The hypothesis would be that English as a second language is not a factor in determining whether or not AI has been used in writing the paper. This would result in a more holistic approach to the study. Furthermore, it would be prudent to rerun this test every six months as the two tools continue to be refined.

## 5. Conclusion

This study shows that more work needs to be done so that educators can rely on the results of these tests when determining plagiarism. The fact that the AI detection tools failed to identify 100% of the AI-written papers as having been written by AI shows that the metrics used by the tools need to be adjusted. It is not unreasonable for the tools to misidentify some of the papers written by both categories of human submissions, but it should have been able to determine the AI papers were 100% AI-written. It is clear that educators cannot use these tools to effectively determine if a paper has or has not been written or is heavily influenced by AI tools like ChatGPT. Educators should continue using other sources to determine if a paper has a high probability of being written by AI or not.

## Ethics Declaration

This paper used secondary data which did not require ethics clearance.

## AI Declaration

This paper did not make use of AI to write content; Covidence was used as a tool to extract and organize data for thematic analysis.

## References

Alzahrani, S. M., Salim, N. & Abraham, A., 2012. Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(2).

Khalil, M & Er, E. 2023. Will ChatGPT get you caught? Rethinking of Plagiarism Detection. Copenhagen, HCI International 2023.

Dien, J., 2023. Editorial: Generative artificial intelligence as a plagiarism problem. Biological Psychology, Volume 181.

Francke, E. & Alexander, B., 2019. The Potential Influence of Artificial Intelligence on Plagiarism: A Higher Education Perspective. Oxford, Academic Conferences, and Publishing-International.

Hutson, J., 2024. Rethinking Plagiarism in the Era of Generative AI. Digital Commons @Lindenwood University.

King, M. R., 2023. A Conversation on Artificial Intelligence, Chatbots, and Plagiarism in Higher Education. Volume 16, pp. 1-2.

Kleebayoon, A. & Wiwanitkit, V., 2023. Artificial Intelligence, Chatbots, Plagiarism and Basic Honesty: Comment. Cellular and Molecular Bioengineering, Volume 16, pp. 173-174.

Leki, I. & Carson, J., 2012. “Completely Different Worlds”: EAP and the Writing Experiences of ESL Students in University Courses. Tesol Quarterly.

Mahajan, A., Anshika, Sharma, A., Sharma, N., Kaur, A., 2023. Challenges to Plagiarism Detection and the Usof Voice Assistants in Everday Life and Education. s.l., Congress on Smart Computer Technologies: Proceedings of CSCT.