Plagiarism in AI empowered world

Aiste Steponenaite[1][0000-0002-1988-3419] and Basel Barakat [2][0000-0001-9126-7613]

[1] Medway School of Pharmacy, University of Kent, UK
[2] School of Computer Science, University of Sunderland, UK
A.Steponenaite@kent.ac.uk
Basel.Barakat@sunderland.ac.uk

Abstract. The use of Artificial Intelligence (AI) has revolutionized many aspects of education and research, but it has also introduced new challenges, including the problem of students using AI to create assignments that cannot be detected by plagiarism checkers. The proliferation of AI tools that can generate original-sounding text has made it easier for students to pass off the work of others as their own, making it more difficult for educators to identify and prevent plagiarism. This paper identifies the problem of plagiarism in the AI empowered world by comparing ChatGPT written assignments for biology and computer science. We have tested the plagiarism of those assignments in freely available tools online as well as in trusted and widely used Turnitin. We show that although the original ChatGPT written assignments sometimes result in relatively high plagiarism level, adding just one additional step of paraphrasing the work with free AI tools online significantly reduces the detected plagiarism with similarity levels laying within the acceptable range. This suggests that educational facilities should rethink of how they are assessing students' knowledge.

Keywords: Artificial Intelligence, ChatGPT, Plagiarism, Education

1 Introduction

Continuously evolving technology has provided various tools to assist with educational process. The tools range from simple methods for sharing documents, to sophisticated tools, like Virtual Learning Environments (VLEs). The last few years had a massive increase in 'smart' technologies that can learn from past data and make decisions without the need to be specifically programmed to do so, simply by applying Artificial Intelligence (AI) algorithms. AI has played a major role in several industries by automating tasks that usually require humans time and effort. It also has a great potential for enhancing the academia experience and the future of education, as using the AI tools can assist with various tasks and processes related to both education and research. These tools are designed to help educators and researchers to improve efficiency, accuracy, and productivity in their work.

Widely used AI tools in academia and teaching include tools for grading and evaluating student work, generating personalized learning materials, and supporting research and data analysis. And although these AI tools have the potential to revolutionize the way we approach education and research, it is important to carefully consider the limitations and ethical implications of them, and to ensure that they are used in a responsible and transparent manner.

ChatGPT - AI powered chatbot from San Francisco company OpenAI - has raised a lot of concerns since its launch on 30th November 2022. It is currently using GPT-3 and later in 2023 it will start using GPT-4, which is the largest language model yet that has 1 trillion parameters. ChatGPT chatbot is able to generate human-like text writing articles, emails, computer codes with little to no input from the users. Because of its high language capabilities and the quality of outputs it can present, ChatGPT, and AI tools in general, are seen as a threat to the education sector and academic integrity in online learning [1, 2].

Generating a high-quality essay with ChatGPT takes only seconds. In this paper we investigated the efficiency of plagiarism detection tools in assignments generated using AI tool ChatGPT (Dec 15 Version, OpenAI, 2022) for biology and computing fields. To add a level of complexity, we paraphrased the generated essays using free to use website paraphrasingtool.ai to test how likely students might succeed in cheating on their assessments with the help of AI and tested the plagiarism in paraphrasingtool.ai, smallseotools.com and Turnitin. We have also tested the written essays in AI generated text detection tool GPTZero [6].

2 Methods

We have created two assignments – biology and computer science based – and generated answers using the ChatGPT Dec 15 Version. First, we copied the exact assignment information, then we paraphrased the question and in both cases used the option to “regenerate response”. This way we obtained four responses for each question and used paraphrasingtool.ai to paraphrase each of the answers. This gave us 8 essays per subject which we then tested in free to use plagiarism checkers: paraphrasingtool.ai plagiarism checker, SmallSEOTools as well as Turnitin. We have used the detected plagiarism percentage to report our results. To test if the generated assignments were likely to have been written by AI, we used the free AI generated text detection tool GPTZero API version 2.0.0.

The created assignments briefs were:

Table 1. The tested assignments.

Major	Assignment Brief
Biology	Critically evaluate the physiology and pathophysiology of Alzheimer's disease and available treatments. Consider disease cause, progression, visible and diagnostically detectable symptoms and possible prognosis. Use scientific papers to write your essay and present your results in 1800-2000 words.
Computer Science	Write a report critically evaluating five machine learning for image classification algorithms. The report should be no more than 2000 words and it should include an introduction, evaluation metrics, methodology and conclusion with 10-15 references.

Data was plotted as mean ± SEM using GraphPad Prism (version 9.5.0) and significance was assessed using two-way ANOVA with Tukey multiple comparisons test.

