Canadian Perspectives on Academic Integrity, 7(1), 1-9.
http://doi.org/10.55016/ojs/cpai.v7i1/77675
Peer-reviewed article

Who wrote this? Detecting artificial intelligence-generated text from human-written text

Rahul Kumar[1]*, & Michael Mindzak[1]

Abstract
This article explores the impact of artificial intelligence (AI) on written compositions in education. Participants' accuracy in distinguishing between texts generated by humans and those produced by generative AI (GenAI) was examined. The study challenges the assumption that the listed author of a paper is the one who wrote it, which has implications for formal educational systems. If GenAI text becomes indistinguishable from human-generated text to a human instructor, marker, or grader, it raises concerns about the authenticity of the submitted work. This is particularly relevant in postsecondary education, where academic papers are crucial in assessing students' learning, application, and reflection. The study had 135 participants who were randomly presented with two passages in one session. The passages were on the topic of "How will technology change education?" and were placed into one of three pools based on the source of origin: written by researchers, generated by AI, and searched and copied from the Internet. The study found that participants were able to identify human-generated texts with an accuracy rate of 63% and with an accuracy of only 24% when the composition was AI-generated. Despite its limitations, such as limited sample size and an older predecessor of the current GenAI software, this study highlights the potential impact of AI on education and the need for further research to evaluate comparisons between AI-generated and human-generated text.

Keywords
academic integrity, artificial intelligence, Canada, detection, GenAI, generative AI, KMR

[1] *Brock University*
***Corresponding author**: rkumar@brocku.ca

Introduction

Recent developments in and availability of artificial intelligence (AI) tools have sparked an unprecedented interest in the field. A question has emerged regarding whether such AI tools can replicate the written compositions of humans, particularly students, in education. This issue of detection is further complicated because generative AI (GenAI) is particularly good at generating previously unseen (i.e., new) text of high quality (Lund & Wang, 2023; Mindzak & Eaton, 2021). This pilot study aims to evaluate the robustness of current text-generating software compared to human-generated compositions by examining participants' accuracy in distinguishing human versus GenAI texts.

Our investigation is an essential undertaking because there is an implied assumption in formal educational systems that when someone submits a paper, that person wrote it. If that assumption is false because GenAI text is indistinguishable from human-generated text, then the listed author might not have produced the text. This is problematic because current postsecondary education (PSE) practices, especially in disciplines such as education, require students to write numerous academic papers to successfully complete program requirements—a practice that assumes students' writing demonstrates their learning, application, reflection, and creativity while integrating course material into their papers. Yet, there is a long-standing tendency for some students to take shortcuts and submit papers they purport to have written when in fact they were written by someone else or copied from elsewhere (Drake, 1941; Jordan, 2001). Many systems are sought and implemented in PSE to detect and subsequently thwart the undesirable practice of misrepresenting authorship (Dawson & Sutherland-Smith, 2018a). Popular strategies include academic integrity workshops, reminders of the importance of academic integrity, posters, honour pledges, and subscription-based text-matching detection software to incline students to uphold academic integrity. And on the detection side, training graders and markers (Dawson & Sutherland-Smith, 2018b) is advocated. Policies to guide action when breaches occur are also a staple method at almost all PSE institutions. The prevalence of AI tools in generating new text will continue to complicate detection efforts if the quality of writing produced by GenAI tools and large language models (LLMs) is comparable to human-generated texts.

Literature Review

Plagiarism is a persistent phenomenon at all levels of formal schooling (Drake, 1941; Vrbanec & Meštrović, 2021),

which continues to undermine academic integrity. PSE has adopted various commercial text-matching software programs that detect plagiarized work since the early 2000s (Brinkman, 2013). However, the increasing sophistication of artificially intelligent writing software poses significant challenges if it can evade detection by conventional techniques (Perkins et al., 2023). This is precisely what is happening with the advancement and integration of AI tools in education (Kerr, 2020; Selwyn, 2019), propelled by the fast-paced development of AI technology (Gray, 2022). As these GenAI tools continue to improve with each iteration, they will likely perform many facets of educational work (Kumar, 2023; McMurtrie, 2023a; Williamson, 2020; Williamson & Enyon, 2020), increasing students' ability to take shortcuts and impinging on academic integrity.

