Journal of Academic Ethics (2024) 22:89–113
https://doi.org/10.1007/s10805-023-09492-6

Detection of GPT-4 Generated Text in Higher Education: Combining Academic Judgement and Software to Identify Generative AI Tool Misuse

Mike Perkins[1] · **Jasper Roe**[2] · **Darius Postma**[3] · **James McGaughran**[1] · **Don Hickerson**[1]

Accepted: 24 October 2023 / Published online: 31 October 2023

Abstract
This study explores the capability of academic staff assisted by the Turnitin Artificial Intelligence (AI) detection tool to identify the use of AI-generated content in university assessments. 22 different experimental submissions were produced using Open AI's ChatGPT tool, with prompting techniques used to reduce the likelihood of AI detectors identifying AI-generated content. These submissions were marked by 15 academic staff members alongside genuine student submissions. Although the AI detection tool identified 91% of the experimental submissions as containing AI-generated content, only 54.8% of the content was identified as AI-generated, underscoring the challenges of detecting AI content when advanced prompting techniques are used. When academic staff members marked the experimental submissions, only 54.5% were reported to the academic misconduct process, emphasising the need for greater awareness of how the results of AI detectors may be interpreted. Similar performance in grades was obtained between student submissions and AI-generated content (AI mean grade: 52.3, Student mean grade: 54.4), showing the capabilities of AI tools in producing human-like responses in real-life assessment situations. Recommendations include adjusting the overall strategies for assessing university students in light of the availability of new Generative AI tools. This may include reducing the overall reliance on assessments where AI tools may be used to mimic human writing, or by using AI-inclusive assessments. Comprehensive training must be provided for both academic staff and students so that academic integrity may be preserved.

Keywords ChatGPT · GPT-4 · Turnitin AI detect · AI detection · Assessment design · Artificial intelligence

Extended author information available on the last page of the article

Introduction

Background

Artificial Intelligence (AI) has seen an upsurge in its application in various domains and offers a wealth of opportunities for society, for example, helping medical professionals diagnose diseases more quickly and enhancing communication across borders (Pichai, 2023). However, AI's ability to seamlessly enable the production and reorganisation of information carries risks. Among academics, existential conversations about the future of higher education and assessment are ongoing, as teachers and students begin to acquaint themselves with the emerging challenges of the post-AI education landscape (Perkins, 2023). Such risks are set to continue to grow, as it is estimated that AI applications are doubling in complexity and processing power every six months (Pichai, 2023).

One example of AI development creating risk in the academy can be seen in the recent public deployment of Large Language Models (LLM), such as GPT-3 and GPT-4, which have led to a 'wave' of attention not only from the public but also among policymakers and scholars (Bowman, 2023). LLM applications such as ChatGPT, which is based on OpenAI's GPT models, are examples of Generative-AI (GenAI) tools which exhibit sophisticated language generation capabilities. From an educational perspective, there are multiple benefits to the use of LLMs, including assisting Non-Native English Speakers (NNES) with structured academic writing (Roe et al., 2023), supporting personalised learning (Kasneci et al., 2023), and providing formative feedback for students (Baidoo-Anu & Owusu Ansah, 2023). However, the sophistication of these tools in producing text that is fluent, detailed, and highly natural in tone means that distinguishing between human-written and AI-generated content is becoming increasingly difficult for faculty members (Abd-Elaal et al., 2022; Köbis & Mossink, 2021). This leaves the opportunity for students to use such tools to submit assessed work that has been partially or fully authored by AI.

In response, several software tool packages have been developed to help with the identification of AI text, which may support faculty in identifying when AI-generated text has been submitted by a student. However, few studies have explored the underlying accuracy of these tools, and no empirical studies have compared how they might be used in practice alongside academic judgement. Such studies are important for establishing an understanding of the capabilities of both human assessors and detection technologies to accurately identify any misuse of GenAI tools. Understanding this can help HEIs craft well-informed institutional responses to Gen-AI use in HE. This is particularly important given the controversy regarding how these tools may be utilised in an academic setting, especially as the potential impact of false accusations of academic misconduct may be severely damaging for students. For example, the release of the AI detection tool by Turnitin in April 2023 was met with concern by some universities that chose not to enable the function (Cassidy, 2023). Recent studies have identified the challenges that these software tools face in accurately identifying AI-generated content (Elkhatat et al., 2023; Weber-Wulff et al., 2023), especially as more advanced LLMs are used to develop the text (Elkhatat et al., 2021) or when techniques are used to attempt to evade detection (Weber-Wulff et al., 2023).

