Preprints.org

Article

Not peer-reviewed version

AI-Generated Text and Plagiarism Detection: Pandora's Tech-Box Unmasked

Edgar Eslit [*]

Posted Date: 2 January 2025

doi: 10.20944/preprints202501.0060.v1

Keywords: AI-generated text; false positives; Pandora's Tech-Box; plagiarism detection

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2025 doi:10.20944/preprints202501.0060.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

AI-Generated Text and Plagiarism Detection: Pandora's Tech-Box Unmasked

Edgar R. Eslit

St. Michael's College of Iligan, Inc.; edgareslit@gmail.com or e.eslit@my.smciligan.edu.ph

Abstract: The use of AI-text generators and plagiarism detectors have made a big challenge to academic community because of the growing occurrence of "false positives" in student's academic output. These mistakes, where original student work is mistakenly flagged as AI-generated or plagiarized, can have damaging effects on students' academic performance and emotional well-being. This research article explores into the root causes of these "false positives", examining the limitations of current AI-generated text and plagiarism detection tools with their impact on the educational context. Drawing on the educational theories like constructivism and cognitive load theory, the researcher promotes for a more nuanced approach that balances the potential benefits of AI with the crucial role of human judgment. Using qualitative content analysis and a critical examination of existing literature, the paper proposes a framework for reasonable and more effective AI-assisted or text plagiarism detection. This includes developing more sophisticated AI models that better understand context and language, fostering human intervention in the evaluation process, and updating policies based on the paper's suggestions to ensure that AI will serve as a supportive tool rather than a retaliatory one. Overall, this article emphasizes the importance of a human-centered approach to AI in education, where technology enhances, rather than hampers, the learning and assessment process.

Keywords: AI-generated text; false positives; Pandora's Tech-Box; plagiarism detection

I. Introduction

Artificial intelligence has revolutionized many aspects of education, from personalized learning platforms to automated grading systems. Among these innovations, AI-powered plagiarism detection tools have become commonplace, offering educators a seemingly efficient way to maintain academic integrity. Platforms like Turnitin and Grammarly leverage sophisticated algorithms to analyze student submissions and identify potential instances of plagiarism or AI-generated text.

Consequently, this technological advancement has also introduced a new set of challenges. A significant concern lies in the increasing occurrence of "false positives," where original student work is mistakenly flagged as plagiarized by these AI systems. The complexities of human language – the subtle nuances of style, the creative use of sources, the unique voice of each individual – often elude these algorithms, leading to inaccurate and potentially damaging results.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2025 doi:10.20944/preprints202501.0060.v1

Try to imagine a student pouring their heart and soul into a creative essay, only to see it flagged as plagiarized by an AI system. This not only undermines their hard work and intellectual effort but also erodes their trust in the assessment process. Furthermore, the widespread reliance on these automated systems, often without critical human review, can lead to miscarriages of justice, unfairly penalizing students for work that is entirely their own.

Here, the paper aims to investigate the root causes of these false positives, examining the limitations of current AI algorithms and their impact on the delicate balance of academic integrity and student well-being. We delve into the complexities of human language and explore how these complexities are often overlooked by current AI-driven detection methods. By analyzing existing practices and identifying key areas for improvement, this paper seeks to propose a more nuanced approach to plagiarism detection – one that combines the power of AI with the wisdom and critical judgment of human educators.

To address this serious subject, the paper explores the following key questions:

(1) What factors contribute to the high rate of false positives in AI-driven plagiarism and AI-generated text checkers detection systems?

(2) How do these "false positives" impact student confidence, motivation, and overall academic performance?

(3) What innovative strategies can be implemented to improve the accuracy and fairness of AI-driven plagiarism detection while safeguarding academic integrity?

This research article endeavors to move beyond simply identifying the problem of "false positives". The paper aims to contribute to a more nuanced understanding of the challenges and opportunities presented by AI in education (an ultimate gap needing attention), ultimately fostering a learning environment that supports both academic integrity and student accomplishment.

II. Literature Review

The artificial intelligence (AI) has profoundly transformed the educational landscape, and plagiarism detection and or AI-generated text notion is no exception. While AI-powered tools offer a valuable resource for maintaining academic integrity, their increasing sophistication has also introduced new challenges. Hence, this literature review examines the evolution of AI-driven plagiarism detection, focusing on the challenges of false positives, AI-generated text, the advancements in AI technology, and the crucial ethical considerations arising from their widespread use in academic settings.

The early AI generated text tools and plagiarism detection systems relied on basic text-matching algorithms, essentially searching for verbatim copies within vast databases. However, as the digital world expanded, these rudimentary methods proved inadequate. To address this, researchers began incorporating more sophisticated techniques, such as natural language processing (NLP) and machine learning, into these tools. This shift allowed for a deeper analysis of text, moving beyond simple keyword matching to identify more subtle forms of plagiarism, including paraphrasing and rephrasing (El Mostafa & Benabbou, 2020). Today, leading platforms like Turnitin, Grammarly, etc., utilize these advanced AI techniques to analyze student work, identifying potential instances of plagiarism and even AI-generated text detection.

In spite of these advancements, a significant concern remains: the high rate of "false positives." These occur when original student work is mistakenly flagged as plagiarized if not AI-generated, often due to the inherent limitations of AI algorithms in understanding the nuances of human language. Imagine a student meticulously crafting an essay or research paper, drawing inspiration from various sources but expressing their ideas in their own unique voice. An AI system, however, might misinterpret their original phrasing as plagiarism or AI generated, leading to unwarranted accusations and potentially damaging consequences for the student. This not only undermines student confidence but also erodes trust in the academic assessment process.

Luckily, continues research is driving advancements in AI-driven plagiarism detection. Machine learning models are now trained on massive datasets, enabling them to better understand context, tone, and writing style. This allows for more nuanced analysis, reducing the likelihood of flagging legitimate original work as plagiarized (Albadarin et al., 2023). Deep learning models, for example, can now effectively detect paraphrased content, a feat that previously eluded many detection tools (Bowen & Watson, 2022).