The perspicacity of AI use has varied from doom and gloom predictions to pronouncements of changing teaching and learning practices (Feldstein, 2022; McMurtrie, 2023b; Warner, 2023). Consequently, technological developments have rapidly reshaped our traditional understanding of plagiarism, academic integrity, and authorship (Mindzak & Eaton, 2021). For example, varying degrees of school closures during the COVID-19 pandemic have further accelerated and exacerbated the expansion of technological integration into educational spheres (Teräs et al., 2020). In short, the proliferation of educational technology like GenAI has increased because it can produce human-like quality text in seconds via natural language generation. Urgent questions have arisen about the nature and purpose of teaching writing and requiring it in courses and programs when it only marginally meets learning outcomes (McKnight, 2021). GenAI is reaching such advanced levels that humans do not need to provide as much input in the writing process.

There is a marked increase in the number of products that use machine learning (Kerr, 2020) and equally significant growth in software and services for essays, research papers, and article generation (Abd-Elaal et al., 2019). With powerhouse companies like Google and Microsoft striving to produce and refine human-like text, algorithmic and non-algorithmic writing software, tools, and platforms continue to gain broad reach and popularity (Abd-Elaal et al., 2019; Marche, 2021). Most notable in this field has been the development of tools released by OpenAI, particularly ChatGPT[1]. The sophistication of this powerful AI based on LLM is becoming increasingly difficult to differentiate from human writing (Dehouche, 2021; Luitse & Denkena, 2021). Currently, a version of ChatGPT is available to anyone with an internet connection and a compatible web browser.

With the release of ChatGPT, there has been polemic debate surrounding the use of GenAI in education—including its ability to take on writing tasks previously limited to humans (Haman & Školník, 2023). With rapid successive releases of GenAI tools and significant improvements in quality, the empirical research in the field concerning both the efficacy as well as detectability of GenAI has not been able to keep pace. Yet the detection of unauthorized work continues to occupy university resources to prevent misconduct (Dawson & Sutherland-Smith, 2018a; Eaton, 2021). The availability of GenAI has raised the stakes of how universities and K-12 schools will gauge what constitutes academic misconduct (Mindzak & Eaton, 2021). This study provides empirical evidence regarding participants' ability to determine whether a text composition was GenAI or human-generated.

[1]While this article references ChatGPT, which has expanded significantly in 2023, the study itself utilized a previous version called GPT-2.

Methodology

To investigate the capability of an AI text generator compared to human writing composition, we created a survey instrument to collect data from volunteers affiliated with the education faculty at an Ontario university. Ethics clearance was obtained prior to the commencement of the research.

The Survey Tool

We created a survey tool, named KMR, with which we collected the following demographic data from participants:

1. gender
2. age range
3. ethnicity
4. highest level of education degree/certification
5. faculty association
6. role

Participants were given the option to not provide the data above, but all questions corresponding to the data collection were compulsory; participants could not proceed if they omitted any of the demographic questions. After collecting demographics, participants were randomly presented with compositions of (at most) 260 words on the topic of "How will technology change education?"

One of the study's objectives was to determine the accuracy of predicting the composition source (i.e., whether a human being or AI composed it). We also wanted to investigate which of the demographics correlated with the accuracy of the prediction. Before the study, we could not predict which demographics would be best correlated; hence, all the demographic items shown above were solicited from participants.

After reading the composition, participants were presented with the following instructions:

1. Rate the writing level of the composition.
2. Assign a mark to the composition (between 0 and 100).
3. Determine whether the composition is human-generated or computer-generated.
4. Determine whether the text composition was obtained from the Internet or not.
5. Provide any qualitative comments.

After evaluating the first composition, participants were presented with a second composition with the same instructions. The survey completion time ranged between 9 and 12 minutes per participant.

Compositions were randomly presented to participants that had been categorized into one of the following pools based on the source of origin: (a) written by researchers, (b) generated by AI, and (c) searched and copied from the internet. Each pool contained two passages. Although passages were randomly presented to participants, they were controlled so that no singular passage appeared inordinately more than others; that is, the passage presentation was distributed equally.

Research team members wrote the passages that were part of the first pool. One of the team members ensured there were no grammatical errors. The second pool consisted of AI-generated passages using OpenAI's GPT-2. We used a commercial service through InferKit, developed by Adam Daniel King[2]. The text was generated using the settings shown in Table 1.

Table 1. Settings Used for Generating AI Passages

Criteria	Setting
Prompt	How will technology change education?
Attempts	3
Platform	InferKit
Length to generate	1,500 characters
Words to include	Education, technology, research, students
Category	Any
Start at the beginning	Yes

No text smoothing was performed on the generated text. The third pool consisted of found text on the Internet using a Google search with the prompt: How will technology change education? We copied two paragraphs from the Google search.

