Elkhatat *et al.*
International Journal for Educational Integrity (2023) 19:17
https://doi.org/10.1007/s40979-023-00140-5

International Journal for Educational Integrity

ORIGINAL ARTICLE Open Access

Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text

Ahmed M. Elkhatat[1*], Khaled Elsaid[2] and Saeed Almeer[3]

*Correspondence:
ahmed.elkhatat@qu.edu.qa

[1] Department of Chemical Engineering, Qatar University, P.O. 2713, Doha, Qatar
[2] Chemical Engineering Program, Texas A&M University at Qatar, P.O. 23874, Doha, Qatar
[3] Department of Chemistry and Earth Sciences, Qatar University, P.O. 2713, Doha, Qatar

Abstract

The proliferation of artificial intelligence (AI)-generated content, particularly from models like ChatGPT, presents potential challenges to academic integrity and raises concerns about plagiarism. This study investigates the capabilities of various AI content detection tools in discerning human and AI-authored content. Fifteen paragraphs each from ChatGPT Models 3.5 and 4 on the topic of cooling towers in the engineering process and five human-witten control responses were generated for evaluation. AI content detection tools developed by OpenAI, Writer, Copyleaks, GPTZero, and CrossPlag were used to evaluate these paragraphs. Findings reveal that the AI detection tools were more accurate in identifying content generated by GPT 3.5 than GPT 4. However, when applied to human-written control responses, the tools exhibited inconsistencies, producing false positives and uncertain classifications. This study underscores the need for further development and refinement of AI content detection tools as AI-generated content becomes more sophisticated and harder to distinguish from human-written text.

Keywords: AI-generated content, Plagiarism, Academic integrity, ChatGPT, AI content detection tools

Introduction

The instances of academic plagiarism have escalated in educational settings, as it has been identified in various student work, encompassing reports, assignments, projects, and beyond. Academic plagiarism can be defined as the act of employing ideas, content, or structures without providing sufficient attribution to the source (Fishman 2009). Students' plagiarism strategies differ, with the most egregious instances involving outright replication of source materials. Other approaches include partial rephrasing through modifications in grammatical structures, substituting words with their synonyms, and using online paraphrasing services to reword text (Elkhatat 2023; Meuschke & Gipp 2013; Sakamoto & Tsuda 2019). Academic plagiarism violates ethical principles and ranks among the most severe cases of misconduct, as it jeopardizes the acquisition and assessment of competencies. As a result, implementing strategies to reduce plagiarism

is vital for preserving academic integrity and preventing such dishonest practices in students' future scholarly and professional endeavors (Alsallal et al. 2013; Elkhatat 2022; Foltýnek et al. 2020). Text-Matching Software Products (TMSPs) are powerful instruments that educational institutions utilize to detect specific sets of plagiarism, attributed to their sophisticated text-matching algorithms and extensive databases containing web pages, journal articles, periodicals, and other publications. Certain TMSPs also enhance their efficacy in identifying plagiarism by incorporating databases that index previously submitted student papers (Elkhatat et al. 2021).

Recently, Artificial Intelligence (AI)-driven ChatGPT has surfaced as a tool that aids students in creating tailored content based on prompts by employing natural language processing (NLP) techniques (Radford et al. 2018). The initial GPT model showcased the potential of combining unsupervised pre-training with supervised fine-tuning for a broad array of NLP tasks. Following this, OpenAI introduced ChatGPT (model 2), which enhanced the model's performance by enlarging the architecture and using a more comprehensive pre-training dataset (Radford et al. 2019). The subsequent launch of ChatGPT (models 3 and 3.5) represented a significant advancement in ChatGPT's development, as it exhibited exceptional proficiency in producing human-like text and attained top results on various NLP benchmark lines. This model's capacity to generate contextually appropriate and coherent text in response to user prompts made it suitable for release of ChatGPT, an AI-driven chatbot aimed at helping users produce text and participate in natural language dialogues(Brown et al. 2020; OpenAI 2022).

The recently unveiled ChatGPT (model 4) by OpenAI on March 14, 2023, is a significant milestone in NLP technology. With enhanced cybersecurity safety measures and superior response quality, it surpasses its predecessors in tackling complex challenges. ChatGPT (model 4) boasts a wealth of general knowledge and problem-solving skills, enabling it to manage demanding tasks with heightened precision. Moreover, its inventive and cooperative features aid in generating, editing, and iterating various creative and technical writing projects, such as song composition, screenplay development, and personal writing style adaptation. However, it is crucial to acknowledge that ChatGPT (model 4)'s knowledge is confined to the cutoff date of September 2021 (OpenAI 2023), although the recently embedded plugins allow it to access current website content.

This development presents potential risks concerning cheating and plagiarism, which may result in severe academic and legal ramifications (Foltýnek et al. 2019). These potentially elevated risks of cheating and plagiarism include but are not limited to the Ease of Access to Information with its extensive knowledge base and ability to generate coherent and contextually relevant responses. In addition, the Adaptation to Personal Writing Style allows for generating content that closely matches a student's writing, making it even more difficult for educators to identify whether a language model has generated the work(OpenAI 2023).

Academic misconduct in undergraduate education using ChatGPT has been widely studied (Crawford et al. 2023; King & chatGpt 2023; Lee 2023; Perkins 2023; Sullivan; et al. 2023). Despite the advantages of ChatGPT for supporting students in essay composition and other scholarly tasks, questions have been raised regarding the authenticity and suitability of the content generated by the chatbot for academic purposes (King & chatGpt 2023). Additionally, ChatGPT has been rightly criticized for generating

incoherent or erroneous content (Gao et al. 2022; Qadir 2022), providing superficial information (Frye 2022), and having a restricted knowledge base due to its lack of internet access and dependence on data up until September 2021 (Williams 2022). Nonetheless, the repeatability (repeatedly generated responses within the same chatbot prompt) and reproducibility (repeatedly generated responses with a new chatbot prompt)of authenticity capabilities in GPT-3.5 and GPT-4 were examined by text-matching software, demonstrating that the generation of responses remains consistently elevated and coherent, predominantly proving challenging to detect by conventional text-matching tools (Elkhatat 2023).