The unauthorised or undeclared use of GenAI tools in student submission is a risk to academic integrity, as this may constitute misrepresentation of authorship and be considered academic misconduct. The consequences of unchecked academic integrity violations grow-

ing may be severe; Bretag et al. (2019) invites us to consider the implications of students who cheat their way to becoming doctors, engineers, and social workers by submitting work that is not their own, while Guerrero-Dib et al. (2020) highlight that intentional violations of academic integrity behaviours can continue into the workplace after students complete their studies. Thus, the risk must be treated as serious, and research is required to understand the accuracy, effectiveness, and potential uses of AI text detection software as a major intervention or adjuvant for the detection of dishonest use of GenAI tools in academic work.

To this end, the European Network of Academic Integrity (ENAI) has released guidelines on the acceptability and use of AI tools in education, highlighting that if AI tools are used to produce assessed work which results in the awarding of credit or progression, this could constitute academic misconduct. However, recognising that AI tools will likely form an important part of students' professional lives in the future, it is necessary for them to make use of these in an educational context and gain familiarity with the potential opportunities for AI (Foltynek et al., 2023). As a result, putting guidelines into place, implementing codes of conduct, developing sanctioning procedures, and providing acceptable use cases are pressing tasks for universities. A part of the strategy to ensure the acceptable use of GenAI tools in academic work is likely to include the use of text detection software and faculty training. Given the high stakes of accusations of academic misconduct for students, it is vital to obtain detailed information about the relative accuracy of AI detection software in identifying cases where this may have been used, especially in real-life scenarios where academic judgement would be key in determining whether academic misconduct has occurred.

Study Objectives

The primary objective of this study is to explore the implications of AI-generated content in academic settings, with a particular focus on the detection and evaluation of such content. The specific objectives are as follows:

1. Assess the quality and perceived authenticity of academic papers generated by AI tools, specifically the GPT-4 model.
2. Evaluate the effectiveness of AI detection tools, particularly Turnitin's AI detection tool, in identifying AI-generated content, despite the use of advanced prompting techniques to attempt to evade detection.
3. Assess academic staff members' ability to use this information to judge whether academic dishonesty may have occurred.
4. Understand the perceptions of academic staff regarding AI-generated content. This involves analysing the marks and feedback provided by academic staff on the AI-generated papers and identifying common themes and patterns.
5. Explore the implications of AI-generated content for academic integrity, specifically for potential adjustments to assessment strategies and methodologies used by Higher Education Institutions (HEIs).

Significance of the Study

Our research provides valuable insights into the expanding body of knowledge in several critical areas. Initially, we concentrate on the evaluation of the quality and perceived authen-

ticity of academic papers generated by AI tools, specifically the GPT-4 model. This helps to gauge the potential impact of these tools on the academic landscape.

Second, we assess the effectiveness of AI detection tools, particularly Turnitin's AI detection tool in identifying AI-generated content. Although other GenAI text detection tools exist and our study is wholly independent, we focus on Turnitin's software for several reasons. The company's existing plagiarism detection software is widespread in higher education. Over 98% of universities in the United Kingdom use Turnitin (Turnitin.com, n.d.-b), and the software is used by over 15, 000 educational institutions in 140 countries worldwide (Turnitin.com, 2021). This high level of visibility and notoriety suggests that AI detection tools will become a mainstay in global higher education. In addition, the institution at which the study took place had already begun to use this software; thus, it provided an opportune moment to investigate the accuracy of the software in the current state of implementation. Furthermore, because the tool is standardised and operates globally, it provides a good basis for studying replicability in other parts of the world. Third, our study explores the ability of academic staff members to use the information provided by these detection tools during the marking process to make judgments regarding possible instances of academic dishonesty. This investigation is vital for maintaining the integrity of the academic process in the GenAI era. Fourth, we aim to understand the perceptions of academic staff regarding AI-generated content through interpretation and analysis of the written feedback given throughout the experiment and identifying common themes and patterns. Finally, we explore the implications of AI-generated content on academic integrity, specifically for potential adjustments to the assessment strategies used by HEIs. This insight may help educators and HEIs design assessment strategies and tasks that accurately reflect students' achievement of course learning outcomes, despite the increased availability and potential prevalence of these GenAI tools.