Moreover, the prevalent use of AI in plagiarism and AI-generated text detection also raises substantial ethical concerns. One primary concern is the potential for bias. If the training data used to develop these AI systems reflects existing societal biases, the resulting algorithms may inadvertently discriminate against certain groups of students, such as those from diverse linguistic or cultural backgrounds (Mishra, 2024). Furthermore, the collection and use of student data for plagiarism and AI-generated text detection raise important privacy concerns. Ensuring the security and ethical use of this sensitive information is paramount.

This review of related literature underscores the critical need for a balanced approach to AI in plagiarism and AI-generated text detections. While these tools offer valuable assistance in identifying potential issues, they should never be considered the final arbiters of academic integrity. Human judgment remains crucial. Educators must critically evaluate the AI-generated reports, considering the unique context of each student's work and using their expertise to make informed decisions. By combining the power of AI with the wisdom and experience of human educators, institutions can create a more equitable and effective system for maintaining academic integrity in the digital age.

III. Theoretical and Conceptual Framework

Looking at the changing educational landscape, AI is significantly changing how people teach and learn, especially when it comes to spotting plagiarism in schools and universities. While AI tools have become crucial for finding potential plagiarism and AI-generated text, it is important to look at the theories and ideas behind how these tools are used. This section explores key educational theories, the concepts surrounding plagiarism detection, and provides a framework for analyzing those false accusations by AI-based text and plagiarism checkers, which helps to further show how AI is used in education today.

Theories on education play a crucial part in guiding how AI is used in learning. One key theory is constructivism, which says that learning is an active process where students build their understanding through experiences. In this way, AI tools can act as a helpful guide, giving students the tools and feedback they need to improve their knowledge and skills. These tools help students engage with their learning more meaningfully by adjusting to individual needs, with AI systems constantly customizing the learning process to fit the student's pace and learning style (Barrett & Pack, 2023). However, how people use AI in education must also consider cognitive load theory, which says that students learn best when they are not overwhelmed by too much information. Cognitive load theory helps in the design of AI tools by making sure they do not overload students, allowing them to focus on learning without excessive effort. When AI tools are used well, they can reduce cognitive load by automating repetitive tasks like AI-generated text and plagiarism checking, allowing students to focus on deeper learning activities, such as critical thinking and analysis (Dehghantanha & Choo, 2024).

The core concept of AI's role in education is AI-generated text and plagiarism detections, which used to just focus on finding when students copied someone else's work word-for-word. But as academic practices have changed, so have the kinds of plagiarism institutions need to find. New forms, like self-plagiarism (where students use their own old work) and paraphrasing (where text is changed but still closely follows other sources) are now harder to detect with old methods (Smith & Doe, 2024). AI tools have made big strides in meeting these challenges. Traditional plagiarism tools compare student work against a database of existing sources by just matching strings of text, but AI-based tools now use natural language processing (NLP) and machine learning to find reworded content and more complex forms of plagiarism (Quidwai, Li, & Dube, 2023). These AI systems go beyond simple text matches, analyzing the meaning and structure of the text, which makes them better at detecting patterns of academic dishonesty. As AI tools get more sophisticated, they allow for more complex analysis, which expands what can be detected and how accurately. One need not wonder why his/her name and that of his/her school is even be marked or flagged as an AI-generated

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2025 doi:10.20944/preprints202501.0060.v1

text if not a plagiarized one once his/her work will undergo a plagiarism scan by the teacher because of these existing AI scanning tools, which, by the way, their fixed algorithms are made for this purpose.

Despite the good view that AI brings to AI-generated text and plagiarism detections, problems still remain, especially when it comes to false positives. These mistakes happen when the AI incorrectly flags content as AI-generated or plagiarized, even when it is an original work. This can cause big issues, as students can face unfair penalties for work they have not copied. It can be gleaned for other studies show these mistakes can negatively impact students' well-being and academic performance (Harrison & Quarterman, 2024). The causes of false positives are numerous, from the limits of current AI models to the complexities of language. For example, rephrasing or using common phrases can trigger false alarms, as AI tools may misinterpret these instances as plagiarism. It is important to consider how AI systems are trained and the data they use to detect plagiarism. Likewise, these schemes must be constantly improved to lessen the blunders and make sure that all academic assessments are fair enough (Johnson & Nguyen, 2019).

Finding out the complexities of these false positives, it helps to have a framework that combines theories about AI in education and how people understand plagiarism. This framework clarifies how AI can be used as a tool in the detection process without being the final decision-maker. Instead, AI should be a resource that flags areas of concern for expert educators to review (Smith & Doe, 2024). This approach allows for a more nuanced understanding of plagiarism, considering the context of each student's work and the specifics of any flagged content. By using AI alongside expert teachers' knowledge and experience, this framework promotes fairness and accuracy in plagiarism detection, lowering the risk of false positives while maintaining academic honesty.

Presumably, this theoretical and conceptual framework offers a vivid view on how AI tools, plagiarism detection, AI text generated, and the potential for false positives are all related. It provides a structured way to understand how AI can help with education without sacrificing fairness, emphasizing the importance of human judgment in assessing academic work. In the succeeding sections of this paper, the researcher explores how these principles can be applied to tackle the challenges of "false positives".

IV. Methodology

To structure the writing process, the researcher employs a qualitative approach, specifically content analysis, to investigate the complex phenomenon of false positives in AI-driven plagiarism detection systems. Unlike quantitative methods that focus on numerical data, content analysis allows for a deeper exploration of the underlying causes and consequences of these errors. By meticulously examining existing scholarly literature, the paper aims to uncover the nuanced experiences of students and educators, the ethical dilemmas surrounding these technologies, and the limitations of current AI algorithms.

The author made use of a comprehensive body of scholarly literature, prioritizing credible and highly-cited sources. This involves a systematic review of academic literature, including a thorough search of reputable databases such as Web of Science, Scopus, and Google Scholar, as well as consulting the SMCII collection and other academic libraries. This research prioritizes articles from peer-reviewed journals, books, and conference proceedings, ensuring that the findings are grounded in rigorous academic inquiry. Furthermore, this research article incorporates a wide range of viewpoints by including articles that utilize both qualitative and quantitative research methods, providing a comprehensive understanding of the issue.

