Examining the Accuracy of AI Detection Software Tools in Education

Mohanad Halaweh*, Ghaleb El Refae†
*College of Business, Al Ain University, Al Ain, UAE
*Email: mohanad.halaweh@aau.ac.ae
†Email: ghalebelrefae@aau.ac.ae

***Abstract*—Educators have raised concerns about the utilization of ChatGPT in generating unoriginal text and the possibility of plagiarism. To address these concerns, various AI text detection software tools have been developed to evaluate whether a text is AI generated or human generated. The aim of this research is to empirically examine the accuracy of AI detection tools in identifying AI-generated texts. An experiment was conducted using textual data generated by ChatGPT, which was assessed using Turnitin and four other AI detection tools. Through multiple iterations and interventions, the text was paraphrased by ChatGPT until it appeared original and could not be detected as AI-generated by Turnitin's AI detection tool. The findings revealed that all the AI detection software tools that were examined failed to detect the AI-generated text by ChatGPT in the final iteration. The findings provide valuable insights that have implications for various stakeholders, including educators, researchers, and AI text detection software developers. Based on the tools examined, educators and researchers do not need to set a specific threshold or percentage, including 0%, to determine what qualifies as acceptable AI-generated text. This is because establishing such a threshold can be misleading, considering the current limitations in the algorithms of these tools. Furthermore, the data generated in this paper can provide a solid basis for replicated research or software testing and assessment. It can be utilized to evaluate the accuracy of alternative AI detection tools and any future advancements in the tools mentioned in this investigation.**

***Keywords*—Artificial Intelligence, Education, Generative AI, AI Detection Tools, Turnitin, ChatGPT**

I. Introduction

Artificial intelligence (AI) has experienced rapid development in recent years and has brought about transformative changes in various fields, including education. One of the most recent disruptive AI-based tools is ChatGPT (Generative Pre-Trained Transformer), which was released by OpenAI in November 2022. Utilizing natural language processing and deep neural network models [1], ChatGPT generates human-like responses to user input. It stands out due to its exceptional capacity to generate and deliver entirely novel content during live interactions with users. It has the ability to consistently maintain a coherent dialog style, effectively engage users, and furnish authentic and pertinent responses, thereby avoiding irrelevant answers to individual queries [2]. ChatGPT has emerged as the most advanced chatbot worldwide [3]. Unlike traditional chatbots, ChatGPT is built upon GPT-3, the third generation of the GPT series developed by OpenAI. The GPT-3 model is significantly more advanced and trained on an extensive dataset with approximately 175 billion parameters, compared to GPT-2's 1.5 billion parameters. This enables it to generate human-like text with high accuracy [4]. In March 2023, OpenAI announced the release of ChatGPT-4, its latest and most advanced version. OpenAI proclaimed it to be more accurate and innovative than all previous versions. GPT-4 is expected to have around 100 trillion parameters, approximately 500 times more than GPT-3, thus approaching the number of neural connections in the human brain [3].

Despite the advantages of ChatGPT in understanding and generating meaningful responses, as demonstrated by Korinek across various use cases, such as ideation, feedback, writing, summarizing text, data analysis, coding, and solving mathematical problems [5], researchers and educators have raised concerns about its potential use for generating unoriginal texts in research papers and assignment reports as well as the risk of plagiarism [6, 7, 8]. In the context of education, concerns have been raised regarding the potential misuse of ChatGPT, which enables students to generate human-like responses that evade detection by plagiarism detection software. A survey conducted in January 2023 that included over 1000 university students revealed that more than one-third of them employed ChatGPT for their assessment writing. Alarmingly, 75% of these students acknowledged that using ChatGPT amounted to cheating, yet they persisted in doing so [9]. The utilization of ChatGPT raises apprehensions about students merely copying and pasting texts without engaging in critical analysis of the selected content from sources. This practice often occurs without proper citation of the original sources, resulting in a failure to recognize the potential for plagiarism and, consequently, compromising academic integrity [6], [10]–[13]. Such concerns highlight the importance of addressing the issue of academic integrity and promoting the responsible use of AI tools, such as ChatGPT, within educational settings.