The body of scholarship on LLMs is notably immature (Bowman, 2023), but understanding the implications of these advanced AI tools in relation to academic integrity is crucial. As the field of AI continues to evolve, educators must adapt their assessment strategies to maintain academic rigor. One way to achieve this is to recognise the concept of cognitive offloading, a process described by Risko and Gilbert (2016) in which individuals rely on external resources to reduce their cognitive load. Dawson (2020) has equally described how assessment tasks can be developed in a way that makes the use of any such cognitive offloading tools transparent, potentially allowing for their legitimate use within the educational context. Dawson (2022, as cited in Sparrow, 2022) has also recommended that universities should clarify the extent to which AI tools can be used in cognitive offloading. However, such methods are yet to gain widespread appeal, and as a result, it is still essential to understand whether an assessment was written by the submitting student or by a GenAI tool.

While this research is predicated on the detection of AI-generated text, we do not assume that an effective method for solving misrepresentation of authorship and academic misconduct by adopting a 'detect-and-punish' approach is desirable. Instances of academic dishonesty do not always represent students deliberately seeking to gain an unfair advantage, but are nuanced, complex, and highly contextual events which can even be symptomatic of personal-social stress or struggle (Roe, 2022). Consequently, we take the epistemological position that by detecting AI-generated text in assessments, educators can find better ways to educate and train students, understand the use of AI, and adapt their assessment methods

to maintain rigor and robustness for the benefit of all. As we acknowledge the irreversible nature of GenAI's integration into our systems, we must design assessment strategies and methods that effectively gauge students' knowledge, while permitting some level of cognitive offloading. The findings of this study will serve as a basis for potential adjustments in the overall assessment strategies, methodologies, and guidelines across HEIs, thereby supporting academic integrity in the face of evolving AI technologies.

Structure of the Paper

This paper is structured into seven main sections: Introduction, literature review, methodology, results, discussion, limitations, future research, and conclusions. The literature review provides an overview of the existing research on AI-generated content and its detection. The methodology section details the experimental design, challenges in generating suitable content, and ethical considerations, while the results section presents the key findings. The Discussion, Limitations, and Future Research sections offer interpretations, implications, and suggestions for further investigation. Finally, the conclusions summarise the study's findings and contributions to the field of academic integrity and assessment design.

Literature Review

AI-generated Content and its Impact on Academic Integrity

Over the past decade, AI-generated content has become a significant concern for academic institutions worldwide. With the increasing sophistication of LLMs, such as OpenAI's GPT-3.5 and GPT-4, questions have been raised about their impact on academic integrity(Abd-Elaal et al., 2022; Köbis & Mossink, 2021). These models can generate convincing content that closely mirrors human-written text, presenting a significant challenge to maintaining academic standards and ensuring a fair assessment.

This advancement in AI has sparked serious ethical concerns related to the potential misrepresentation of authorship and authenticity of academic work (Cotton et al., 2023; Perkins, 2023; Rahman & Watanobe, 2023; Rodgers et al., 2023; Rudolph et al., 2023). Although tools have been designed to detect the presence of GenAI content, their effectiveness has limitations (Malinka et al., 2023; Rudolph et al., 2023; Uzun, 2023), especially when users deliberately try and evade detection methods (Sadasivan et al., 2023). Consequently, students may generate essays, research papers, or other academic assignments without acknowledging the assistance of AI or properly attributing the sources used by the AI model (Perkins, 2023; Strzelecki, 2023). Such actions undermine the core principles of academic integrity, erode the educational value of assignments, and compromise academic institutions' credibility. The debate on whether ChatGPT is a 'bullshit spewer' (Rudolph et al., 2023) highlights another concern: the propensity of LLMs to provide inaccurate or false information (Azaria & Mitchell, 2023). This can be an indicator of the use of LLMs in text production (Perkins, 2023).