Moving forward, the thematic analysis is then employed to identify key themes and patterns within the saturated collected data. This involves a meticulous process of closely examining the literature to identify recurring themes, such as the ethical implications of false positives, the impact on student well-being, and the limitations of current AI algorithms. The coded data is then organized and synthesized into meaningful categories of themes to develop a comprehensive understanding of the research problem. Finally, the identified saturated themes are analyzed to draw meaningful

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2025 doi:10.20944/preprints202501.0060.v1

conclusions about the causes, consequences, and potential solutions to the problem of false positives in AI-driven plagiarism detection.

Ensuring the validity and reliability of the findings, the paper incorporates several strategies. Triangulation, utilizing multiple sources of data, including research articles, case studies, and published expert reports, is employed to cross-validate findings and ensure a comprehensive understanding of the issue. Additionally, this work seeks feedback from academic peers and experts in the field to enhance the rigor and validity of the research findings through peer review.

Observing these rigorous research methods, the paper aims to provide a comprehensive and insightful analysis of false positives in AI-driven plagiarism detection, contributing to a deeper understanding of this critical issue and informing the development of more equitable and effective solutions for the future.

V. Results and Discussion

After completing the current landscape of AI-driven and plagiarism detections and their inherent challenges, the researcher delves deeper into the realities of this evolving technology. This discussion examines real-world scenarios where AI-generated accusations have impacted students and institutions, explores innovative technological solutions that can mitigate these challenges, and critically analyzes the ethical considerations that must guide the responsible integration of AI into academic integrity assessments. By examining these facets, the researcher aims to gain a deeper understanding of the complexities surrounding AI-generated text and plagiarism detections and proposes a path forward that balances the power of technology with the importance of human judgment and ethical considerations.

A. Reflective Case Studies

As reflected in the volumes of literature about AI and its use in education, the rise of AI-driven plagiarism and AI-generated text detection tools has revolutionized how academic integrity is maintained in educational institutions. However, this technological advancement has also introduced a significant challenge: the occurrence of "false positives," where original student work is mistakenly flagged as plagiarized if not AI-generated. These mishaps, while seemingly technical in nature, have profound implications for students, educators, and the overall integrity of the academic process. Examining real-world case studies provides valuable insights into the limitations of current AI systems and highlights the crucial role of human judgment.

In-depth reading unveils illustrative example that involved a big university in the Northeastern United States that heavily relied on an AI plagiarism checker. Several student essays, despite demonstrating original thought and proper citation, were flagged as plagiarized. Upon closer inspection by human reviewers, it became evident that the flagged content often consisted of common academic phrases or properly cited sources, showcasing the limitations of AI in accurately differentiating between legitimate academic writing and instances of plagiarism. This case study underscores the importance of human oversight, emphasizing that AI should serve as a tool to assist, not replace, human judgment in academic integrity assessments. This was emphasized in the study conducted by Patel (2024).

On the same token, the article by Merod (2023) showed that students' essays were mistakenly flagged as plagiarized by an AI tool due to the presence of common academic language. This scenario reflects a common occurrence in many educational settings. Despite the student's original work, the AI system, unable to fully grasp the nuances of academic writing, flagged the essay due to the presence of widely used phrases and concepts. This highlights the crucial need for AI systems to be trained on diverse datasets and to be sensitive to the complexities of human language and academic discourse.

Looking at these situations, many institutions all over the world have implemented strategies to mitigate the impact of false positives. One common approach involves establishing formal appeal processes, allowing students to challenge AI-generated accusations. For example, a university in the Midwestern United States, along with many other institutions globally, has implemented a review

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2025 doi:10.20944/preprints202501.0060.v1

process where faculty members carefully examine flagged content, providing students with an opportunity to defend their work. This human-centered approach ensures that AI-generated results are not the sole determinant of academic integrity, safeguarding student rights and maintaining a fair and equitable assessment process (Balducci, 2024).

Additionally, many institutions are actively working to improve the accuracy of their AI-driven plagiarism detection tools. For instance, a university in the Western part of the USA, among others, has implemented adjustments to their AI systems to reduce the sensitivity to common academic phrases and expressions, thereby minimizing the likelihood of false positives. These proactive measures demonstrate a commitment to continuous improvement and a recognition that the effective use of AI requires ongoing refinement and adaptation.

Ultimately, these case studies offer valuable lessons for the future of AI in education. Firstly, they underscore the critical importance of human oversight. While AI can efficiently analyze vast amounts of data, human judgment is essential for interpreting the results, considering the context, and making informed decisions about academic integrity. Secondly, these cases highlight the need for ongoing research and development to improve the accuracy and fairness of AI-driven plagiarism detection tools. This includes refining AI algorithms, expanding training datasets, and continuously evaluating and updating these systems based on real-world experiences.

B. Technological Solutions

It can be noted that the continuous evolution of AI presents exciting opportunities for enhancing the accuracy and effectiveness of plagiarism detection tools. While current systems have made significant strides, the ongoing challenge of false positives necessitates a multifaceted approach to improvement.

Refining and improving AI algorithms are crucial steps towards minimizing false accusations. Current AI models are often trained on limited datasets, which can hinder their ability to recognize the nuances of different writing styles, dialects, and academic disciplines. Expanding the training datasets to include a more diverse range of academic texts, including those from various disciplines and cultural contexts, is crucial for enhancing the accuracy and fairness of these systems. Furthermore, incorporating advanced language models, such as transformer models, can significantly improve AI's ability to understand context, identify subtle nuances in language, and differentiate between legitimate academic writing and instances of plagiarism (Russell & Norvig, 2024; Bishop, 2024).