Recruitment and Participants

Participants were recruited via an email sent to Faculty of Education instructors, staff, and students. The email contained a link to the survey. Participation in the study was voluntary, and participants could withdraw at any time. Participants were not compensated monetarily or otherwise for their participation. Only the completed survey results were included in the analysis.

Analysis and Findings

The data were collected between February and June 2022, and 135 responses were recorded, which included 17 partial responses. Thirty-six (26.7%) of the respondents were self-identified males and 98 (72.6%) respondents were self-identified females. The age ranges into which respondents fell are shown in Table 2.

[2]https://inferkit.com/

Table 2. Age Ranges of Participants

Age range	Number	Percent
18–24	74	54.8
25–34	25	18.5
35–44	13	9.6
45–54	17	12.6
55–64	4	3.0
65–74	2	1.5
Total	135	100.0

Participants' education levels varied, ranging from high school certificates to PhDs. The distribution is shown in Figure 1. The roles that participants most identified with are illustrated in Figure 2.

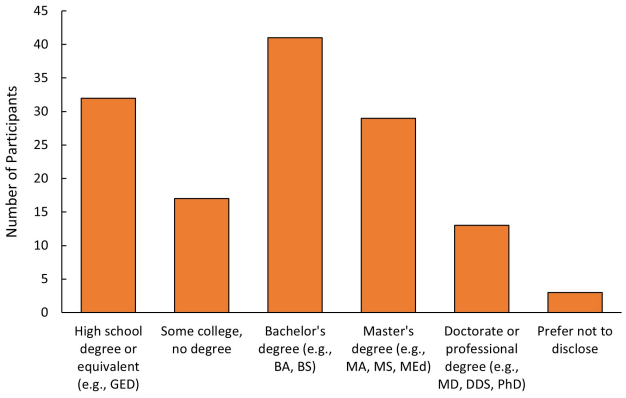

Figure 1. Distribution of the Highest Degrees of Participants

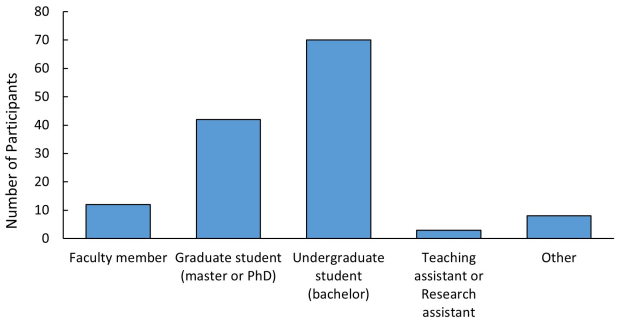

Figure 2. Distribution of the Primary Role of Participants

The survey design was such that each participant was presented with only two passages. Hence, the sample sizes for each passage available for analysis were smaller than the total number of participants in the study (n = 135). The low numbers in each of these categories resulted in the inability to run many inferential statistical tests to determine whether any demographics interacted or correlated with the determination of the correct authorship or origin of the composition. Nevertheless, base models revealed interesting observations that are presented in Table 4.

Table 3. Estimates of the Passage-Specific Intercept-Only Models

Passage	Correct source	Percent correct	Wald	B	e^B	p
1	Human	66.7	4.164	0.693	2.000	0.041
2	Human	64.9	3.170	0.613	1.846	0.075
3	Human	55.0	0.399	0.201	1.222	0.528
4	AI	20.0	10.762	-1.386	0.250	0.001
5	AI	27.0	7.199	-0.993	0.370	0.007
6	Human	66.7	4.164	0.693	2.000	0.041

Note. Percent correct = proportion of guesses that correctly identified the authorship of a given passage out of the total attempts to identify that message.; e^B = odds of correct guesses about the authorship of each passage, calculated by dividing the percentage of correct guesses by the percentage of incorrect guesses.

Intercept-only logistic regression was performed for each type of passage to examine if the proportion of correct guesses to the authorship of the passages was significantly different from the proportion of incorrect guesses. As per the odds values in Table 3, for human-generated passages, participants were 1.2 to 2 times as likely (i.e., 20% to 100% *more* likely) to provide correct responses than incorrect ones. This increased likelihood of correct responses (vs. incorrect responses) was statistically significant for two out of four human-generated messages. In contrast, for AI-generated passages, participants were 0.25 to 0.37 times as likely (i.e., 63% to 75% *less* likely) to provide correct over incorrect guesses of the authorship. This compromised likelihood of correct responses (vs. incorrect responses) was statistically significant for the two AI-generated messages. This observation demonstrates that participants are better at identifying the authorship of a human-generated passage than an AI-generated passage.