Recently, Open AI classifier tools have become relied upon for distinguishing between human writing and AI-generated content, ensuring text authenticity across various applications. For instance, OpenAI, which developed ChatGPT, introduced an AI text classifier that assists users in determining whether an essay was authored by a human or generated by AI. This classifier categorizes documents into five levels based on the likelihood of being AI-generated: very unlikely, unlikely, unclear, possibly, and likely AI-generated. The OpenOpen AI classifier has been trained using a diverse range of human-written texts, although the training data does not encompass every type of human-written text. Furthermore, the developers' tests reveal that the classifier accurately identifies 26% of AI-written text (true positives) as "likely AI-generated" while incorrectly labeling 9% of the human-written text (false positives) as AI-generated (Kirchner et al. 2023). Hence, OpenAI advises users to treat the classifier's results as supplementary information rather than relying on them exclusively for determining AI-generated content (Kirchner et al. 2023). Other AI text classifier tools include Writer.com's AI content detector, which offers a limited application programming interface API-based solution for detecting AI-generated content and emphasizes its suitability for content marketing. Copyleaks, an AI content detection solution, claims a 99% accuracy rate and provides integration with many Learning Management Systems (LMS) and APIs. GPTZero, developed by Edward Tian, is an Open AI classifier tool targeting educational institutions to combat AI plagiarism by detecting AI-generated text in student assignments. Lastly, CrossPlag's AI content detector employs machine learning algorithms and natural language processing techniques to precisely predict a text's origin, drawing on patterns and characteristics identified from an extensive human and AI-generated content dataset.

The development and implementation of AI content detectors and classifier tools underscore the growing importance and need to differentiate between human-written and AI-generated content across various fields, such as education and content marketing. To date, no studies have comprehensively examined the abilities of these AI content detectors and classifiers to distinguish between human and AI-generated content. The present study aims to investigate the capabilities of several recently launched AI content detectors and classifier tools in discerning human-written and AI-generated content.

Methodology

The ChatGPT chatbot generated two 15-paragraph responses on "Application of Cooling Towers in the Engineering Process." The first set was generated using ChatGPT's Model 3.5, while the second set was created using Model 4. The initial prompt was to "write

around 100 words on the application of cooling towers in the engineering process." Five human-written samples were incorporated as control samples to evaluate false positive responses by AI detectors, as detailed in Table 1. These samples were chosen from the introduction sections of five distinct lab reports penned by undergraduate chemical engineering students. The reports were submitted and evaluated in 2018, a planned selection to ensure no interference from AI tools available at that time.

Five AI text content detectors, namely OpenAI, Writer, Copyleaks, GPTZero, and CrossPlag, were selected and evaluated for their ability to differentiate between human and AI-generated content. These AI detectors were selected based on extensive online research and valuable feedback from individual educators at the time of the study. It is important to note that this landscape is continually evolving, with new tools and websites expected to be launched shortly. Some tools, like the Turnitin AI detector, have already been introduced but are yet to be widely adopted or activated across educational institutions. In addition, the file must have at least 300 words of prose text in a long-form writing format (Turnitin 2023).

It is important to note that different AI content detection tools display their results in distinct representations, as summarized in Table 2. To standardize the results across all detection tools, we normalized them according to the OpenAI theme. This normalization was based on the AI content percentage. Texts with less than 20% AI content were classified as "very unlikely AI-generated," those with 20–40% AI content were considered "unlikely AI-generated," those with 40–60% AI content were deemed "unclear if AI-generated," those with 60–80% AI content were labeled "possibly AI-generated." Those with over 80% AI content were categorized as "likely AI-generated." Statistical analysis and capabilities tests were conducted using Minitab (Minitab 2023).

The diagnostic accuracy of AI detector responses was classified into positive, negative, false positive, false negative, and uncertain based on the original content's nature

Table 1 Codings of AI-generated and Human-written content

AI-generated content		Human-written content
ChatGPT (Model 3.5)	ChatGPT (Model 4)	
GPT 3.5_1	GPT 4_1	Human 1
GPT 3.5_2	GPT 4_2	Human 2
GPT 3.5_3	GPT 4_3	Human 3
GPT 3.5_4	GPT 4_4	Human 4
GPT 3.5_5	GPT 4_5	Human 5
GPT 3.5_6	GPT 4_6	
GPT 3.5_7	GPT 4_7	
GPT 3.5_8	GPT 4_8	
GPT 3.5_9	GPT 4_9	
GPT 3.5_10	GPT 4_10	
GPT 3.5_11	GPT 4_11	
GPT 3.5_12	GPT 4_12	
GPT 3.5_13	GPT 4_13	
GPT 3.5_14	GPT 4_14	
GPT 3.5_15	GPT 4_15	

Table 2 Results representation of AI content detectors

AI detection tool	Results representation
The Open AI classifier	Very unlikely AI-generated
	Unlikely AI-generated
	Unclear if AI-generated
	Possibly AI-generated
	Likely AI-generated
Writer's	The percentage of human-written content
Corssplag's	The percentage of AI-generated content $\geq$ 80% primarily AI-generated 20–80% mixed origin $\leq$ 20% primarily human-written
Copyleaks	Binary result of the probability percentage for AI or human authorship
GPTzero	Likelihood of being entirely a human-written or an AI-generated

(AI-generated or human-written). The AI detector responses were classified as positive if the original content was AI-generated and the detector output was "Likely AI-generated" or, more inclusively, "Possibly AI-generated." Negative responses arise when the original content is human-generated, and the detector output is "Very unlikely AI-generated" or, more inclusively, "Unlikely AI-generated." False positive responses occur when the original content is human-generated, and the detector output is "Likely AI-generated" or "Possibly AI-generated." In contrast, false negative responses emerge when the original content is AI-generated, and the detector output is "Very unlikely AI-generated" or "Unlikely AI-generated." Finally, uncertain responses are those where the detector output is "Unclear if it is AI-generated," regardless of whether the original content is AI-generated or human-generated. This classification scheme assumes that "Possibly AI-generated" and "Unlikely AI-generated" responses could be considered borderline cases, falling into either positive/negative or false positive/false negative categories based on the desired level of inclusivity or strictness in the classification process.

This study evaluated these five detectors, OpenAI, Writer, Copyleaks, GPTZero, and CrossPlag, focusing on their Specificity, Sensitivity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV). These metrics are used in biostatistics and machine learning to evaluate the performance of binary classification tests. Sensitivity (True Positive Rate) is the proportion of actual positive cases which are correctly identified. In this context, sensitivity is defined as the proportion of AI-generated content correctly identified by the detectors out of all AI-generated content. It is calculated as the ratio of true positives (AI-generated content correctly identified) to the sum of true positives and false negatives (AI-generated content incorrectly identified as human-generated) (Nelson et al. 2001; Nhu et al. 2020).