Given these concerns, it is essential for institutions to establish clear guidelines and policies regarding the use of LLMs in the assessed work. Comprehensive guidelines should outline the appropriate use of AI-generated content, citation requirements, and limitations

(Crawford et al., 2023; Sullivan et al., 2023). The importance of proper attribution and students' responsibilities in maintaining the integrity of their work should be explicitly emphasised. With clear guidance, institutions can help students navigate potential academic compliance issues associated with AI-generated content and promote responsible practices (Okonkwo & Ade-Ibijola, 2021; Sohail et al., 2023; Strzelecki, 2023).

AI-generated Content and its Detection

Assessing the ability of higher education institutions (HEIs) to detect the use of LLMs in student work is crucial for maintaining academic integrity. Research on AI-generated content and its detection has primarily focused on earlier versions of GPT software, with no currently available examples of studies exploring the detection of GPT-4 generated content. Here we summarise the current situation regarding the detection of GenAI content.

In relation to GPT-2, Abd-Elaal et al. (2022) demonstrated the challenges faced by academic staff in identifying LLM-generated content and the potential benefits of training to improve detection. The results showed that participants could correctly identify LLM-generated samples 59.5% of the time, which was only slightly better than chance. Similarly, Köbis and Mossink (2021) and Gunser et al. (2021) found that LLM-generated poems are difficult for experts to differentiate from human-written work. Fröhling and Zubiaga (2021) introduced a low-cost detection model that accurately identifies GPT-2- and GPT-3-generated text. However, they noted the ethical challenges of deploying such detectors, which may inadvertently discriminate against NNES.

Bidermann and Raff (2022) demonstrated that more advanced models like GPT-J can evade detection tools like MOSS (Measure of Software Similarity), suggesting that newer LLMs might be even harder to detect. Gehrmann et al. (2019) proposed a tool called GLTR to improve LLM-generated text detection by humans, increasing the accuracy of untrained subjects from 54 to 74%. Solaiman et al. (2019) and Ippolito et al. (2020) also presented tools for detecting machine-generated text, although their applicability to newer LLMs remains untested. Research on GPT-3 detection suggests that human detection abilities decline as LLM complexity increases. For example, Kumar et al. (2022) found that participants from diverse backgrounds struggled to discern GPT-3-generated texts from human writing. Although not focused solely on academia, this study underscores broader detection challenges. Clark et al. (2021) reported that non-expert evaluators identified GPT-2-generated text with 57.9% accuracy, but GPT-3-generated text with only 49.9% accuracy. Training using LLM-generated examples marginally improved detection rates; however, Liang et al. (2023) found that by prompting GPT-3.5 to self-edit, detection rates for machine-generated scientific abstracts decreased from 100 to 13% across seven different detector tools.

There is an ever-growing number of commercial AI detection tools available (Uzun, 2023), including providers such as OpenAI (the developer of GPT-4 and ChatGPT) and Turnitin. GPTZero, a free-to-use tool, is the self-proclaimed most popular current AI text detector with over one million registered accounts (GPTZero, n.d.-b). They claim to be able to classify 99% of human-written articles correctly and 85% of AI-written articles correctly from several LLMs, including both ChatGPT (GPT-4) and Bard (LaMDA) (GPTZero, n.d.-a). OpenAI's detector proclaimed a far more marginal success rate, claiming to correctly identify only 26% of AI written text with a 9% false-positive rate, and has been discontinued at the time of writing (OpenAI, 2023b). Other services make much bolder claims Original-

ity.AI claim to be the most accurate available detector of ChatGPT generated content with 95.93% true positive detection rates, and 1.56% false positive rates (Originality.AI, 2023). Similarly, Turnitin's detection tool, released in April 2023, claims to have 98% confidence in the ability to detect AI-generated content while retaining a false positive rate of less than 1% (Turnitin.com, 2023). Other services, including Netus AI and stealthwriter openly promote themselves as being able to paraphrase GenAI content to evade detection by other services. Netus AI state '*As AI content generation evolves, we continue to stay ahead of the curve'* (Netus AI, n.d.).