Another way to report findings is to determine "true positives" and "true negatives" for text passages that were AI-generated and human-written. In turn, these would permit calculating the true positive and the true negative rates. This is shown in Table 4.

Table 4. Cross-Tabulation of Participants' Guesses on Authorship and True Authorship

		Authorship of Passages		
		AI	Human	Total
Guesses about	AI	17	57	74
passage	Human	55	99	154
authorship	Total	72	156	228

Note. Authorship of passages was determined by adding the correct entries of AI and human guesses. The incorrect responses included those identified incorrectly and the participants were unsure about.

Using the numbers in Table 4, formulas (1) to (4) were used to calculate the numbers shown in Table 5. The calculation of true positive rates (TPR), true negative rates (TNR), positive predictive value (PPV), and negative predictive value (NPV) annuls the variance associated with only a few participants having assessed each passage.

$$TPR = \frac{Truepositive(TP)}{(Truepositive(TP) + Falsenegative(FN))} \quad (1)$$

$$TNR = \frac{Truenegative(TN)}{(Truenegative(TN) + Falsepositive(FP))} \quad (2)$$

$$PPV = \frac{Truepositive(TP)}{(Truepositive(TP) + Falsepositive(FN))} \quad (3)$$

$$NPV = \frac{Truenegative(TN)}{(Truenegative(TN) + Falsenegative(FN))} \quad (4)$$

Since there were participants in each category (see Table 4) showing 17 passages related to AI and 57 by humans, only descriptive comparisons are possible. Applying the above formulas to AI-generated text passages yields the values reported in Table 5. Similarly, the data can be presented across all passages, as shown in Table 6.

Table 5. Descriptive Comparison of Accuracy Rates of AI Passages

Category	Values
TPR for AI detection or Sensitivity	0.24
TNR for AI detection or Specificity	0.63
PPV for AI detection	0.23
NPV for AI detection	0.64

Note. TPR = true positive rate; TNR = true negative rate; PPV = positive predictive value; NPV = negative predictive value

To further evaluate participants' performance in differentiating between AI- and human-generated passages, we calculated a series of diagnostic metrics quantifying different rates for correct and incorrect responses (see Tables 4 and 6). Overall, participants were worse at identifying AI-generated than human-generated messages. For example, out of the total number of times that AI-generated messages were presented, participants made correct guesses only on 24% of them (see Table 6, Row 1). In contrast, out of the total number of times that human-generated messages were presented, participants correctly identified 63% of them (see Table 6, Row 2). Furthermore, out of the total times participants guessed that a message was AI-generated, only 23% of these guesses were correct (see Table 6, Row 3). Out of the total times participants guessed that a message was human generated, 64% of the guesses were correct (see Table 6, Row 4).

In short, participants were less likely to label a message as AI-generated than human-generated. Out of the total attempts to guess the authorship of the passages, participants believed that a passage was AI-generated 32% of the time, and that a passage was human-generated 68% of the time (see Table 6, Rows 5-6). Therefore, it seems that participants assume human agency when determining the authorship of passages.

Table 6. Diagnostics Metrics Quantifying Correct and Incorrect Guesses About the Authorship of Passages

Formulas for diagnostic metrics	Value
True guesses (AI)/total messages (AI)	0.24
True guesses (human)/total messages (human)	0.63
True guesses (AI)/total guesses (AI)	0.23
True guesses (human)/total guesses (human)	0.64
Total guesses (AI)/total guesses	0.32
Total guesses (human)/total guesses	0.68

Note. These values were derived from the diagnostic metrics that are commonly used in data science (e.g., sensitivity, specificity, positive predictive rate, and negative predictive rate). See Table 4 for the specific values used for calculating these matrices.

Discussion

Overall, results from the study indicate that when the passage was composed by humans, participants correctly identified it two-thirds of the time (66.7%)–Passages 1 and 6. But when the passage presented was AI-generated (Passages 4 and 5), participants had difficulty identifying it as such. They either identified it as human-generated or marked it as unsure. In this study, the unsure results were marked as incorrect because in practical terms when a grader (faculty or teaching assistant) is unsure, they are more likely to err on the side of not doing anything rather than raising the flag as suspicious. In effect, the passage would pass as human-written. If the detection rates of AI-composed text were so low (as compared with human-written passages), it raises important sets of questions about better understanding the impact of GenAI in education, which extends to students, educators, and institutions because AI permeates PSE. Chief among them are: How can we reliably detect an AI-written passage? How do we design assignments and assessments where the boundaries between human and AI writing are moot? That is, when, if ever, are we going to get to the postplagiarism world that Eaton (2021; 2023) envisions? Until we reach those tranquil waters, we will have to confront the issue of academic integrity in imaginative ways. Just as in the past with contract cheating, for instance, GenAI blurs, if not crosses, the line between what is original, individual work and what is conceived as assistance, if not overt plagiarism.