On the other hand, Specificity (True Negative Rate) is the proportion of actual negative cases which are correctly identified. In this context, it refers to the proportion of human-generated content correctly identified by the detectors out of all actual human-generated content. It is computed as the ratio of true negatives (human-generated content correctly identified) to the sum of true negatives and false positives

(human-generated content incorrectly identified as AI-generated) (Nelson et al. 2001; Nhu et al. 2020).

Predictive power, a vital determinant of the detectors' efficacy, is divided into positive predictive value (PPV) and negative predictive value (NPV). Positive Predictive Value (PPV) is the proportion of positive results in statistics and diagnostic tests that are actually positive results. In this context, it is the proportion of actual AI-generated content among all content identified as AI-generated by the detectors. It is calculated as the ratio of true positives to the sum of true and false positives. Conversely, Negative Predictive Value (NPV) is the proportion of negative results in statistics and diagnostic tests that are accurate negative results.in this context, it is the proportion of actual human-generated content among all content identified as human-generated by the detectors. It is calculated as the ratio of true negatives to the sum of true and false negatives (Nelson et al. 2001; Nhu et al. 2020). These metrics provide a robust framework for evaluating the performance of AI text content detectors; collectively, they can be called "Classification Performance Metrics" or "Binary Classification Metrics."

Results

Table 3 outlines the outcomes of AI content detection tools implemented on 15 paragraphs generated by ChatGPT Model 3.5, 15 more from ChatGPT Model 4, and five control paragraphs penned by humans. It is important to emphasize that, as stated in the methodology section and detailed in Table 2, different AI content detection tools display their results in distinct representations. For instance, GPTZERO classifies the content into two groups: AI-Generated or Human-Generated content. In contrast, the Open-Open AI classifier divides the content into a quintuple classification system: Likely AI-Generated, Possibly AI-Generated, Unclear if it is AI-Generated, Unlikely AI-Generated, and Very Unlikely AI-Generated. Notably, both GPTZERO and the OpenOpen AI classifier do not disclose the specific proportions of AI or human contribution within the content. In contrast, other AI detectors provide percentages detailing the AI or human contribution in the submitted text. Therefore, to standardize the responses from all AI detectors, the percentage data were normalized to fit the five-tier classification system of the OpenOpen AI classifier, where each category represents a 20% increment. The table also includes the exact percentage representation of AI contribution within each category for enhanced clarity and specificity.

Table 4, on the other hand, demonstrates the diagnostic accuracy of these AI detection tools in differentiating between AI-generated and human-written content. The results for GPT 3.5-generated content indicate a high degree of consistency among the tools. The AI-generated content was often correctly identified as "Likely AI-Generated." However, there were a few instances where the tools provided an uncertain or false-negative classification. GPT 3.5_7 and GPT 3.5_14 received "Very unlikely AI-Generated" ratings from GPTZERO, while WRITER classified GPT 3.5_9 and GPT 3.5_14 as "Unclear if AI-Generated." Despite these discrepancies, most GPT 3.5-generated content was correctly identified as AI-generated by all tools.

The performance of the tools on GPT 4-generated content was notably less consistent. While some AI-generated content was correctly identified, there were several false negatives and uncertain classifications. For example, GPT 4_1, GPT 4_3,

Table 3 The responses of five AI text content detectors for GPT-3.5, GPT-4, and Human-written contents

Generated Content	OpenOpen AI classifier	WRITER		CROSSPLAG		COPYLEAKS		GPTZERO
	Response	Response	% of AI Content	Response	% of AI Content	Response	% of AI Content	Response
GPT 3.5_1	Likely AI-Generated	Likely AI-Generated	100%	Likely AI-Generated	100%	Likely AI-Generated	99.9%	Likely AI-Generated
GPT 3.5_2	Likely AI-Generated	Likely AI-Generated	98%	Likely AI-Generated	100%	Likely AI-Generated	99.9%	Likely AI-Generated
GPT 3.5_3	Likely AI-Generated	Possibly AI-Generated	66%	Likely AI-Generated	99%	Likely AI-Generated	99.9%	Likely AI-Generated
GPT 3.5_4	Likely AI-Generated	Likely AI-Generated	97%	Likely AI-Generated	100%	Likely AI-Generated	99.9%	Likely AI-Generated
GPT 3.5_5	Likely AI-Generated	Likely AI-Generated	93%	Likely AI-Generated	100%	Likely AI-Generated	99.9%	Likely AI-Generated
GPT 3.5_6	Likely AI-Generated	Likely AI-Generated	87%	Likely AI-Generated	100%	Likely AI-Generated	99.9%	Likely AI-Generated
GPT 3.5_7	Likely AI-Generated	Very unlikely AI-Generated	6%	Likely AI-Generated	100%	Likely AI-Generated	99.9%	Likely AI-Generated
GPT 3.5_8	Likely AI-Generated	Likely AI-Generated	88%	Likely AI-Generated	100%	Likely AI-Generated	99.9%	Likely AI-Generated
GPT 3.5_9	Likely AI-Generated	Unclear if it is AI-Generated	49%	Likely AI-Generated	100%	Likely AI-Generated	99.9%	Likely AI-Generated
GPT 3.5_10	Likely AI-Generated	Likely AI-Generated	100%	Likely AI-Generated	100%	Likely AI-Generated	99.9%	Likely AI-Generated
GPT 3.5_11	Likely AI-Generated	Likely AI-Generated	100%	Likely AI-Generated	100%	Likely AI-Generated	99.9%	Likely AI-Generated
GPT 3.5_12	Likely AI-Generated	Likely AI-Generated	96%	Likely AI-Generated	100%	Likely AI-Generated	99.9%	Likely AI-Generated
GPT 3.5_13	Likely AI-Generated	Likely AI-Generated	96%	Likely AI-Generated	100%	Likely AI-Generated	99.9%	Likely AI-Generated
GPT 3.5_14	Likely AI-Generated	Unclear if it is AI-Generated	52%	Likely AI-Generated	100%	Likely AI-Generated	99.9%	Very unlikely AI-Generated
GPT 3.5_15	Likely AI-Generated	Likely AI-Generated	82%	Likely AI-Generated	100%	Likely AI-Generated	99.9%	Likely AI-Generated
GPT 4_1	Likely AI-Generated	Very unlikely AI-Generated	9%	Very unlikely AI-Generated	12%	Likely AI-Generated	82.6%	Very unlikely AI-Generated
GPT 4_2	Likely AI-Generated	Unclear if it is AI-Generated	47%	Likely AI-Generated	88%	Likely AI-Generated	99.3%	Very unlikely AI-Generated
GPT 4_3	Possibly AI-Generated	Very unlikely AI-Generated	4%	Very unlikely AI-Generated	16%	Likely AI-Generated	84.7%	Very unlikely AI-Generated
GPT 4_4	Possibly AI-Generated	Very unlikely AI-Generated	5%	Unlikely AI-Generated	32%	Likely AI-Generated	86.0%	Very unlikely AI-Generated
GPT 4_5	Unclear if it is AI-Generated	Very unlikely AI-Generated	3%	Very unlikely AI-Generated	1%	Possibly AI-Generated	76.3%	Very unlikely AI-Generated
GPT 4_6	Possibly AI-Generated	Very unlikely AI-Generated	1%	Likely AI-Generated	87%	Likely AI-Generated	99.5%	Very unlikely AI-Generated
GPT 4_7	Likely AI-Generated	Likely AI-Generated	95%	Very unlikely AI-Generated	6%	Likely AI-Generated	83.8%	Very unlikely AI-Generated