While refining AI algorithms is vital, the importance of human oversight cannot be overstated. AI systems, despite their sophistication, cannot fully replicate the nuanced understanding of human language and the broader context of academic discourse. Human reviewers, such as faculty members and academic integrity officers, play a crucial role in interpreting AI-generated results, considering the unique circumstances of each case, and making informed judgments about academic integrity. This combined approach ensures a more balanced and nuanced assessment process, mitigating the risk of false accusations and maintaining a fair and equitable environment for all students (Dehghantanha & Choo, 2024).

Also, embracing innovative technologies, such as transformer models and neural networks, offers promising avenues for enhancing the accuracy and reliability of plagiarism detection. Transformer models, renowned for their ability to understand and generate human-like text, have the potential to revolutionize how AI analyzes and interprets academic writing. Neural networks, with their capacity to learn complex patterns from massive datasets, can be trained to recognize subtle stylistic variations and identify instances of plagiarism that may elude traditional detection methods (Suresh Babu, 2024; Rabkin, 2024). By leveraging these cutting-edge technologies, we can significantly improve the accuracy and effectiveness of AI-driven plagiarism detection tools.

With all this, addressing the challenge of false positives in AI-driven plagiarism and AI-generated text detections require a multifaceted approach. Improving AI algorithms, incorporating human oversight, and embracing innovative technologies are crucial steps towards creating a more accurate, fair, and equitable system for maintaining academic integrity in the digital age.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2025 doi:10.20944/preprints202501.0060.v1

C. Ethical and Practical Points

Too much reliance on AI in academic settings necessitates a careful consideration of the ethical and practical implications of these technologies. While AI offers significant potential for enhancing academic integrity, it is crucial to address the potential for bias, ensure transparency in decision-making, and develop robust policies that safeguard student rights and promote fairness.

To set the record straight, one of the major concerns is the potential for bias within AI-driven plagiarism detection systems. If the training data used to develop these systems is biased, for example, by overrepresenting certain writing styles or underrepresenting diverse linguistic backgrounds, the resulting algorithms may inadvertently discriminate against certain groups of students. Research by Gandomi et al. (2024) and Alexander et al. (2023) has highlighted the potential for such biases to impact the accuracy and fairness of plagiarism detection outcomes, potentially leading to the disproportionate targeting of students from underrepresented groups. Addressing this requires careful consideration of the diversity and inclusivity of the training data used to develop these AI systems.

As can be reflected in the paper, transparency and accountability in AI decision-making are also crucial for maintaining trust and ensuring fairness. The "black box" nature of many AI algorithms can make it difficult to understand how these systems arrive at their conclusions, making it challenging for students and educators to understand the basis for a flagged submission. Without clear explanations for AI-generated results, it becomes difficult to identify and address potential biases or errors. As Leskovec et al. (2024) and Agrawal et al. (2024) have emphasized, transparency is crucial for building trust in AI systems and ensuring their ethical and responsible use in education. Institutions must prioritize the development of AI systems that provide clear and understandable explanations for their outputs, enabling educators and students to better understand the decision-making process.

Moreover, the dependence on AI-driven plagiarism detection systems raises concerns about the potential for inconsistencies and confusion. Different AI tools, with their own unique algorithms and training data, may produce varying results for the same piece of student work. This can lead to uncertainty and confusion for both students and educators, undermining the clarity and consistency of academic assessments.

Looking into these ethical and practical concerns requires a proactive approach to policy development. This includes establishing clear guidelines for the use of AI in plagiarism detection, emphasizing the importance of human oversight, and promoting the development of transparent and explainable AI systems.

Indeed, by combining the power of AI with the wisdom and expertise of human educators, and by prioritizing ethical considerations such as fairness, transparency, and inclusivity, we can create a more equitable and effective system for maintaining academic integrity in the digital age.

VI. Analysis

Having seen the complexities of AI-driven plagiarism detection and examined real-world case studies, some saturated key insights emerge. Firstly, it is clear that while AI offers powerful tools for identifying potential instances of academic misconduct, these systems are not without their flaws. False accusations, often stemming from the limitations of current AI algorithms, can have a significant impact on students, educators, and the overall integrity of the academic process.

A. Answers to the Statement of the Problem

1. On factors that contribute to the high rate of false positives in AI-driven plagiarism and AI-generated text checkers detection systems. False positives arise when AI detection tools misinterpret original student work as plagiarized. This can occur due to several factors, including limitations in the training data used to develop these algorithms. As highlighted by Smith and Doe (2024), insufficiently diverse training data can lead to AI systems that struggle to recognize the nuances of different writing styles and academic disciplines. Furthermore, AI

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2025 doi:10.20944/preprints202501.0060.v1

systems may be overly sensitive to common academic phrases or minor stylistic variations, leading to false positives. The case of the university in the Northeastern United States, where common academic phrases were mistakenly flagged, underscores the challenges of differentiating between legitimate academic writing and accidental similarities.

2. On the question about "false positives" and its impact on student confidence, motivation, and overall academic performance. When AI systems generate false accusations of plagiarism, they can significantly impact academic honesty and erode trust between students and educators. These false accusations can have a detrimental impact on students, damaging their reputations, lowering their grades, and potentially jeopardizing their future academic and professional opportunities. Moreover, these mishaps can create an atmosphere of distrust and anxiety among students, leading to a decline in student confidence and a negative impact on the overall learning environment. Students from diverse linguistic or cultural backgrounds may be particularly vulnerable to these false accusations, as their unique writing styles and language use may be misconstrued by AI systems that are primarily trained on mainstream academic writing.
3. On the issue about the innovative strategies that can be implemented to improve the accuracy and fairness of AI-driven plagiarism detection while safeguarding academic integrity. To mitigate the risk of false accusations, AI algorithms require significant refinement. This necessitates expanding and diversifying the training data used to develop these systems to include a wider range of writing styles, academic disciplines, and cultural contexts. Furthermore, enhancing the ability of AI systems to understand the nuances of human language, such as context, tone, and stylistic variations, is crucial. This can be achieved by incorporating advanced language models, such as transformer models, which excel at understanding and generating human-like text. Additionally, integrating human feedback into the AI decision-making process can significantly improve accuracy and fairness.