Gray (2022) stated that the contract cheating industry encompasses "a devaluing and de-prioritization of original work when it comes to the classroom space, predatory relationships between consumers and service providers, and a cavalier approach to handling data privacy" (p. 190). The relationships between consumers and service providers lack accountability, ethics, and academic integrity. The antithesis or absence of academic integrity and contract cheating have been granted unprecedented avenues to thrive. If the results from this study hold up against a larger study, the implications would be grim–the detection of who did the work would remain an ongoing concern. Automated detection products currently available do a poor job as they produce false positives and false negatives when human compositions are introduced (Köbis & Mossink, 2021). New technologies for automated detection are being explored, but they are in their infancy (Lancaster, 2023).

While the momentum and advocacy surrounding the fight against contract cheating is growing, the contract-cheating industry is experiencing stiff competition from GenAI (McColl, 2023), partly because the quality of GenAI-produced writing is decent enough to dupe the assessors and is done privately. The findings from this study support McColl's position. The AI-composed passages were sufficiently human-like, which is why humans' detection rate was so low (24%). Suppose the results from this study are corroborated in larger studies; in that case, human detective powers, professed by many, will need to be questioned. As the quality of GenAI improves, the automated detection will likely falter, as is evident in a recent investigation by Elkhatat et al. 2023.

This conundrum squarely poses the problem of verifying whether a text is human-generated or AI-generated. The solutions, like the problem itself, are emerging, as is evident in Lancaster's 2023 work. Eaton and Christensen Hughes 2022 claim there is a demonstrated need for more urgent measures to be taken regarding academic integrity, and the findings from this study of human assessors' inability to identify the author of texts only bolsters that call. No silver-bullet solution can resolve the predicament of accurately assessing authorship. This does not mean banning the use of AI, nor does it mean letting it run amok. A reasonable middle ground will emerge in time and with collaboration amongst those who are dealing with the issues. Eaton and Christensen Hughes 2022 propose cultivating a strong and sustainable community of Canadian researchers and practitioners to address academic misconduct and cheating: developing resources to be shared within the academic integrity community is necessary. In keeping with her own recommendation, Eaton hosts a weekly academic integrity hour for interested parties from the PSE sector in Canada to share approaches on practice, policy, training, and emergent issues.

Identifying work from GenAI creates additional challenges and burdens for researchers, teachers, and anti-plagiarism tools (Abd-Elaal et al., 2022). This will likely continue as GenAI tools become more sophisticated. While some of the characteristics and prescriptions to combat contract cheating apply to GenAI, there are ample differences. Therefore, conflating contract cheating- and GenAI-produced results should be avoided. The challenges in using GenAI tools productively and ethically while preserving academic integrity are numerous (Harper et al., 2021) and evolving, with faculty members and institutions caught up in an unending game of catch-up. Results from this study pose questions for educators and instructors regarding the accuracy of deciphering and gauging GenAI text without tools. Of course, the use of GenAI detection tools is also uncharted territory (Dalalah & Dalalah, 2023), and research on detecting GenAI is still evolving (Elkhatat et al., 2023; Köbis & Mossink, 2021; Lancaster, 2023; Weber-Wulff et al., 2023).

A notable theme in the emergent literature (Abd-Elaal et al., 2022; Gehrmann et al., 2019) is the call for increased awareness and the development of educational practices that could help educators identify AI-generated text to minimize the persistent threat of academic misconduct. Concerns have grown and will continue to grow as GenAI tools become more sophisticated and their proliferation among students increases. Dawson and Sutherland-Smith 2018a and Curtis et al. 2021 are concerned that because many educators are unaware of various AI writers, there is a pressing need to raise awareness and that training may assist in more accurately detecting this form of plagiarism and cheating. In accordance with their assertion and the results of this study, we need robust ongoing AI literacy for educators, administrators, and students to preserve academic integrity. Concomitantly, this study should be independently verified on a larger scale.

Limitations

This pilot study had several limitations, with the most notable being the sample size. With only 135 participants and each passage evaluated by fewer than 40 participants, the results did not permit logical regression against demographic variables. Nevertheless, with limited empirical work available on the topic, further research can use a similar survey instrument to evaluate comparisons between AI-generated and human-generated text.