Table 3 (continued)

Generated Content	OpenOpen AI classifier	WRITER		CROSSPLAG		COPYLEAKS		GPTZERO
	Response	Response	% of AI Content	Response	% of AI Content	Response	% of AI Content	Response
GPT 4_8	Possibly AI-Generated	Unclear if it is AI-Generated	57%	Possibly AI-Generated	77%	Likely AI-Generated	97.3%	Very unlikely AI-Generated
GPT 4_9	Possibly AI-Generated	Unlikely AI-Generated	23%	Possibly AI-Generated	63%	Likely AI-Generated	95.7%	Very unlikely AI-Generated
GPT 4_10	Possibly AI-Generated	Very unlikely AI-Generated	13%	Very unlikely AI-Generated	3%	Likely AI-Generated	81.9%	Likely AI-Generated
GPT 4_11	Possibly AI-Generated	Unclear if it is AI-Generated	51%	Likely AI-Generated	81%	Likely AI-Generated	97.2%	Likely AI-Generated
GPT 4_12	Possibly AI-Generated	Very unlikely AI-Generated	16%	Very unlikely AI-Generated	1%	Possibly AI-Generated	80.0%	Likely AI-Generated
GPT 4_13	Unclear if it is AI-Generated	Very unlikely AI-Generated	0%	Very unlikely AI-Generated	1%	Possibly AI-Generated	80.0%	Very unlikely AI-Generated
GPT 4_14	Possibly AI-Generated	Very unlikely AI-Generated	18%	Very unlikely AI-Generated	2%	Very unlikely AI-Generated	0.7%	Very unlikely AI-Generated
GPT 4_15	Possibly AI-Generated	Unlikely AI-Generated	32%	Possibly AI-Generated	69%	Likely AI-Generated	96.1%	Likely AI-Generated
Human 1	Likely AI-Generated	Likely AI-Generated	92%	Very unlikely AI-Generated	3%	Very unlikely AI-Generated	7.6%	Very unlikely AI-Generated
Human 2	Likely AI-Generated	Likely AI-Generated	98%	Very unlikely AI-Generated	5%	Likely AI-Generated	99.9%	Likely AI-Generated
Human 3	Possibly AI-Generated	Very unlikely AI-Generated	0%	Very unlikely AI-Generated	1%	Very unlikely AI-Generated	0.1%	Very unlikely AI-Generated
Human 4	Likely AI-Generated	Very unlikely AI-Generated	2%	Unlikely AI-Generated	28%	Unlikely AI-Generated	20.2%	Very unlikely AI-Generated
Human 5	Likely AI-Generated	Unclear if it is AI-Generated	54%	Very unlikely AI-Generated	1%	Very unlikely AI-Generated	4.2%	Very unlikely AI-Generated

and GPT 4_4 received "Very unlikely AI-Generated" ratings from WRITER, CROSSPLAG, and GPTZERO. Furthermore, GPT 4_13 was classified as "Very unlikely AI-Generated" by WRITER and CROSSPLAG, while GPTZERO labeled it as "Unclear if it is AI-Generated." Overall, the tools struggled more with accurately identifying GPT 4-generated content than GPT 3.5-generated content.

When analyzing the control responses, it is evident that the tools' performance was not entirely reliable. While some human-written content was correctly classified as "Very unlikely AI-Generated" or "Unlikely AI-Generated," there were false positives and uncertain classifications. For example, WRITER ranked Human 1 and 2 as "Likely AI-Generated," while GPTZERO provided a "Likely AI-Generated" classification for Human 2. Additionally, Human 5 received an "Uncertain" classification from WRITER.

Table 4 The diagnostic accuracy of AI detector responses

Response	WRITER	CROSSPLAG	GPTZERO	COPYLEAKS	OpenOpen AI calssifier
GPT 3.5_1	Positive	Positive	Positive	Positive	Positive
GPT 3.5_2	Positive	Positive	Positive	Positive	Positive
GPT 3.5_3	Positive	Positive	Positive	Positive	Positive
GPT 3.5_4	Positive	Positive	Positive	Positive	Positive
GPT 3.5_5	Positive	Positive	Positive	Positive	Positive
GPT 3.5_6	Positive	Positive	Positive	Positive	Positive
GPT 3.5_7	False Negative	Positive	Positive	Positive	Positive
GPT 3.5_8	Positive	Positive	Positive	Positive	Positive
GPT 3.5_9	Uncertain	Positive	Positive	Positive	Positive
GPT 3.5_10	Positive	Positive	Positive	Positive	Positive
GPT 3.5_11	Positive	Positive	Positive	Positive	Positive
GPT 3.5_12	Positive	Positive	Positive	Positive	Positive
GPT 3.5_13	Positive	Positive	Positive	Positive	Positive
GPT 3.5_14	Uncertain	Positive	False Negative	Positive	Positive
GPT 3.5_15	Positive	Positive	Positive	Positive	Positive
GPT 4_1	False Negative	False Negative	False Negative	Positive	Positive
GPT 4_2	Uncertain	Positive	False Negative	Positive	Positive
GPT 4_3	False Negative	False Negative	False Negative	Positive	Positive
GPT 4_4	False Negative	False Negative	False Negative	Positive	Positive
GPT 4_5	False Negative	False Negative	False Negative	Positive	Uncertain
GPT 4_6	False Negative	Positive	False Negative	Positive	Positive
GPT 4_7	Positive	False Negative	False Negative	Positive	Positive
GPT 4_8	Uncertain	Positive	False Negative	Positive	Positive
GPT 4_9	False Negative	Positive	False Negative	Positive	Positive
GPT 4_10	False Negative	False Negative	Positive	Positive	Positive
GPT 4_11	Uncertain	Positive	Positive	Positive	Positive
GPT 4_12	False Negative	False Negative	Positive	Positive	Positive
GPT 4_13	False Negative	False Negative	False Negative	Positive	Uncertain
GPT 4_14	False Negative	False Negative	False Negative	False Negative	Positive
GPT 4_15	False Negative	Positive	Positive	Positive	Positive
Human 1	False Positive	Negative	Negative	Negative	False Positive
Human 2	False Positive	Negative	False Positive	False Positive	False Positive
Human 3	Negative	Negative	Negative	Negative	False Positive
Human 4	Negative	Negative	Negative	Negative	False Positive
Human 5	Uncertain	Negative	Negative	Negative	False Positive