To address these concerns, several AI text-detection tools have been developed to determine whether a text is AI generated or human generated. Previous research attempts relied on online tools, such as GPT-2 Output Detector and GPTZero, to identify AI-generated text or utilized Turnitin to assess the similarity index (matching text) of ChatGPT text without specifically targeting the detection of AI-generated text [11], [14]–[17]. In contrast, the current study employs Turnitin's AI detection tool, which became available in April 2023 [18], along with other tools. This paper aims to evaluate the accuracy of AI detection tools in identifying plagiarism in both text similarity and AI-generated text, an area that has not previously been empirically explored. Importantly, the most popular academic tool, Turnitin, did not possess the capability to detect AI-generated texts until April 2023, which highlights

979-8-3503-5475-1/24/$31.00 ©2024 IEEE

the novelty and significance of this research and its findings, at least at the time of writing.

Hence, it is imperative to assess the effectiveness of AI detection tools, such as Turnitin, in identifying texts generated by ChatGPT. In light of this, this study aims to address the following research questions:

1. Can AI-generated text by ChatGPT bypass AI text detection software tools?
2. What are the levels of accuracy exhibited by AI detection software tools (e.g., Turnitin) in identifying AI-generated text generated specifically by ChatGPT?

The aim of this study is to assess the accuracy of AI detection tools in identifying texts created by AI. To achieve this, an experiment was conducted using textual data generated by ChatGPT in April 2023, which was then evaluated using Turnitin and four other AI detection tools (GPT-2 Output Detector, AI Text Classifier, ZeroGPT, and GPTZero).

Turnitin was selected in this experiment as the primary tool for AI-generated text/plagiarism detection due to its widespread recognition as one of the leading and most widely utilized tools in academia employed by educators and researchers [19], [20]. With the largest market share, Turnitin claims to be trusted by more than 15 000 higher education institutions in over 140 countries [21]. Notably, in April 2023, Turnitin incorporated a new feature specifically designed for detecting AI-generated text, which further enhanced its capabilities. Apart from Turnitin, the other tools were included because they were among the first to be released and made freely accessible. These tools also assert their ability to identify AI-generated text effectively.

II. Related Work

Several recent studies have been conducted to evaluate AI detection tools. Khalil and Er assessed the originality of 50 essays generated by ChatGPT using the plagiarism detection tools iThenticate and Turnitin [11]. They found that ChatGPT had the potential to generate sophisticated text outputs with high originality, making it challenging for plagiarism-checking software to detect. However, it is important to mention that their study focused only on text similarity matching rather than AI-generated text. Referring to several studies, Pegoraro et al. assessed the effectiveness of multiple tools and various approaches in detecting ChatGPT-generated content [17]. They found that the most successful tool achieved a success rate of less than 50% in detecting content. However, the methods, the analytical details, how the text was generated, and how it deceived the tools were not clearly described, and it was presented as a black box without specific details. Furthermore, the study did not specifically evaluate Turnitin's AI capabilities.

Elkhatat et al. [16] investigated the capabilities of various AI content detection tools in discriminating between human and AI-authored content. For evaluation, 15 paragraphs each from ChatGPT Models 3.5 and 4 and five human-written responses were generated. AI content detection tools from OpenAI, as well as Writer, Copyleaks, GPTZero, and CrossPlag, were used for evaluation, and it was shown that the tools were more accurate in identifying content generated by GPT 3.5 than by GPT 4. However, it should be noted that the Turnitin AI detector was not used in their study because it had not been widely adopted or activated across educational institutions. Anderson et al. [14] used GPT-2 Output Detector to detect AI-generated texts and found that with additional paraphrasing, the tool showed a shift in the "real" text percentage. For instance, when using GPT-2 Output Detector for one essay, the percentage went from 0.02% to 99.52%, which indicated that the text was human generated. The study provided a sample of text as supplementary data showing the generated and paraphrased texts, but it did so without detailing how the paraphrasing intervention was conducted. In addition, Turnitin was not used in this study.