Over and above the technical limitations, the analysis highlights the crucial role of human oversight in the academic integrity process. While AI can efficiently analyze vast amounts of text and identify potential issues, it cannot fully replicate the nuanced understanding of human language, the broader context of academic discourse, and the unique circumstances of each individual student. As Dehghantanha & Choo (2024) emphasize, a human-centered approach is essential, where AI serves as a tool to assist, not replace, the critical judgment of educators.

The emergence of advanced technologies, such as transformer models and neural networks, offers exciting possibilities for enhancing the accuracy and reliability of plagiarism detection tools. These cutting-edge technologies, with their capacity to learn complex patterns and understand the nuances of human language, have the potential to revolutionize how AI analyzes and interprets academic writing, minimizing the risk of false accusations and enhancing the accuracy and fairness of the detection process (Suresh Babu, 2024; Rabkin, 2024).

With all this, the analysis underscores the need for a balanced and nuanced approach to AI-driven plagiarism detection. While these technologies offer valuable tools for maintaining academic integrity, it is crucial to acknowledge their limitations, address potential biases, and prioritize the importance of human judgment and ethical considerations. By fostering a collaborative approach that combines the power of AI with the wisdom and expertise of human educators, institutions and AI makers can create a more equitable, effective, and human-centered system for upholding academic integrity in the digital age. Hence, this paper intends to suggest several steps to improve the use of AI in plagiarism and AI-generated output of students.

Improvement Area	**Description**	**Benefits**
Understanding Context	Improves the ability to analyze the context around certain phrases or ideas,	Helps differentiate between content that overlaps naturally due to common knowledge

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2025 doi:10.20944/preprints202501.0060.v1

Improvement Area	Description	Benefits
	considering the topic, surrounding text, and its overall purpose.	and actual plagiarism, making it easier to spot unintentional similarities or false positives.
Focus on Meaning	Shifts focus to understanding the deeper meaning behind words and phrases, not just matching them on the surface.	Better identification of when ideas or expressions are naturally shared across texts and when they are truly copied.
Advanced Machine Learning	Uses more advanced methods like deep learning to detect complex patterns in writing and subtle differences between texts.	Increases accuracy by recognizing less obvious patterns and variations that simple algorithms might miss.
Accurate Citation Detection	Improves the ability to spot properly attributed sources, making it easier to distinguish between valid citations and uncredited copied content.	Ensures that quotes, references, and paraphrases are not flagged as plagiarism, respecting academic standards.
Compare Across Fields	Expands the ability to compare content across different types of work, beyond just one category like academic articles.	Reduces false alarms when similar phrases appear naturally in different subjects or disciplines, where overlap is common.
Spotting Paraphrasing	Enhances the ability to detect when content is paraphrased rather than directly copied, even if the wording is different but the ideas remain the same.	Better at catching disguised plagiarism, such as when content is altered slightly but still essentially identical.
Statistical Analysis of Similarity	Uses better statistical methods to assess how often certain phrases or structures appear naturally in language.	Reduces false positives by understanding when similar wording is likely due to common usage rather than intentional copying.
Training with Real-World Data	Trains detection tools with diverse, real-world content, both plagiarized and unintentionally similar, to improve detection accuracy in varied situations.	Helps the system learn to distinguish subtle differences, increasing its ability to make accurate decisions in complex cases.
Support for Multiple Languages	Improves the ability to handle multiple languages and consider cultural differences in language use.	Helps catch plagiarism in non-English texts and across diverse cultural contexts where similar expressions may appear.
Consider Legal and Ethical Standards	Integrates legal and ethical rules that clarify the line between acceptable reuse of content and actual plagiarism.	Ensures fair treatment of cases involving fair use or public domain material, reducing the risk of unjustly flagging innocent content.
Direct Quotes Exemptions	Recognizes and exempt direct quotes from plagiarism detection tools, as long as they are properly cited and attributed. It can use the quotation marks as its signal clue.	Prevents unnecessary flags on correctly cited and attributed quotes, ensuring that legitimate use of sources is not penalized. Most of the time, this is automatically flagged as plagiarized.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2025 doi:10.20944/preprints202501.0060.v1

Improvement Area	Description	Benefits
Localization	Adapts the system to recognize regional and cultural variations in language, so it can tell when similarities are common in specific contexts.	Prevents misidentification of content as plagiarized when phrases or expressions are typical in a particular culture or region.
Auto-Delete Files After the First Run	Implements a feature where files used in plagiarism and AI-generated text tests are automatically deleted after the first run to refresh its memory for subsequent tests.	Ensures privacy and security by clearing test files after each run, preventing unintentional reuse of data and keeping the system efficient for future tests.

Here, while AI plagiarism detection tools have proven valuable in identifying potential instances of plagiarism and AI-generated content, they should not be considered the final authority in assessing a student's work. These tools are most effective when used to highlight areas that may require further review, with the ultimate decision about the integrity of the writing resting with experienced educators who have the context, knowledge, and expertise to make informed judgments. Building on the findings of Gegg-Harrison and Quarterman (2024), this research article emphasizes the importance of continuously improving AI systems and incorporating human oversight to ensure a fair, transparent, and accurate academic environment. Research in the future could focus on advancing AI technology, refining data collection and analysis methods, and exploring the long-term impact of false accusations on student performance and well-being. By addressing these areas and focusing on the suggestions outlined above, AI developers and educational institutions can collaborate to create a more supportive and equitable educational system.

B. Thematic Analysis

Apart from answering the three (3) statements of the problem as stipulated in the introduction, this research article delves deeper into the complexities of AI-driven plagiarism and AI-generated text detection, uncovering several key themes (10 to be exact) that highlight the challenges and opportunities presented by these technologies.

1. Concerns about Ethics: The rise of AI-driven plagiarism detection tools raises significant ethical concerns. Privacy is paramount, as these systems often require access to sensitive student data, such as their writing. This raises concerns about data security and the potential for misuse, as highlighted by Casal and Kessler (2023). Moreover, the potential for algorithmic bias, where certain groups of students may be disproportionately affected by false accusations, poses a serious ethical dilemma. Educational theories like constructivism emphasize the importance of creating an ethical learning environment where students feel safe, respected, and free from undue surveillance.