This study had only six text passages: Two were AI-generated and four were human-generated. The research team members wrote two of the four human-generated passages, and two were obtained from the internet. The participants were randomly presented with two passages in one session, so a participant could have encountered two human-written passages. We cannot confirm if this approach may have skewed the results, and we thus regard it as a limitation of this study. Furthermore, since the ratio of human- to AI-written passages was two to one (2:1), the study relies on descriptive analysis comparison of accuracy ratio and hence the measures of sensitivity and specificity are reported.

While many demographic data were collected, we did not determine if English was participants' native language, which might be important because fluency could be an essential factor in deciphering authorship. Regardless, the next version of the tool should include native languages as an important criterion.

Lastly, this study utilized GPT-2, an earlier version of what is now known as ChatGPT or GPT-4, but that was the current technology at the time of this study. Advancements with this technology have resulted in significant observed improvements, such as successfully passing standardized exams and submitting graduate-level written compositions (van Boom, 2023). Hence, it is fair to assume that ongoing developments in GenAI will continue to result in improved text compositions compared to those tested against human compositions. Works of Weber-Wulff et al. 2023 and Elkhatat et al. 2023 also suggest that with the improvements in the GenAI technology, automated detection will also continue to falter.

Conclusion

Zawacki-Richer et al. 2019 assert that AI has been used for decades in the computing and industrial sectors, but its use is bound to increase since the launch of ChatGPT. ChatGPT has captured educators' and learners' imaginations and has proven to do a decent job of mimicking human writing. If automated detection tools like GPTZero, Originality.AI, and Turnitin are prone to be erroneous (Weber-Wulff et al., 2023) and raise concerns about privacy, student consent, and data mining, then human detection seems the only viable alternative. If the results from this study were to hold up in similar empirical research studies, then human detection cannot be relied upon either. Granted that work on automated detection is underway (Lancaster, 2023), the current predicament requires that PSE institutions re-examine the kinds of assessments assigned. The educational landscape is liable to change whether deliberate ameliorative actions are taken or not. If educators pay no heed, then students could be awarded false credentials without corresponding skills; that would be disastrous. Or suppose educators take deliberative action to refine assessments such that GenAI output is only an intermediary step in the final artifact a student submits; in that case, students can still demonstrate their true learning and we can bypass the detection problem. In either of the two scenarios, the future of PSE is on the precipice of monumental change.

In accordance with Kerr's 2020 prediction, it is safe to assume that AI integration into teaching, learning, and assessment practices will forge ahead. More research and interventions are needed to build the capacities of learners and teachers alike so that these tools are used ethically, efficiently, and effectively. GenAI can educate, train, and improve human-level tasks, like summarizing (Yang et al., 2023), but simultaneously, it poses challenges and risks (Southgate, 2021). This study reveals one such challenge—-human detection work that is not yet honed. It is conceivable that the participants in this study were unable to detect AI-generated passages as effectively as human-generated prose because they were unable to imagine that GenAI could write decent prose. However, human capacity has also improved because of the proliferation of ChatGPT and other LLMs. This assumption can only be validated by a larger, similar study.

Acknowledgments

The authors would like to thank the internal faculty funding provided to conduct the research. We would also like to thank the attendees at the Digital Pedagogies conference who asked interesting questions about implications for assessment at the session and help develop this manuscript to its current form. We would also like to thank Dr. Phillip Dawson from Deakin University for suggesting we conduct the statistical diagnostic tests to verify the results and present the aggregates. And

finally, we would also like to thank Mr. Xiaoyang Xia for providing valuable consultation with statistical analysis.

Author Information

Dr. Rahul Kumar is an Assistant Professor in the Department of Educational Studies at Brock University. His scholarship is primarily in the field of Higher Education. He also has written in the fields of educational technologies, artificial intelligence, quality of education at PSE level, and international education.

Dr. Michael Mindzak is an Assistant Professor in the Department of Educational Studies at Brock University. His scholarship is primarily in the field of labour relations. He has also written in the fields of educational technologies, artificial intelligence, and the nature of work of the professoriate.