Although there have been some attempts to evaluate AI detection tools, these attempts had certain limitations, as indicated previously. The current research evaluates the accuracy of Turnitin as a key trustworthy tool, not only for similarity checks (text matching) but also for AI detection. This particular feature has not been clearly assessed in previous studies, as it is a relatively new addition to Turnitin. Moreover, this research contributes a dataset and describes methodological steps for generating AI text, showcasing interventions for paraphrasing supported by evidence. This aspect was frequently absent in most previous studies, affecting the reliability and creditability of the results. In addition, the implications of the current research outputs for various stakeholders are discussed.

III. Research Method/Experiment

This experimental investigation was carried out utilizing two tools: ChatGPT, as an AI text generator, and Turnitin, as an AI detection software tool. The investigation began by providing a prompt to ChatGPT requesting an argument highlighting the unreliability of AI detection software tools and the importance of educators not solely relying on them. All prompts and texts generated by ChatGPT are consolidated in File 1 (attached in the supplementary material at https://data.mendeley.com/datasets/8d8npp4f94/1), which is an HTML source file exported from OpenAI's ChatGPT. It is important to note that each iteration of the text (response), including any typos, was preserved without any editing, and the prompts and responses from the conversation with ChatGPT were retained. The generated texts were then submitted to Turnitin to assess the percentages of similarity and AI detection. The experiment involved manually recording the percentages of AI-generated text and similarities in each iteration, which are shown in Table 1.

TABLE I. Turnitin Data (Percentages of Similarity and AI Detection) with ChatGPT Interventions

	Turnitin		ChatGPT
Iteration	**Similarity %**	**AI detection %**	**Intervention**
Iteration 0	NA	NA	Prompt ChatGPT to generate an argument discussing the unreliability of AI detection software and the importance of educators not solely depending on AI detection software tools.
Iteration 1	3%	100%	Rewrite the following (response from iteration 0) so that AI detection

			software cannot detect it.
Iteration 2	0%	48%	Paraphrase the following so that it looks like text generated by a human not an AI.
Iteration 3	3%	38%	Paraphrase the following text in academic way (only two paragraphs from six).
Iteration 4	0%	29%	Paraphrase the following text in academic way (only two paragraphs from six).
Iteration 5	0%	5%	Paraphrase the following text (only two sentences).
Iteration 6	0%	0%	No further action. The text (final output) is attached in the supplementary material.

IV. RESEARCH RESULTS

As anticipated, in the first iteration (the initial responses generated by ChatGPT), the Turnitin AI detection tool detected the texts as 100% AI generated, while the similarity percentage was 3% (see File 2 in the supplementary material at https://data.mendeley.com/datasets/8d8npp4f94/1). To reduce the AI percentage, multiple interventions were implemented in subsequent iterations. In the second iteration, an intervention was introduced in which ChatGPT was prompted to rewrite the texts generated in the initial iteration in a way that would avoid detection using AI detection software. Table I shows that the AI percentage dropped to 48% in the second iteration with this intervention.

This process of intervention and iteration continued with various interventions applied in each iteration until the AI percentage reached 0% and the similarity reached 0% (see File 3 in the supplementary material). Notably, in the later iterations, the changes were focused on specific paragraphs or sentences flagged by Turnitin as AI generated rather than entire generated texts being rewritten, as in the first and second iterations. The results showed that the Turnitin AI detection tool was unable to identify the AI-generated text. Furthermore, the generated textual data on the investigated subject (for the final output, see File 3 in the supplementary materials) not only completely evaded plagiarism detection (with 0% similarity and 0% AI detection) but also demonstrated meaningfulness and relevance.