Despite these claims of accuracy by producers of AI detection tools, Weber-Wulff et al. (2023) found that by testing 12 publicly available tools and two commercial tools for AI-generated text detection, there was a bias towards classifying text as composed by humans rather than AI, and techniques to hide AI-generated text can be highly effective. Elkhatat et al. (2023) presented the results of a range of AI text detector accuracies in determining whether samples (n = 20) were human written, and found that there was a variation amongst the detectors' ability to accurately determine whether the text was machine or human generated, with the likelihood of accurate detection falling when GPT-4 generated content was tested.

Current methods used to identify machine-generated text can be broadly categorised into three groups: (1) identifying the presence of 'watermarked' content, (2) statistical outlier detection methods which seek out irregularities in the text generated, and (3) classifiers that have been trained to distinguish between text generated by a machine and text written by a human (Krishna et al., 2023). Watermarking is promoted as a solution to increase detection capabilities (Kirchenbauer et al., 2023). Sadasivan et al. (2023) demonstrated that the application of paraphrasing tools in addition to generative text production can invalidate the use of detection software, leading to the conclusion that even with watermarked text, highly sophisticated detection software can be easily evaded. Krishna et al. (2023) determined that before using a paraphrasing tool, DetectGPT accurately identified 70.3% of model-generated sequences from GPT2-XL. However, after using a paraphrasing tool to manipulate the content, the detection rate dropped to only 4.6%. It is critical to note that this was achieved using nominal semantic alterations. Solaiman et al. (2019) and Ippolito et al. (2020) also presented tools for detecting machine-generated text, although their applicability to newer LLMs remains untested.

Although these studies exploring detection methods all present challenges, the detection of AI-generated text is not necessarily an unsolvable problem. Chakraborty et al. (2023) derived a complexity-bound formula to demonstrate that detection of AI-generated text is almost always possible in certain scenarios and with adequate sample sizes, and watermarking of text for example, has shown potential in detection of AI generated content. However, as Lancaster (2023) points out, it is likely that detection will not be a single, all-encompassing solution; rather, a whole-of-system approach that includes an interlinked web of principles and practices, which includes detection as one component, will be required. The limited success of both human participants and detection tools in accurately detecting GenAI content underscores the threat to academic integrity in HEIs and the need for increased understanding of how detection works in practice, especially when advanced prompting techniques are used to obfuscate the nature of the text.

Advances in AI LLMs: GPT-4

The release of the GPT-4 LLM in March 2023 marked a major step in the journey of AI models in coming closer to achieving a true Artificial General Intelligence (AGI) system which can, in theory, perform many of the tasks that a human can (Bubeck et al., 2023), and is the underlying model behind ChatGPT Plus, a paid-for-service offered on a subscription basis by OpenAI, as well as Bing search (Microsoft, 2023).

As with previous versions of the GPT LLM, GPT-4 utilises a transformer-based model with self-attention for natural language processing (NLP) (Vaswani et al., 2017) and autoregressive modelling for time-dependent data analysis (Brown et al., 2020). However, it can accept inputs up to 16 times larger than GPT-3.5, can generate outputs of up to 24,000 words: more than eight times larger than GPT-3.5 (Koubaa, 2023), and is now able to recognise and interpret visual inputs such as videos and images (Campello de Souza et al., 2023). It has also shown improvements in reducing so-called hallucinations and inappropriate responses, although no data on the extent of this reduction is given (OpenAI, 2023a) whilst demonstrating advanced expertise in completing standardised tests such as a simulated bar exam and Graduate Records Examination (OpenAI, 2023a), and standardised IQ tests (Campello de Souza et al., 2023). Recently, Zhang et al. (2023) showed that GPT3.5 was able to solve a third of the assessment problems curated for undergraduate degrees in mathematics, electrical engineering, and computer science courses, while this increased to a perfect solve rate for GPT-4.