In order to effectively illustrate the distribution of discrete variables, the Tally Individual Variables function in Minitab was employed. This method facilitated the visualization of varying categories or outcomes' frequencies, thereby providing valuable insights into the inherent patterns within the dataset. To further enhance comprehension, the outcomes of the Tally analysis were depicted using bar charts, as demonstrated in Figs. 1, 2, 3, 4, 5 and 6. Moreover, the classification performance metrics of these five AI text content are demonstrated in Fig. 7, indicating a varied performance across different metrics. Looking at the GPT 3.5 results, the OpenAI Classifier displayed the highest sensitivity, with a score of 100%, implying that it correctly identified all AI-generated content. However, its specificity and NPV were the lowest, at 0%, indicating a limitation in correctly identifying human-generated content and giving pessimistic predictions

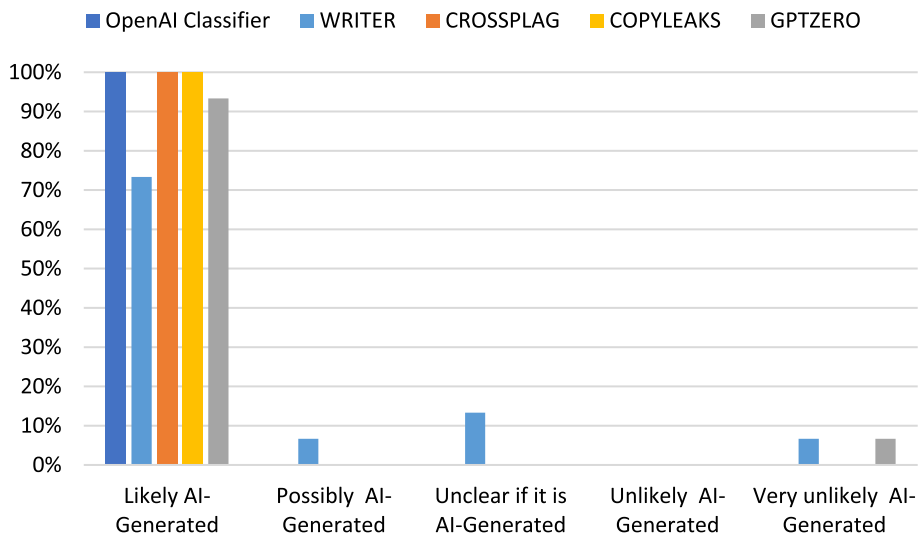

Fig. 1 The responses of five AI text content detectors for GPT-3.5 generated contents

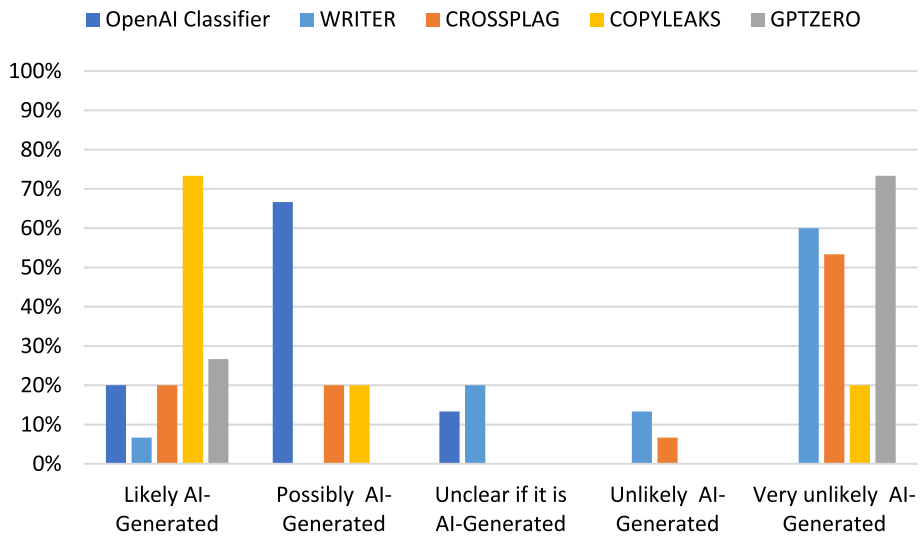

Fig. 2 The diagnostic accuracy of the AI text content detectors' responses for GPT-3.5 generated contents

when it was genuinely human-generated. GPTZero exhibited a balanced performance, with a sensitivity of 93% and specificity of 80%, while Writer and Copyleaks struggled with sensitivity. The results for GPT 4 were generally lower, with Copyleaks having the highest sensitivity, 93%, and CrossPlag maintaining 100% specificity. The OpenAI Classifier demonstrated substantial sensitivity and NPV but no specificity.

Discussion

The analysis focuses on the performance of five AI text content detectors developed by OpenAI, Writer, Copyleaks, GPTZero, and CrossPlag corporations. These tools were utilized to evaluate the generated content and determine the effectiveness of each detector in correctly identifying and categorizing the text as either AI-generated or

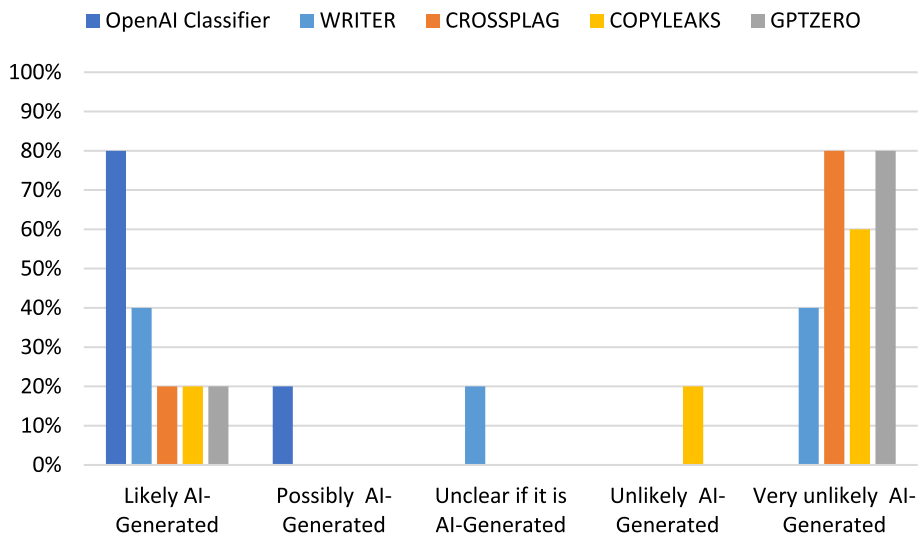

Fig. 3 The responses of five AI text content detectors for GPT-4 generated contents