3 Results

3.1 Plagiarism detection

Assignments written using the ChatGPT were relatively well written and covered all the aspects requested in the original assignment. However, ChatGPT generated responses did not meet the word limit requirements, references were not always to the best academic standard and in some cases, there were no references at all, and the content was shallow. Nevertheless, if a student would submit an assignment of similar quality, it would deserve a pass mark.

The plagiarism score was the highest in the text written by ChatGPT only. However, detectable text similarity was significantly reduced after using paraphrasingtool.ai in both assignment groups as shown in Fig. 1 and Fig. 2. Two-way ANOVA results for the biology assignment (see Fig. 1) showed that there is a statistically significant interaction between a text source (original ChatGPT vs paraphrased) and the tool used to check for plagiarism ($F_{(2, 18)}$=4.83, P=0.021). Simple main effects analysis showed that a text source $F_{(1, 18)}$=15.12, P=0.0011) and a plagiarism detection tool ($F_{(2, 18)}$=8.58, P=0.0024) had a statistically significant effect on plagiarism levels.

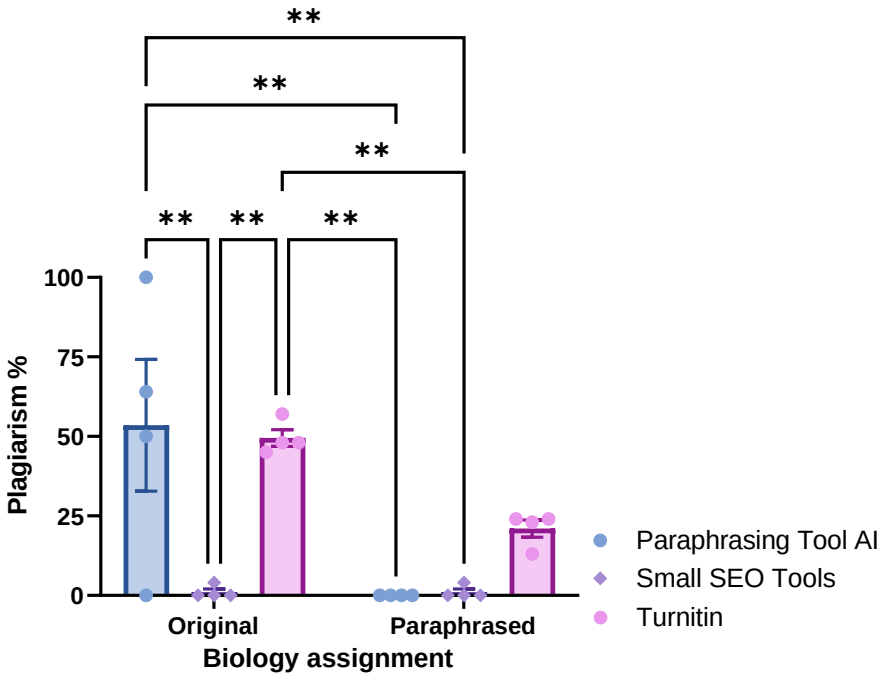

Fig. 1. The quality of AI generated biology assignments assessed in plagiarism checkers. Text generated using ChatGPT had the highest plagiarism results when tested with Paraphrasing Tool AI plagiarism detection website and Turnitin. Running the essays through paraphrasingtool.ai. has significantly reduced plagiarism. Two-way ANOVA with Tukey's multiple comparison test, *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. Data plotted as mean ± SEM.

The same trend was observed in the computing assignment (see Fig. 2). There is a statistically significant interaction between a text source and plagiarism detection tool ($F_{(2,\ 18)}$=9.82, P=0.0013) with significant effects on plagiarism arising from a text source ($F_{(1,\ 18)}$=14.79, P=0.0012) and plagiarism detection tool ($F_{(2,\ 18)}$=35.16, P<0.0001).