Received
June 24, 2023

Accepted
December 10, 2023

Published online
January 10, 2024

References

Abd-Elaal, E., Gamage, S. H. P. W., & Mills, J. E. (2019). Artificial intelligence is a tool for cheating academic integrity [paper presentation. In *Aaee 2019 annual conference.* Queensland, Australia. https://aaee.net.au/wp-content/uploads/2020/07/AAEE2019_Annual_Conference_paper_180.pdf

Abd-Elaal, E., Gamage, S. H. P. W., & Mills, J. E. (2022). Assisting academics to identify computer generated writing. *European Journal of Engineering Education*, *47*(5), 725–745. https://doi.org/10.1080/03043797.2022.2046709

Brinkman, B. (2013). An analysis of student privacy rights in the use of plagiarism detection systems. *Science and Engineering Ethics*, *19*, 1255–1266. https://doi.org/10.1007/s11948-012-9370-y

Curtis, G. J., Slade, C., Bretag, T., & McNeill, M. (2021). Developing and evaluating nationwide expert-delivered academic integrity workshops for the higher education sector in australia. *Higher Education Research and Development*, *41*(3), 665–680. https://doi.org/10.1080/07294360.2021.1872057

Dalalah, D., & Dalalah, O. M. A. (2023). The false positives and false negatives of generative ai detection tools in education and academic research: The case of chatgpt. *The International Journal of Management Education*, *21*(2). https://doi.org/10.1016/j.ijme.2023.100822

Dawson, P., & Sutherland-Smith, W. (2018a). Can markers detect contract cheating? results from a pilot study. *Assessment and Evaluation in Higher Education*, *43*(2), 286–293. https://doi.org/10.1080/02602938.2017.1336746

Dawson, P., & Sutherland-Smith, W. (2018b). Can training improve marker accuracy at detecting contract cheating? a multi-disciplinary pre-post study. *Assessment and Evaluation in Higher Education*, *44*(5), 715–725. https://doi.org/10.1080/02602938.2018.1531109

Dehouche, N. (2021). Plagiarism in the age of massive generative pre-trained transformers (gpt-3. *Ethics in Science and Environmental Politics*, *21*, 17–23. http://doi.org/10.3354/esep00195

Drake, C. A. (1941). Why students cheat. *The Journal of Higher Education*, *12*(8), 418–420. https://doi.org/10.1080/00221546.1941.11773211

Eaton, S. E. (2021). *Plagiarism in higher education: Tackling tough topics in academic integrity*. Libraries Unlimited.

Eaton, S. E. (2023). 6 tenets of postplagiarism: Writing in the age of artificial intelligence. *Learning, Teaching and Leadership*. https://drsaraheaton.wordpress.com/2023/02/25/6-tenets-of-postplagiarism-writing-in-the-age-of-artificial-intelligence/

Eaton, S. E., & Christensen-Hughes, J. (Eds.). (2022). *Academic integrity in canada: An enduring and essential challenge*. Springer. https://doi.org/10.1007/978-3-030-83255-1

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of ai content detection tools in differentiating between human and ai-generated text. *International Journal for Educational Integrity*, *19*. https://doi.org/10.1007/s40979-023-00140-5 (Article 17.)

Feldstein, M. (2022). *I would have cheated in college using chatgpt. eliterate.* https://eliterate.us/i-would-have-cheated-in-college-using-chatgpt/

Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). *Gltr: Statistical detection and visualization of generated text.* https://doi.org/10.48550/arXiv.1906.04043 (arXiv.)

Gray, B. C. (2022). Ethics, edtech, and the rise of contract cheating. In J. Eaton S.E. & Hughes (Ed.), *Academic integrity in canada: An enduring and essential challenge* (p. 189–201). Springer. https://doi.org/10.1007/978-3-030-83255-1_9

Haman, M., & Školník, M. (2023). Using chatgpt to conduct a literature review. *Accountability in Research*. https://doi.org/10.1080/08989621.2023.2185514

Harper, R., Bretag, T., & Rundle, K. (2021). Detecting contract cheating: Examining the role of assessment type. *Higher Education Research and Development*,

40(2), 263–278. https://doi.org/10.1080/07294360.2020.1724899

Jordan, A. E. (2001). College student cheating: The role of motivation, perceived norms, attitudes, and knowledge of institutional policy. *Ethics and Behavior*, *11*(3), 233–247. https://www.doi.org/10.1207/S15327019EB1103_3

Kerr, K. (2020). Ethical considerations when using artificial intelligence-based assistive technologies in education. In B. Brown, V. Roberts, M. Jacobsen, & C. Hurrell (Eds.), *Ethical use of technology in digital learning environments: Graduate student perspectives* (p. 9–14). University of Calgary. https://doi.org/10.7939/r3-yanz-fe63

Kumar, R. (2023). Faculty members' use of artificial intelligence to grade student papers: A case of implications. *International Journal for Educational Integrity*, *19, Article 9*. https://doi.org/10.1007/s40979-023-00130-7