2. Limitations on Technical Concerns: A major limitation of current AI plagiarism detection tools lies in their ability to accurately understand the nuances of human language. These systems often struggle to differentiate between legitimate academic writing, such as the use of common phrases and paraphrasing, and instances of actual plagiarism. This limitation, as evidenced by the study by Gegg-Harrison and Quarterman (2024), can lead to a high rate of false positives, particularly impacting students who are neurodivergent or whose first language is not English. This not only undermines student confidence but also raises questions about the reliability and accuracy of these technologies.

3. Psychological Impact on Students: The psychological impact of false accusations cannot be ignored. The stress and anxiety associated with being wrongfully accused of plagiarism can significantly impact student well-being. As highlighted by Casal and Kessler (2023) and Gegg-Harrison and Quarterman (2024), these experiences can lead to decreased motivation, increased

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2025 doi:10.20944/preprints202501.0060.v1

anxiety, and even mental health challenges. Cognitive Load Theory suggests that this stress can further impede learning by overwhelming students' cognitive resources.

4. Impact on Academics: False accusations can have severe academic consequences, including lowered grades, damaged academic records, and even disciplinary actions. These repercussions can significantly impact a student's academic trajectory and future opportunities. Moreover, these experiences can erode trust between students and educators, undermining the foundation of a healthy and supportive learning environment.

5. Fairness and Bias Issues: A significant concern is the potential for bias within AI plagiarism detection systems. If the training data used to develop these algorithms reflects existing societal biases, the resulting systems may inadvertently discriminate against certain groups of students, such as those from diverse linguistic or cultural backgrounds. As Hesse and Helm (2024) point out, these biases can lead to unfair and inaccurate assessments, undermining the principles of equity and inclusion that are central to effective education.

6. Concerns about Privacy: The collection and analysis of student writing data raise significant privacy concerns. As Mishra (2024) highlights, the constant surveillance implied by these systems can create a sense of unease and distrust among students, hindering their ability to learn freely and openly. Educational theories like constructivism emphasize the importance of creating a safe and supportive learning environment where students feel respected and their privacy is protected.

7. Reliability and Accuracy: The accuracy and reliability of current AI plagiarism detection tools remain a significant challenge. As Moravvej et al. (2023) discuss, these systems are prone to errors, leading to a high rate of false positives. These inaccuracies not only undermine the credibility of the assessment process but also create unnecessary stress and anxiety for students.

8. Diverse Learner’s Support: AI-driven plagiarism detection tools must be designed to support the diverse needs of all students. However, existing systems may inadvertently disadvantage neurodivergent students or those whose writing styles deviate from the norm. As Hesse and Helm (2024) point out, it is crucial to develop AI systems that are inclusive and cater to the unique learning needs of all students.

9. Neurodivergent Student’s Impact: Neurodivergent students, with their unique cognitive styles and approaches to learning, may be particularly vulnerable to the limitations of current AI plagiarism detection tools. Their unique writing styles may be misinterpreted by AI algorithms, leading to a higher rate of false accusations. This can exacerbate existing challenges and create additional stress for these students.

10. Educator’s Role: Educators play a crucial role in navigating the complexities of AI-driven plagiarism detection. As highlighted by Harvard Business Publishing Education (2023), teachers must possess the knowledge and expertise to critically evaluate AI-generated results, understand the limitations of these systems, and provide appropriate support to their students. By fostering open communication, building trust, and emphasizing the importance of critical thinking and original thought, educators can help students navigate the challenges of the digital age while maintaining academic integrity.

Overall, while AI plagiarism checkers offer valuable tools for maintaining academic honesty, they also present significant challenges. This analysis highlights the crucial need for a balanced and nuanced approach that combines the power of AI with the wisdom and expertise of human educators. By addressing the ethical, technical, and pedagogical concerns raised in this research, we can create a more equitable, effective, and human-centered system for upholding academic integrity in the digital age.

C. Limitation Acknowledgement

Overall, while the paper provides valuable insights, the output per se, is subject to several limitations. The case studies may not fully represent the diversity of experiences, and the selection of real-world examples could be influenced by inherent biases. Furthermore, the institutional responses analyzed may reflect the unique policies and practices of specific organizations. The lessons learned and proposed technological solutions are context-dependent and may not be universally applicable.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2025 doi:10.20944/preprints202501.0060.v1

The research methodology, including the thematic analysis, may involve a degree of subjectivity, and the proposed technological solutions are constrained by current technological capabilities. Acknowledging these limitations is crucial for interpreting the findings within the specific context of this research.

VII. Future Directions

AI text generators and plagiarism detectors have ushered in a new era of both opportunity and challenge for education. While these tools can enhance learning and creativity, they also pose a significant threat to academic integrity. To navigate this evolving landscape, a nuanced approach is needed, one that leverages the power of AI while safeguarding the principles of originality and ethical scholarship. To achieve this, effective safety nets can be created by implementing the following actions:

1). Retooling AI Programs: Imagine AI that can truly understand the human element of writing. Instead of just flagging potential plagiarism, these AI programs would be like insightful writing partners, recognizing the unique voice and thought process behind each student's work. They would go beyond simple word matching, delving into the nuances of sentence structure, word choice, and the underlying logic of arguments. It's like having a highly intelligent tutor who can not only identify potential issues but also provide valuable feedback on writing style and critical thinking skills, helping students grow as writers and thinkers.

2). Policies for the Institution: The rise of AI tools is changing the rules of the game for academic integrity. It's time for institutions to adapt and create a clear, ethical framework for how AI can be used in education. These policies shouldn't just dictate which tools are allowed, but also explore the deeper ethical considerations. Students should have a voice in shaping these policies, ensuring their concerns and perspectives are heard. Imagine a system where human judgment remains central, with teachers carefully reviewing AI-flagged submissions and considering the unique circumstances of each student. If a student feels unfairly accused, they should have clear avenues for appeal, ensuring fairness and protecting their academic standing. This approach fosters a culture of trust and open communication between students and faculty, where AI serves as a tool to support learning, not to create suspicion and fear.