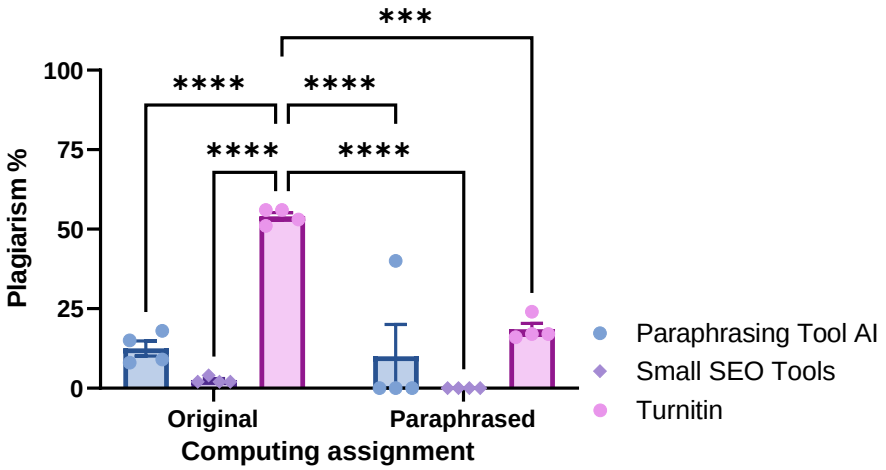

Fig. 2. The quality of AI generated computing assignments assessed in plagiarism checkers. Most of the assignments had a low level of plagiarism detected when using all tools, apart from Turnitin in the original ChatGPT version. Two-way ANOVA with Tukey's multiple comparison test, *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. Data plotted as mean ± SEM.

When comparing original ChatGPT generated text versus the paraphrased version in both assignment groups, paraphrasing the text resulted in acceptable plagiarism levels with similarity values <25% when tested using Turnitin – a plagiarism drop from 49.5% to 21% for biology and 54% to 18.5% drop for computing assignments. This plagiarism drop is evident when assessing the reports with highlighted sections that were plagiarised in biology (Fig. 3) and computing (Fig. 4) assignments.

Fig. 3. A snippet of plagiarism report from Turnitin for biology assignment highlighting all the copied sections in the original ChatGPT text (A) and the paraphrased version of the same assignment (B).

A

Introduction:

Image classification is a fundamental task in the field of machine learning, with a wide range of applications such as facial recognition, object detection, and medical image analysis. In this report, we will critically evaluate five machine learning algorithms for image classification: k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), Decision Trees, Random Forests, and Convolutional Neural Networks (CNN).

Evaluating Metrics:

There are several metrics that can be used to evaluate the performance of a machine learning algorithm for image classification. The most commonly used metric is accuracy, which is defined as the percentage of correct predictions made by the model. However, accuracy can be misleading if the data is imbalanced, meaning that there are significantly more samples of one class compared to the others. In such cases, it is important to also consider other metrics such as precision, recall, and F1 score.

Precision is defined as the number of true positive predictions divided by the total number of positive predictions made by the model. Recall is the number of true positive predictions divided by the total number of positive samples in the dataset. The F1 score is the harmonic mean of precision and recall, and it is a good metric to use when we want to balance precision and recall.

B

Machine learning's core task is image classification, which has numerous applications in areas like facial recognition, object detection, and medical image analysis. k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), Decision Trees, Random Forests, and Convolutional Neural Networks (CNN) are five machine learning algorithms for image classification that we will evaluate critically in this report.

Metric Evaluation:

A machine learning algorithm for image classification can be measured using a variety of metrics. Accuracy, which is the percentage of accurate predictions made by the model, is the most frequently used metric. But if the data are unbalanced—that is, if one class of samples is represented by a disproportionately large number of samples—accuracy can be deceptive. In such circumstances, it is crucial to also take other metrics into account, such as precision, recall, and F1 score.

The ratio of the number of accurate positive predictions to all the positive predictions the model made is known as precision. Recall is calculated by dividing the total number of positive samples in the dataset by the proportion of true positive predictions. When we want to balance precision and recall, the F1 score, which is the harmonic mean of the two, is a useful metric to use.

Fig. 4. A snippet of plagiarism report from Turnitin for computing assignment highlighting all the copied sections in the original ChatGPT text (A) and the paraphrased version of the same assignment (B).

3.2 AI generated text detection

Any written text can be assessed by measuring the content's perplexity and burstiness. Perplexity tests how well a language model can predict a sequence of words. The lower the value, the better the language model at predicting the next word. This perplexity score can then be used to find out burstiness – variation in the randomness in the text.

All of the biology assignments written by ChatGPT had low perplexity and burstiness scores, strongly indicating that the text is entirely written by AI (see Table 2). Only one of the assignments (paraphrased ChatGPT text with the regenerated response from original assignment question) was suggested to be only partially written by the AI.

It was surprising to see that all computing assignments written by ChatGPT (without paraphrasing the text) had much higher scores for perplexity and burstiness compared to biology assignments. It was also suggested that all assignments were only partially written by the AI. Paraphrasing the text resulted in even higher scores, with "paraphrased ChatGPT + paraphrased assignment + regenerate response" assignment tricking even the AI detection tool, as it was suggested that the work is likely to be written entirely by a human.

Table 2. Perplexity, burstiness and likely text source analysis of all assignments. Tested with GPTZero API.