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus maya angelou: Experimental evidence that people cannot differentiate ai-generated from human-written poetry. *Computers in Human Behavior*, *114, Article 106553*. https://doi.org/10.1016/j.chb.2020.106553

Lancaster, T. (2023). Artificial intelligence, text generation tools and chatgpt—does digital watermarking offer a solution? *International Journal for Educational Integrity*, *19, Article 10*. https://doi.org/10.1007/s40979-023-00131-6

Luitse, D., & Denkena, W. (2021). The great transformer: Examining the role of large language models in the political economy of ai. *Big Data & Society*, *8*(2). https://doi.org/10.1177/20539517211047734

Lund, B. D., & Wang, T. (2023). Chatting about chatgpt: How may ai and gpt impact academia and libraries? *Library Hi Tech News*, *40*(3), 26–29. https://doi.org/10.1108/LHTN-01-2023-0009

Marche, S. (2021). The chatbot problem. *The New Yorker*. https://www.newyorker.com/culture/cultural-comment/the-chatbot-problem

McColl, B. (2023). *Chegg shares plunge after company warns that chatgpt is impacting growth*. Investopedia. https://www.investopedia.com/chegg-shares-plunge-after-company-warns-that-chatgpt-is-impacting-growth-7487968

McKnight, L. (2021). Electric sheep? humans, robots, artificial intelligence, and the future of writing. *Studies in Culture and Education*, *28*(4), 442–455. https://doi.org/10.1080/1358684X.2021.1941768

McMurtrie, B. (2023a). *Are professors ready for ai?* The Chronicle of Higher Education. https://www.chronicle.com/newsletter/teaching/2023-05-25

McMurtrie, B. (2023b). *Will chatgpt change the way you teach?* The Chronicle of Higher Education. https://www.chronicle.com/newsletter/teaching/2023-01-05

Mindzak, M., & Eaton, S. E. (2021). *Artificial intelligence is getting better at writing, and universities should worry about plagiarism*. The Conversation. https://theconversation.com/artificial-intelligence-is-getting-better-at-writing-and-universities-should-worry-about-plagiarism-160481

Perkins, M., Roe, J., Postma, D., McGaughran, J., Hickerson, D., & Cook, J. (2023). Game of tones: Faculty detection of gpt-4 generated content in university assessments. *arXiv*. https://www.doi.org/10.48550/arXiv.2305.18081

Selwyn, N. (2019). *Should robots replace teachers? ai and the future of education*. Wiley.

Southgate, E. (2021). Artificial intelligence and machine learning: A practical and ethical guide for teachers. In C. Wyatt-Smith, B. Lingard, & E. Heck (Eds.), *Digital disruption in teaching and testing: Assessments, big data, and the transformation of schooling* (2nd ed., p. 60–74). Routledge. https://www.doi.org/10.4324/9781003045793-3

Teräs, M., Suoranta, J., Teräs, H., & Curcher, M. (2020). Post-covid-19 education and education technology "solutionism": A seller's market. *Postdigital Science and Education*, *2*, 863–878.

van Boom, D. (2023). Chatgpt can pass the bar exam. *Does that actually matter? CNet.*. https://www.cnet.com/tech/chatgpt-can-pass-the-bar-exam-does-that-actually-matter/

Vrbanec, T., & Meštrović, A. (2021). Taxonomy of academic plagiarism methods. *Zbornik Veleučilišta u Rijeci*, *9*(1), 283–300. http://doi.org/10.31784/zvr.9.1.17

Warner, J. (2023). *How about we put learning at the center?* Inside Higher Ed. https://www.insidehighered.com/blogs/just-visiting/how-about-we-put-learning-center

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., ... Waddington, L. (2023). Testing of detection tools for ai-generated text. *International Journal for Educational Integrity*, *19*(26). https://doi.org/10.1007/s40979-023-00146-z

Williamson, B. (2020). The social life of artificial intelligence in education. *Code Acts in Education*. https://codeactsineducation.wordpress.com/2020/08/18/social-life-artificial-intelligence-education/

Williamson, B., & Enyon, R. (2020). Historical trends, missing links, and future directions in ai in education. *Learning, Media and Technology*, *45*(3), 223–235. https://doi.org/10.1080/17439884.2020.1798995

Yang, X., Li, Y., Zhang, X., Chen, H., & Cheng, W. (2023). *Exploring the limits of chatgpt for query or aspect-based text summarization.* https://doi.org/10.48550/arXiv.2302.08081 (arXiv.)

Zawacki-Richer, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? *International Journal of Educational Technology in Higher Education*, *16, Article 39*. http://doi.org/10.1186/s41239-019-0171-0