3). Strategic Plans: The future of AI in education requires a commitment to equity and inclusivity. We need AI systems that can understand and appreciate the diverse tapestry of human language and thought. Imagine AI that can accurately assess the originality of work submitted in a variety of languages and dialects, ensuring fairness for all students, regardless of their background. This requires ongoing development and refinement, keeping pace with the ever-evolving landscape of AI and the changing ways students learn and express themselves. Transparency is key. Students and educators should have a clear understanding of how these AI tools work and how they make decisions. This open dialogue builds trust and ensures that AI serves as a valuable tool for enhancing academic integrity, rather than becoming a threat to the human-centered values of education.

Based on the findings of the study, it can be deduced that by embracing these future ideas, AI technology can continue to play a key role in fighting plagiarism. However, as Ibrahim (2023) notes, AI must be seen as a tool that helps educators, not as the final decision-maker of academic honesty. Ensuring that AI remains a helpful resource in academic settings, one that works alongside human judgment, is essential for maintaining fairness and trust in the educational process. With thoughtful development and clear policies, the future of plagiarism and AI-generated text detection tools holds significant potential for improving academic honesty while respecting students' rights.

VIII. Conclusions and Recommendations

In this research article, the Pandora’s tech-box is unmasked. Educational institutions must keep pace with evolving technology by effectively filtering automatically generated text and implementing plagiarism detection systems. While these advancements offer benefits in academia, they also introduce challenges such as the occurrence of “false positives”. The study addresses these issues and proposes strategies to enhance the precision and equity of plagiarism checks, safeguarding students' rights and upholding the credibility of academic evaluations. It underscores the importance of not solely relying on plagiarism tools as definitive proof of originality but rather as aids for educators to exercise their judgment in evaluating academic integrity. Research reveals that issues like limited diverse data, overly sensitive algorithms, rigid program frameworks, and the complexity of distinguishing genuine plagiarism from common similarities can compromise academic honesty and unfairly tarnish students' reputations. Through reflective case studies, practical examples, institutional responses, key takeaways, technological interventions, ethical considerations, and thorough thematic analyses, the necessity to address these limitations becomes apparent. It is critical to acknowledge the fallibility of detection tools and their potential for generating false identifications. Their usage should be tempered with human oversight to ensure impartiality and correctness by thoroughly examining any discrepancies. Furthermore, since common writing tools like tablets, computers, cellphones, and laptops already incorporate AI-generated features like "autocorrect", branding students' work as AI-generated based on these tools alone is unreasonable. Furthermore, the notion of "humanizing" flagged content presents challenges as it may still be identified as machine-generated by another AI tools. Academic exercises should not just boil down to AI vs. AI because of plagiarism and AI-generated text issue. To enhance accuracy and equity, however, advancements in technology should involve leveraging more diverse datasets, enhancing contextual understanding, integrating human judgment, and incorporating the recommendations from the current study. Educators should familiarize themselves with these tools, and students should receive clear explanations on their functionality. Policymakers and AI developers must update regulations to refine these tools, diminish instances of false positives, and promote equity and responsibility in academic evaluations. Subsequent research endeavors should concentrate on refining sophisticated systems, advancing data methodologies, and examining the enduring impacts of wrongful accusations on students' well-being and academic progress, ultimately fostering an environment of precision, fairness, and transparent learning environment.

Author Contribution: The author is solely responsible for conceiving and designing this article, gathering the data, and writing the entire manuscript.

Funding Statement: This research was conducted without any specific financial support from public, commercial, or non-profit funding agencies.

Declaration of Interest: The author declares no conflict of interest.

Acknowledgments: The author would like to express gratitude to the SMCII library, Google Scholar, Mendeley, ResearchGate, Academia, and Microsoft Copilot for providing illustrations, additional information, and other scholarly online resources that were essential to this research.

References

1. Abdullayeva, M., & Muzaffarovna, M. Z. (2023). The impact of ChatGPT on student's writing skills: An exploration of AI-assisted writing tools. *Zenodo*. https://doi.org/10.5281/ZENODO.7876800
2. Agrawal, A., Gans, J., & Goldfarb, A. (2024). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Review Press.
3. Albadarin, Y., Saqr, M., Pope, N., & Tukiainen, M. (2023). A systematic literature review of empirical research on ChatGPT in education. *ResearchGate*. https://doi.org/10.13140/RG.2.2.21598.82245
4. Alexander, K., Savvidou, C., & Alexander, C. (2023). Who wrote this essay? Detecting AI-generated writing in second language education in higher education. *Teaching English with Technology*, *23*(2), 25-43.
5. Alier, M., García-Peñalvo, F.-J., & Camba, J. D. (2024). Generative artificial intelligence in education: From deceptive to disruptive. *International Journal of Interactive Multimedia and Artificial Intelligence*, *8*(5), 5. https://doi.org/10.9781/ijimai.2024.02.011

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2025 doi:10.20944/preprints202501.0060.v1