Text source	Biology			Computing		
	Average Perplexity Score	Burstiness Score	Likely text source	Average Perplexity Score	Burstiness Score	Likely text source
ChatGPT + original assignment	19	12	AI	38	44	partially AI
ChatGPT + original assignment + regenerate response	17	19	AI	81	163	partially AI
ChatGPT + paraphrased assignment	17	17	AI	43	36	partially AI
ChatGPT + paraphrased assignment + regenerate response	18	17	AI	93	184	partially AI
paraphrased ChatGPT + original assignment	24	15	AI	116	255	partially AI
paraphrased ChatGPT + original assignment + regenerate response	30	18	partially AI	289	868	partially AI
paraphrased ChatGPT + paraphrased assignment	28	18	AI	188	517	partially AI
paraphrased ChatGPT + paraphrased assignment + regenerate response	23	12	AI	452	1174	human

4 Discussion

Artificial intelligence (AI) will likely continue to play an increasingly important role in academia in the future as we are seeing a rapid increase in the tools available in the market. Just in one week since the release of ChatGPT, one million people have tried it out [3]. However, it might negatively impact students' learning experience as AI might make cheating more accessible and harder to detect than ever before. Writing

essays or online exams with AI requires minimal effort and it results in high quality outputs.

As seen in the results, using AI powered tools can generate human like essays that are hard to detect as plagiarized. We used three plagiarism detection websites - paraphrasingtool.ai plagiarism checker, SmallSEOTools and Turnitin - to test the plagiarism of essays written by ChatGPT in the fields of biology and computer science. In both cases, the essays could easily pass plagiarism detection tools after they were paraphrased with an AI tool paraphrasingtool.ai and when testing with free to use SmallSEOTools, there was almost no plagiarism detected in all of the tested assignments.

Much to our surprise, running the assignments through AI generated text detection tool GPTZero API did not always detect that all the text is written entirely by the AI. With the improvements in the language models used in AI text generation (the launch of GPT-4 later in 2023), it will likely become harder and harder to detect works written entirely by chatbots.

As argued by [4], using ChatGPT can put an end to the way academics assess student's progress. It is predicted that more sophisticated tools will be developed, which will make it harder to detect plagiarism and AI generated text. Having an easy way to produce high quality essays, reports, and coding assignments will affect the integrity of education and the learning process.

For now, academics might need to reconsider how they assess student learning and increase the use of in-class and skills-based assessments. Changing the assignments by incorporating personal opinion part would also require more input from students minimising the use of AI chatbots, as the AI cannot generate answers for such questions.

Although AI tools might sometimes be seen as a danger very rapidly invading academia, correct use of them can bring great benefits. The use of AI can improve the quality of educational processes through increased quality of educational resources with reduced need for human power and resources, by helping students to learn some basic concepts of the topics they are studying, encouraging them to explore additional resources, or even assisting with career choices [5]. Seeing that AI is getting more and more integrated into our daily lives, our role as educators should be to teach students how to remain critical when using AI tools, as very often they can provide wrong information, how to check if the information provided by AI chatbot is right and where to look for reliable, scientific, peer reviewed information.

It is worth remembering that the role of education is not about memorizing information, but about building the professional skills needed to succeed in life as an individual. Time management, critical thinking, communication, research and interpersonal skills are just a few examples of skills that employers are looking for, and these skills require a range of activities and experiences for their development. As educators, we should use a wide range of tools that are out there to prepare students for their future success and teaching them the best practice of safe and critical AI tool use should be part of the curriculum.

5 References

1. Surahman, E., Wang, T.H.: Academic dishonesty and trustworthy assessment in online learning: A systematic literature review. J Comput Assist Learn. 38, 1535–1553 (2022). https://doi.org/10.1111/jcal.12708
2. Abd-Elaal, E.S., Gamage, S.H.P.W., Mills, J.E.: Assisting academics to identify computer generated writing. European Journal of Engineering Education. 47, 725–745 (2022). https://doi.org/10.1080/03043797.2022.2046709
3. Stokel-Walker, C.: AI bot ChatGPT writes smart essays — should professors worry? Nature. (2022). https://doi.org/10.1038/d41586-022-04397-7
4. Susnjak, T.: ChatGPT: The End of Online Exam Integrity? (2022)
5. Chen, Y., Jensen, S., Albert, L.J., Gupta, S., Lee, T.: Artificial Intelligence (AI) Student Assistants in the Classroom: Designing Chatbots to Support Student Success. Information Systems Frontiers. (2022). https://doi.org/10.1007/s10796-022-10291-4
6. *GPTZero*. Retrieved February 9, 2023, from https://gptzero.me/