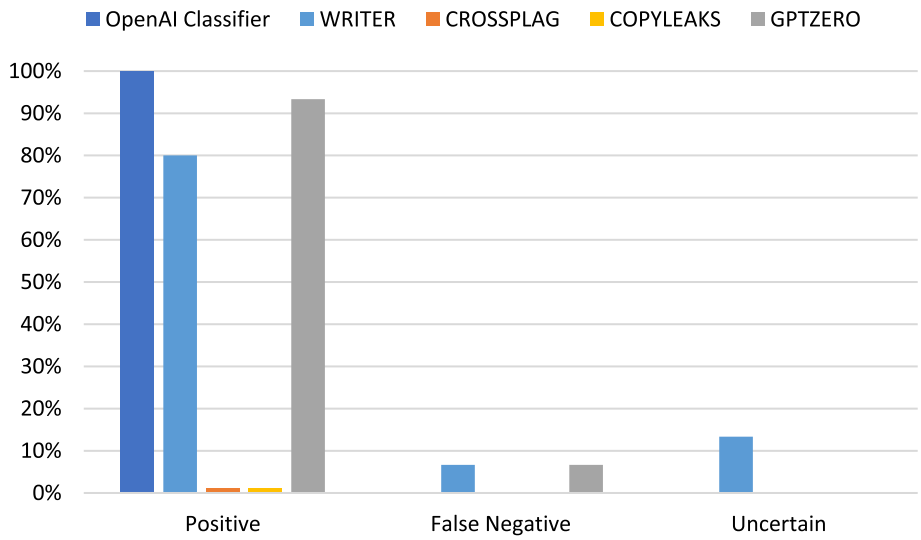

Fig. 4 The diagnostic accuracy of the AI text content detectors' responses for GPT-4 generated contents

human-written. The results indicate a variance in the performance of these tools across GPT 3.5, GPT 4, and human-generated content. While the tools were generally more successful in identifying GPT 3.5-generated content, they struggled with GPT 4-generated content and exhibited inconsistencies when analyzing human-written control responses. The varying degrees of performance across these AI text content detectors highlight the complexities and challenges associated with differentiating between human and AI-generated content.

The OpenAI Classifier's high sensitivity but low specificity in both GPT versions suggest that it is efficient at identifying AI-generated content but might struggle to identify human-generated content accurately. CrossPlag's high specificity indicates its ability to identify human-generated content correctly but struggles to identify

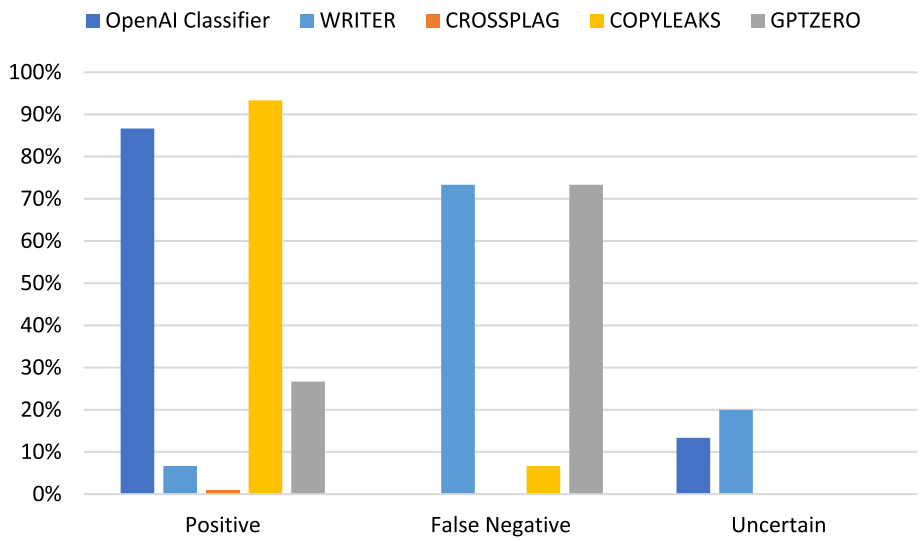

Fig. 5 The responses of five AI text content detectors for human-written contents

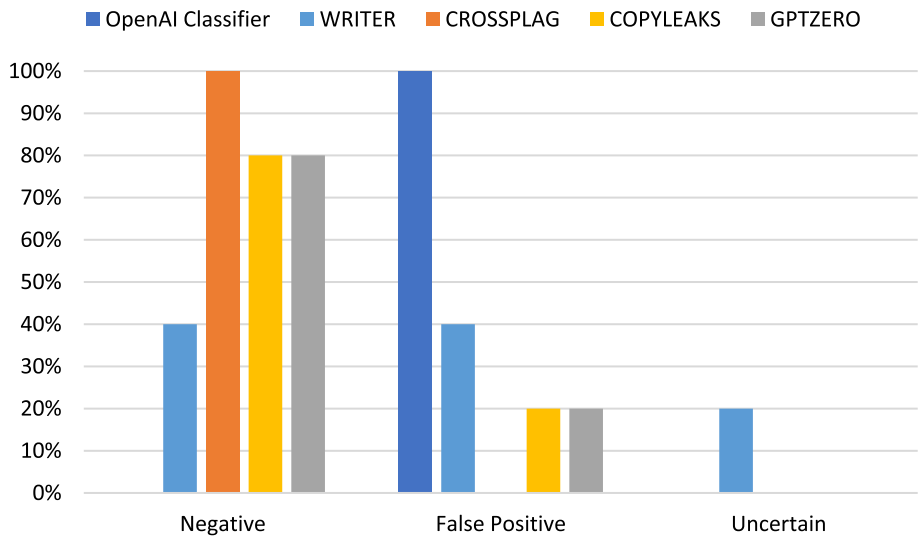

Fig. 6 The diagnostic accuracy of the AI text content detectors' responses for the human-written contents

AI-generated content, especially in the GPT 4 version. These findings raise questions about its effectiveness in the rapidly advancing AI landscape.