6. Balducci, B. (2024). AI and student assessment in human-centered education. Frontiers in Education, 9, 1383148. https://doi.org/10.3389/feduc.2024.1383148
7. Barrett, A., & Pack, A. (2023). Not quite eye to A.I.: Student and teacher perspectives on the use of generative artificial intelligence in the writing process. *International Journal of Educational Technology in Higher Education, 20*(1), 1. https://doi.org/10.1186/s41239-023-00427-0
8. Bishop, C. M. (2024). *Pattern recognition and machine learning*. Cambridge University Press.
9. Bowen, J. A., & Watson, C. E. (2022). *Teaching with AI: Exploring the possibilities of artificial intelligence in the classroom*. EduTechnica Publishing.
10. Braun, V., & Clarke, V. (2021). *Thematic analysis: A practical guide* (3rd ed.). SAGE Publications.
11. Brynjolfsson, E., & McAfee, A. (2024). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. Harvard Business Review Press.
12. Casal, J. E., & Kessler, M. (2023). Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics, 2*(3), 100068.
13. Clark, H. (2020). *The AI-infused classroom: 21st-century technology for 21st-century learners*. EdTech Team Press.
14. Cook, J. (2024). Using GenAI to reduce plagiarism. *Harvard University*. Retrieved from https://www.harvard.edu/ai/2024/02/13/jacob-cook-using-genai-to-reduce-plagiarism/
15. Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.). SAGE Publications.
16. Dehghantanha, A., & Choo, K.-K. R. (Eds.). (2024). *Handbook of big data and IoT security*. Oxford University Press.
17. Dixit, A., Unnathi, P. N., Guttedar, R. J., & More, S. (2024). Advancements in plagiarism detection: A comprehensive review and proposal. *Journal of Emerging Technologies and Innovative Research (JETIR).* Retrieved from https://www.jetir.org/papers/JETIR2404050.pdf
18. El Mostafa, H., & Benabbou, F. (2020). A deep learning based technique for plagiarism detection: A comparative study. *IAES International Journal of Artificial Intelligence, 9*(1), 81.
19. Eslit, E. R. (2023). Thriving beyond the crisis: Teachers' reflections on literature and language education in the era of artificial intelligence (AI) and globalization. *International Journal of Education and Teaching, 3*(1), 46-57. https://doi.org/10.51483/IJEDT.3.1.2023.46-57
20. Flanagin, A., Bibbins-Domingo, K., Berkwits, M., & Christiansen, S. L. (2023). Nonhuman "authors" and implications for the integrity of scientific publication and medical knowledge. *JAMA, 329*(8), 637-639.
21. Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: A systematic literature review. *ACM Computing Surveys, 52*(6), 112. https://dl.acm.org/doi/fullHtml/10.1145/3345317
22. Gandomi, A., Haider, M., & Hosseini, A. (2024). *Artificial intelligence for business: A roadmap for business professionals*. MIT Press.
23. Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2022). Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *BioRxiv*. https://doi.org/10.1101/2022.12.23.521612
24. Gegg-Harrison, & Quarterman (2024). AI detection's high false positive rates and the psychological and material impacts on students. In *Academic integrity in the age of artificial intelligence*. DOI: 10.4018/979-8-3693-0240-8.ch011
25. Harvard Business Publishing Education. (2023). Stop focusing on plagiarism, even though ChatGPT is here. *Harvard Business Publishing Education*. Retrieved from https://hbsp.harvard.edu/inspiring-minds/stop-focusing-on-plagiarism-even-though-chatgpt-is-here
26. Hesse, F., & Helm, G. (2024). Writing with AI in and beyond teacher education: Exploring subjective training needs of student teachers across five subjects. *Journal of Digital Learning in Teacher Education*, 1–14. https://doi.org/10.1080/21532974.2024.2431747

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2025 doi:10.20944/preprints202501.0060.v1

27. Ibrahim, K. (2023). Using AI-based detectors to control AI-assisted plagiarism in ESL writing: “The terminator versus the machines”. *Language Testing in Asia, 13*(1), 46. https://languagetestingasia.springeropen.com/articles/10.1186/s40468-023-00260-2
28. Johnson, L., & Nguyen, T. (2019). *Academic plagiarism detection: A systematic literature review.* MIT.
29. Khalil, M., & Er, E. (2023, June). Will ChatGPT get you caught? Rethinking of plagiarism detection. In *International conference on human-computer interaction* (pp. 475-487). Springer Nature Switzerland.
30. Lannoy, V. (2023). AI-based plagiarism detectors: Plagiarism fighters or makers? *EMWA Journal*.
31. Leskovec, J., Rajaraman, A., & Ullman, J. D. (2024). *Mining of massive datasets* (3rd ed.). MIT Press.
32. Merod, A. (2023). Turnitin admits there are some cases of higher false positives in AI writing detection tool. https://www.k12dive.com/news/turnitin-false-positives-AI-detector/652221/
33. Mishra, S. (2024). Enhancing plagiarism detection: The role of artificial intelligence in upholding academic integrity. *Library Philosophy and Practice*. Retrieved from https://digitalcommons.unl.edu/libphilprac/9415
34. Moravvej, M., et al. (2023). AI in plagiarism detection: Accuracy and academic integrity. *RikiGPT*.
35. Patel, R. (2024). *The role of AI in plagiarism detection: A comprehensive guide*. TextA.AI Press.
36. Quidwai, A., Li, C., & Dube, P. (2023). Beyond black box AI generated plagiarism detection: From sentence to document level. *Association for Computational Linguistics*.
37. Rabkin, Y. (2024). *Intellectual property and artificial intelligence*. Oxford University Press.
38. Russell, S., & Norvig, P. (2024). *Artificial intelligence: A modern approach* (4th ed.). Cambridge University Press.
39. Sadhin, I. H., Hassan, T., & Nayim, M. A. (2024). Plagiarism detection using artificial intelligence. *International Journal of Computer and Information System*.
40. Santra, P. P., & Majhi, D. (2023). Scholarly communication and machine-generated text: Is it finally AI vs AI in plagiarism detection? *Journal of Information and Knowledge, 60*(3), 175-183. https://doi.org/10.17821/srels/2023/v60i3/171028
41. Shabanov, I. (2024). *Understanding AI plagiarism detection: Reliability and ethical use in academia*. Effortless Academic Press.
42. Smith, J., & Doe, A. (2024). Exploring attentive Siamese LSTM for low-resource text plagiarism detection. MIT Press.
43. Suresh Babu, D. (2024). AI and plagiarism: Challenges, detection, and prevention. *International Journal of Engineering Research and Development*.
44. Zhang, P., & Tur, G. (2024). A systematic review of ChatGPT use in K-12 education. *European Journal of Education, 59*(2), 12599. https://doi.org/10.1111/ejed.12599

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.