It is important to highlight that the final output (File 3 in the supplementary material) did not contain any words contributed by the author; it was entirely generated by AI (i.e., ChatGPT) through multiple interventions and iterations.

To ensure a more comprehensive and robust conclusion, additional AI detection tools were utilized to evaluate accuracy compared to Turnitin. Specifically, the first and final outputs generated by ChatGPT were tested (File 4 in the supplementary materials) using other tools developed by OpenAI (GPT-2 Output Detector and AI Text Classifier) as well as ZeorGPT and GPTZero. Interestingly, all of these tools failed to identify the outputs as AI generated, particularly the final output, as depicted in Table 2. The data revealed that Turnitin exhibited greater accuracy compared to the other tools. However, Turnitin also failed to detect AI-generated texts in the final iteration (the sixth iteration), particularly after the text had undergone multiple rounds of paraphrasing using ChatGPT. Table II provides a comparison of the results from the various AI detection tools used to check the outputs of ChatGPT (initial and final iterations), offering insights into the performance of the AI detection tools. As the results show, the final output (the sixth iteration), which was entirely generated by ChatGPT, was evaluated as "0% AI-generated text" by Turnitin, "99.97% real" text by GPT-2 Output Detector, "Unlikely to be AI generated" by AI Text Classifier, "Human Written" by ZeroGPT, and "Likely to be written entirely by a human" by GPTZero. The tools' results were incorrect and misleading (as they were AI-generated text), and if instructors rely solely on them, this could lead to unfair judgment of students' work.

TABLE II. COMPARISON OF THE RESULTS OF AI DETECTION TOOLS IN DETECTING CHATGPT OUTPUTS (FIRST AND FINAL ITERATIONS)

	AI Detection %				
AI detection tools/ ChatGPT output	**Turnitin**	**GPT-2 Output Detector**	**AI Text Classifier**	**ZeroGPT**	**GPTZero**
ChatGPT first output – first iteration	100% AI generated	0.02% fake, 99.98% real	Unclear if it is AI generated	59.5% AI generated	Likely to be written entirely by AI
ChatGPT final output – sixth iteration	0% AI generated	0.03% fake, 99.97% real	Unlikely to be AI generated	Human written	Likely to be written entirely by a human

V. DISCUSSION

The findings answer the two main questions addressed in the Introduction: 1) they provide evidence that paraphrased texts generated by ChatGPT can bypass AI text-detection tools and 2) they demonstrate different levels of accuracy in AI-generated text detection, depending on the number of interventions/iterations made in paraphrasing by ChatGPT. For example, using Turnitin, one of the most popular and trusted tools among academic institutions, the accuracy was high (100% correct) in detecting texts initially generated by ChatGPT (first iteration). However, after paraphrasing by ChatGPT in the sixth iteration, Turnitin's accuracy dropped to zero, and it incorrectly detected the text as 0% AI generated, thereby indicating that the text was real and created by a human.

The experimental findings provide valuable insights to educators regarding the accuracy of AI detection tools in identifying AI-generated text. They highlight the need for educators to thoroughly evaluate students' work using various assessment methods and to exercise critical judgment. This is because researchers and students can generate fake theses, research papers, or assignment reports with entire texts generated by ChatGPT that can deceive AI detection tools, such as Turnitin. This poses a profound dilemma for educators, as they face the critical and challenging task of assessing such works, especially when students possess the

ability to convincingly defend their AI-generated submissions by effectively answering questions and presenting coherent arguments. In these circumstances, educators may find themselves lacking the necessary evidence to fail a student who has successfully defended their work, despite having suspicions that the work is AI generated. On the other hand, it would be unfair to penalize a student after detecting their work as AI generated, especially when the detection tool is inaccurate. This scenario raises critical ethical issues and challenges educators to adapt their assessment methods to address emerging complexities in the era of AI-generated content in the context of education.