The differences between the GPT 3.5 and GPT 4 results underline the evolving challenge of AI-generated content detection, suggesting that detector performance can significantly vary depending on the AI model's sophistication. These findings have significant implications for plagiarism detection, highlighting the need for ongoing advancements in detection tools to keep pace with evolving AI text generation capabilities.

Notably, the study's findings underscore the need for a nuanced understanding of the capabilities and limitations of these technologies. While this study indicates that AI-detection tools can distinguish between human and AI-generated content

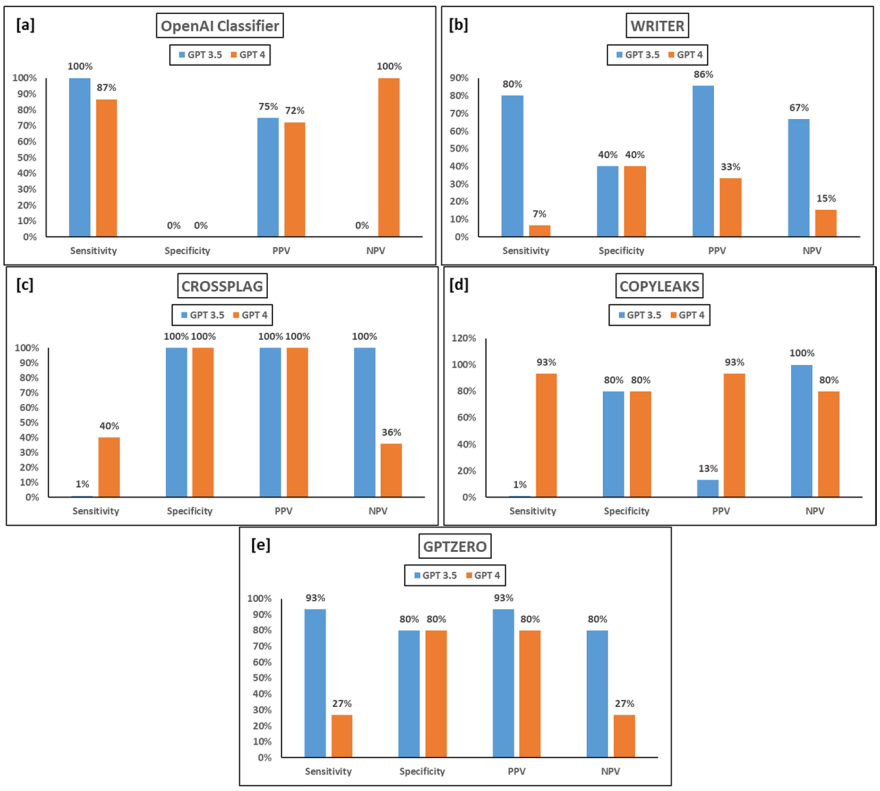

Fig. 7 The Classification Performance Metrics of (a) OpenAI Classifier, (b) WRITER, (c) CROSSPLAG, (d) COPYLEAKS, and (e) GPTZERO

to a certain extent, their performance is inconsistent and varies depending on the sophistication of the AI model used to generate the content. This inconsistency raises concerns about the reliability of these tools, especially in high-stakes contexts such as academic integrity investigations. Therefore, while AI-detection tools may serve as a helpful aid in identifying AI-generated content, they should not be used as the sole determinant in academic integrity cases. Instead, a more holistic approach that includes manual review and consideration of contextual factors should be adopted. This approach would ensure a fairer evaluation process and mitigate the ethical concerns of using AI detection tools.

It is important to emphasize that the advent of AI and other digital technologies necessitates rethinking traditional assessment methods. Rather than resorting solely to methods less vulnerable to AI cheating, educational institutions should also consider leveraging these technologies to enhance learning and assessment. For instance, AI could provide personalized feedback, facilitate peer review, or even create more complex and realistic assessment tasks that are difficult to cheat. In addition, it is essential to note that academic integrity is not just about preventing cheating but also about fostering a culture of honesty and responsibility. This involves educating students about the importance of academic integrity and the consequences of academic misconduct and providing them with the necessary skills and resources to avoid plagiarism and other forms of cheating.

Limitation

The limitations of this study, such as the tools used, the statistics included, and the disciplinary specificity against which these tools are evaluated, need to be acknowledged. It should be noted that the tools analyzed in this study were only those developed by OpenAI, Writer, Copyleaks, GPTZero, and CrossPlag corporations. These AI detectors were selected based on extensive online research and valuable feedback from individual educators at the time of the study. It is important to note that this landscape is continually evolving, with new tools and websites expected to be launched shortly. Some tools, like the Turnitin AI detector, have already been introduced but are yet to be widely adopted or activated across educational institutions. In addition, the file must have at least 300 words of prose text in a long-form writing format. Moreover, the content used for testing the tools was generated by ChatGPT Models 3.5 and 4 and included only five human-written control responses. The sample size and nature of content could affect the findings, as the performance of these tools might differ when applied to other AI models or a more extensive, more diverse set of human-written content.

It is essential to mention that this study was conducted at a specific time. Therefore, the performance of the tools might have evolved, and they might perform differently on different versions of AI models that have been released after this study was conducted. Future research should explore techniques to increase both sensitivity and specificity simultaneously for more accurate content detection, considering the rapidly evolving nature of AI content generation.

Conclusion

The present study sought to evaluate the performance of AI text content detectors, including OpenAI, Writer, Copyleaks, GPTZero, and CrossPlag. The results of this study indicate considerable variability in the tools' ability to correctly identify and categorize text as either AI-generated or human-written, with a general trend showing a better performance when identifying GPT 3.5-generated content compared to GPT 4-generated content or human-written content. Notably, the varying performance underscores the intricacies involved in distinguishing between AI and human-generated text and the challenges that arise with advancements in AI text generation capabilities.

The study highlighted significant performance differences between the AI detectors, with OpenAI showing high sensitivity but low specificity in detecting AI-generated content. In contrast, CrossPlag showed high specificity but struggled with AI-generated content, particularly from GPT 4. This suggests that the effectiveness of these tools may be limited in the fast-paced world of AI evolution. Furthermore, the discrepancy in detecting GPT 3.5 and GPT 4 content emphasizes the growing challenge in AI-generated content detection and the implications for plagiarism detection. The findings necessitate improvements in detection tools to keep up with sophisticated AI text generation models.