The data obtained from this experiment provide insights for education policymakers, enabling them to develop policy regulations that promote students' awareness of the practice of generating AI-generated fake text and its implications for academic integrity. Information on awareness about this issue could be shared with students, informing them that instructors are aware of the potential for deception of AI detection tools. However, instructors should not rely solely on written text to assess students' work. Instead, they should employ a variety of assessment methods, including presentations, viva (oral examinations), and live critical thinking exercises, encouraging students to report unique ideas that may not have been provided by ChatGPT. By incorporating different assessment approaches, instructors can better measure students' understanding, creativity, and ability to apply knowledge, which goes beyond what AI-generated text can offer. Providing evidence (e.g., outputs generated by ChatGPT) to support their ideas would further demonstrate students' originality and thoughtful engagement in their academic work. This approach aims to foster a culture of academic integrity and promote authentic learning. It is worth mentioning that ChatGPT could be used for English editing and paraphrasing for students' own written ideas and texts, but not for those of others or for AI-generated texts.

The findings also provide valuable insights for developers of AI detection tools, highlighting the importance of positioning these tools primarily for initial screening purposes to identify potentially unoriginal text. It is important to avoid promoting these tools as capable of detecting all AI-generated text, as evidenced by the data from Turnitin and other tools. In this experiment, the Turnitin AI tool detected 0% of an entirely AI-generated text from the final output of ChatGPT. This suggests that relying solely on such tools for evaluating and determining the originality of content can be misleading. The experimental data serve as valuable feedback for the company behind Turnitin, enabling it to improve the capabilities of the tool for detecting AI-generated text in future versions, following previously reported assessments [18]. Furthermore, the results of the experiment—specifically, the ChatGPT final output—can be utilized by developers of other AI detection software tools to assess the accuracy of future improved versions of their software in detecting AI-generated text.

VI. CONCLUSION

This study aimed to examine the accuracy of AI detection software tools in identifying AI-generated texts. The results clearly demonstrate that the tested tools consistently failed to detect AI-generated texts produced by ChatGPT after multiple rounds of paraphrasing. This highlights the remarkable capability of the ChatGPT tool as a language model in creating human-like responses. Simultaneously, it underscores the limited effectiveness of tools such as Turnitin in identifying such content. These findings offer valuable insights with implications for various stakeholders, including educators, researchers, and developers of AI text-detection software.

Despite the valuable insights and contributions made in this study, it is important to acknowledge certain limitations that may have influenced the findings. This experiment examined only four tools to evaluate their effectiveness in identifying AI-generated texts. However, it should be noted that there may be other tools with varying levels of accuracy, so it cannot be conclusively claimed that all other tools would have failed in the context of this research. However, it is possible to use the final output of this experiment to further assess the accuracy of detecting AI-generated text using different tools. It is worth mentioning that different AI detection tools may utilize different algorithms, resulting in potential variations in their outcomes. Additionally, the text generated by ChatGPT for the experiment was limited to around 500 words. When evaluating longer texts, the accuracy of AI detection results may vary.

REFERENCES

[1] A. B. Mbakwe, I. Lourentzou, L. A. Celi, O. J. Mechanic, and A. Dagan, "ChatGPT passing USMLE shines a spotlight on the flaws of medical education," PLOS Digit. Health, vol. 2, no. 2, e0000205, 2023.

[2] R. M. Mostafizer and Y. Watanobe, "ChatGPT for education and research: Opportunities, threats, and strategies," Appl. Sci., vol. 13, no. 9, p. 5783, 2023, https://doi.org/10.3390/app13095783.

[3] J. Rudolph, S. Tan, and S. Tan, "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?" J. Appl. Learn. Teach., vol. 6, no. 1, 2023.

[4] T. B. Brown et al., "Language models are few-shot learners," arXiv:2005.14165, 2022.

[5] A. Korinek, "Generative AI for economic research: Use cases and implications for economists," J. Econ. Lit., vol. 61, no. 4, pp. 1281–1317, 2023.

[6] D. R. Cotton, P. A. Cotton, and J. R. Shipway, "Chatting and cheating: Ensuring academic integrity in the era of ChatGPT," Innov. Educ. Teach. Int., vol. 19, no. 1–2, 2023, https://doi.org/10.1080/14703297.2023.2190148.

[7] M. Perkins, "Academic Integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond," J. Univ. Teach. Learn. Pract., vol. 20, no. 2, 2023.

[8] B. A. Anders, "Is using ChatGPT cheating, plagiarism, both, neither, or forward thinking?" Patterns, vol. 4, no. 3, 2023.

[9] Intelligent, "Nearly 1/3 college students have used ChatGPT on written assessments." Intelligent.com. https://www.intelligent.com/nearly-1-in-3-collegestudents-have-used-chatgpt-on-written-assignments/ (accessed Jul. 25, 2023).

[10] M. Sullivan, A. Kelly, and P. McLaughlan, "ChatGPT in higher education: Considerations for academic integrity and student learning," J. Appl. Learn. Teach., vol. 6, no. 10, pp. 1–10, 2023, https://doi.org/10.37074/jalt.2023.6.1.17.

[11] M. Khalil and E. Er, "Will ChatGPT get you caught? Rethinking of plagiarism detection," arXiv, 2023, https://doi.org/10.35542/osf.io/fnh48.

[12] M. A. Afnan et al., "ChatGPT as an educational tool: Opportunities, challenges, and recommendations for communication, business writing, and composition courses," J. Artif. Intell. Technol., 2023, https://doi.org/10.37965/jait.2023.0184.

[13] F. J. García-Peñalvo, "The perception of artificial intelligence in educational contexts after the launch of ChatGPT: Disruption or panic?" Educ. Knowl. Soc., vol. 24, Article e31279, 2023, https://doi.org/10.14201/eks.31279.

[14] N. Anderson et al., "AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in sports & exercise medicine manuscript generation," BMJ Open Sport Exerc. Med., vol. 9, e001568, 2023, https://doi.org/10.1136/bmjsem-2023-001568.

[15] R. J. M. Ventayen, “OpenAI ChatGPT generated results: Similarity index of artificial intelligence-based contents," Adv. Intell. Syst. Comput., 2023, https://ssrn.com/abstract=4332664 or http://dx.doi.org/10.2139/ssrn.4332664.

[16] A. M. Elkhatat, K. Elsaid, and S. Almeer, “Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text,” Int. J. Educ. Integr., vol. 19, no. 17, 2023, https://doi.org/10.1007/s40979-023-00140-5.

[17] A. Pegoraro et al., “To ChatGPT, or not to ChatGPT: That is the question!” arXiv, 2023, https://doi.org/10.48550/arXiv.2304.01487.

[18] Turnitin, “Turnitin announces AI writing detector and AI writing resource center for educators.” Turnitin.com. Accessed: Apr. 20, 2023. https://www.turnitin.com/press/turnitin-announces-ai-writing-detector-and-ai-writing-resource-center-for-educators#:~:text=OAKLAND%2C%20Calif.,1%2F100%20false%20positive%20rate.

[19] M. N. Halgamuge, “The use and analysis of anti-plagiarism software: Turnitin tool for formative assessment and feedback,” Comput. Appl. Eng. Educ., vol. 25, no. 6, pp. 895–909, 2017.

[20] S. A. Meo and M. Talha, “Turnitin: Is it a text matching or plagiarism detection tool?” Saudi J. Anaesth., vol. 13, Suppl 1, pp. S48–S51, 2019, https://doi.org/10.4103/sja.SJA_772_18.

[21] Turnitin, “A new path and purpose for Turnitin.” Turnitin.com. Accessed: Apr. 20, 2023. https://www.turnitin.com/blog/a-new-path-and-purpose-for-turnitin