Notably, while AI detection tools can provide some insights, their inconsistent performance and dependence on the sophistication of the AI models necessitate a more holistic approach for academic integrity cases, combining AI tools with manual review and contextual considerations. The findings also call for reassessing traditional educational

methods in the face of AI and digital technologies, suggesting a shift towards AI-enhanced learning and assessment while fostering an environment of academic honesty and responsibility. The study acknowledges limitations related to the selected AI detectors, the nature of content used for testing, and the study's timing. Therefore, future research should consider expanding the selection of detectors, increasing the variety and size of the testing content, and regularly evaluating the detectors' performance over time to keep pace with the rapidly evolving AI landscape. Future research should also focus on improving sensitivity and specificity simultaneously for more accurate content detection.

In conclusion, as AI text generation evolves, so must the tools designed to detect it. This necessitates continuous development and regular evaluation to ensure their efficacy and reliability. Furthermore, a balanced approach involving AI tools and traditional methods best upholds academic integrity in an ever-evolving digital landscape.

Abbreviations

AI	Artificial Intelligence
LMS	Learning Management Systems
NLP	Natural Language Processing
NPV	Negative Predictive Value
PPV	Positive Predictive Value
TMSP	Text-Matching Software Product

Acknowledgements
The publication of this article was funded by the Qatar National Library.

Authors' contributions
Ahmed M. Elkhatat: Conceptionizaion, Conducting the experiments discussing the results, Writing the first draft. Khaled Elsaid: Validating the concepts, contributing to the discussion, and writing the second Draft. Saeed Almeer: project administration and supervision, proofreading, improving, and writing the final version.

Funding
N/A.

Availability of data and materials
All data and materials are available.

Declarations

Competing interests
The authors declare that they have no conflict of interest.

Received: 30 April 2023 Accepted: 30 June 2023
Published online: 01 September 2023

References

Alsallal M, Iqbal R, Amin S, James A (2013) Intrinsic Plagiarism Detection Using Latent Semantic Indexing and Stylometry. 2013 Sixth International Conference on Developments in eSystems Engineering

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901

Crawford J, Cowling M, Allen KA (2023) Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). J Univ Teach Learning Pract 20(3). https://doi.org/10.53761/1.20.3.02

Elkhatat AM (2023) Evaluating the Efficacy of AI Detectors: A Comparative Analysis of Tools for Discriminating Human-Generated and AI-Generated Texts. Int J Educ Integr. https://doi.org/10.1007/s40979-023-00137-0

Elkhatat AM, Elsaid K, Almeer S (2021) Some students plagiarism tricks, and tips for effective check. Int J Educ Integrity 17(1). https://doi.org/10.1007/s40979-021-00082-w

Elkhatat AM (2022) Practical randomly selected question exam design to address replicated and sequential questions in online examinations. Int J Educ Integrity 18(1). https://doi.org/10.1007/s40979-022-00103-2

Fishman T (2009) "We know it when we see it" is not good enough: toward a standard definition of plagiarism that transcends theft, fraud, and copyright 4th Asia Pacific Conference on Educational Integrity, University of Wollongong NSW Australia

Foltýnek T, Meuschke N, Gipp B (2019) Academic Plagiarism Detection. ACM Comput Surv 52(6):1–42. https://doi.org/10.1145/3345317
Foltýnek T, Meuschke N, Gipp B (2020) Academic Plagiarism Detection. ACM Comput Surv 52(6):1–42. https://doi.org/10.1145/3345317
Frye BL (2022) Should Using an AI Text Generator to Produce Academic Writing Be Plagiarism? Fordham Intellectual Property, Media & Entertainment Law Journal. https://ssrn.com/abstract=4292283
Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, Pearson AT (2022) Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. https://doi.org/10.1101/2022.12.23.521610
King MR, chatGpt (2023) A Conversation on Artificial Intelligence, Chatbots, and Plagiarism in Higher Education. Cell Mol Bioeng 16(1):1–2. https://doi.org/10.1007/s12195-022-00754-8
Kirchner JH, Ahmad L, Aaronson S, Leike J (2023) New AI classifier for indicating AI-written text. OpenAI. Retrieved 16 April from https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text
Lee H (2023) The rise of ChatGPT: Exploring its potential in medical education. Anat Sci Educ. https://doi.org/10.1002/ase.2270
Meuschke N, Gipp B (2013) State-of-the-art in detecting academic plagiarism. Int J Educ Integrity 9(1). https://doi.org/10.21913/IJEI.v9i1.847
Minitab (2023). https://www.minitab.com/en-us/
Nelson EC, Hanna GL, Hudziak JJ, Botteron KN, Heath AC, Todd RD (2001) Obsessive-compulsive scale of the child behavior checklist: specificity, sensitivity, and predictive power. Pediatrics 108(1):E14. https://doi.org/10.1542/peds.108.1.e14
Nhu VH, Mohammadi A, Shahabi H, Ahmad BB, Al-Ansari N, Shirzadi A, Clague JJ, Jaafari A, Chen W, Nguyen H (2020) Landslide Susceptibility Mapping Using Machine Learning Algorithms and Remote Sensing Data in a Tropical Environment. Int J Environ Res Public Health, 17(14). https://doi.org/10.3390/ijerph17144933
OpenAI (2022) Introducing ChatGPT. Retrieved March 21 from https://openai.com/blog/chatgpt/
OpenAI (2023) GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. Retrieved March 22 from https://openai.com/product/gpt-4
Perkins M (2023) Academic integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. J Univ Teach Learning Pract 20(2). https://doi.org/10.53761/1.20.02.07
Qadir J (2022) Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.21789434.v1
Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. OpenAI Blog 1(8):9
Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training
Sakamoto D, Tsuda K (2019) A Detection Method for Plagiarism Reports of Students. Procedia Computer Science 159:1329–1338. https://doi.org/10.1016/j.procs.2019.09.303
Sullivan M, Kelly A, Mclaughlan P (2023) ChatGPT in higher education: Considerations for academic integrity and student learning. J Appl Learning Teach 6(1). https://doi.org/10.37074/jalt.2023.6.1.17
Turnitin (2023) AI Writing Detection Frequently Asked Questions. Retrieved 21 June from https://www.turnitin.com/products/features/ai-writing-detection/faq
Williams C (2022) Hype, or the future of learning and teaching? 3 Limits to AI's ability to write student essays. The University of Kent's Academic Repository, Blog post. https://kar.kent.ac.uk/99505/

Publisher's Note