The existing literature suggests that a major challenge in maintaining academic integrity is the ongoing advancement in AI LLMs, given the inability to accurately differentiate between human and AI-generated content. Although detection tools offer a degree of promise, their effectiveness can vary considerably, particularly when the content has been significantly altered or paraphrased. Notably, this research presents a key gap: there are no studies evaluating the ability of academic staff to identify AI-generated content produced by modern LLMs. Given the improvements in the GPT-4 LLM identified above, this means that GPT-4 based LLMs should be able to perform at an even higher level than earlier models when it comes to developing outputs that will evade detection by academic staff. Furthermore, no studies have examined the efficacy of AI detection tools under conditions that mirror the real-world scenarios often encountered during student-paper assessment. This gap is particularly problematic for academic integrity, as it questions the principles of fairness and equality inherent in assessment methods used in HEIs across the globe. This study aimed to address this gap by exploring areas that have not yet been explored in the literature.

Methodology

Aims and Context

This study, conducted between March 27th and April 07th 2023 was designed to test whether academic staff members in an HEI could correctly identify whether text purporting to be student-written was created using GenAI tools. By doing so, we can identify the robustness of the current assessment strategies commonly used by HEIs, in light of the new era of GenAI tool availability. To do so, experimental submissions responding to 22 assignment

briefs were created by a research team and marked by 15 different academic staff members alongside genuine submissions by students during the usual university grading period.

The study was conducted at a private university located in Southeast Asia with approximately 2000 students in four Schools. The university follows a UK-based curriculum and has achieved international quality accreditation through the UK Quality Assurance Agency (QAA). The language of instruction is entirely English, with most students being Non-Native English Speakers (NNES), studying and writing in a second language. Assignment briefs from the School of Business were chosen for use in this study because it is the largest School, both in terms of the number of assessments from different modules being due for submission at any one time, and the number of students submitting work.

Experimental Design

The following methodology was used to develop experimental submissions and test the ability of the academic staff members to detect AI-generated content.

1. A list of all end-of-semester assessments for the January 2023 semester in the School of Business was reviewed (n = 70) to identify take-home assignments where GenAI tools could potentially be used by students.
2. Assessment methodologies such as presentations, demonstrations, and exams were excluded, and the assignment briefs for the remaining module assessments were downloaded by the research team (n = 25).
3. Using GPT-4 accessed through the ChatGPT Plus interface, the research team attempted to develop experimental submissions in response to assignment briefs. These included essays, reports, and case analyses over a diverse spectrum of topics, including e-commerce portfolios, development of personal development plans, marketing strategies, consumer behaviour reports, reflective pieces of work experience, and essays on supply chain management. The nature of some tasks meant that not every brief led to the creation of an experimental submission (n = 22). The challenges in creating experimental responses and the various techniques used to reduce the likelihood of detection by AI detection tools are discussed in the following section.
4. A fabricated student profile was created using the university Learning Management System. This student was then enrolled in all modules, where experimental submissions could be created by the research team. These responses were submitted for grading by the study participants along with genuine student submissions (n = 963.
5. The study participants (n = 15) consisted of the academic staff members responsible for marking the 22 different module assessments. The participants were informed about the presence of the experimental cases in their grading portfolio and were asked to carry out all grading as usual but to report any submission they suspected of being AI generated through the standard academic misconduct process used by the university.
6. Blind grading of all submissions was carried out by the participants using the Turnitin Feedback Studio as per standard university procedures. All participants had experience using this system to grade student work. The detected percentage of AI-generated content was available to participants through the Turnitin user interface, and all academic staff members in the university received guidance about what these scores indicated.

7. Following the marking process, debriefing was provided to all participants to inform them of which paper was the experimental submission, and training was provided to all faculty members to support them in identifying the AI-generated content.

The choice to use the GPT-4 model accessed through ChatGPT Plus for the creation of the experimental submissions was due to the novelty of this software and the lack of any prior research which had been carried out using this tool. This model was released to the public on March 14th 2023, and all experimental submissions were created between the 27th Match and the 7th April (the submission deadline for all assignment briefs). Given the recent release of the software prior to the start of the study, it is likely that the Turnitin AI detection tool used in the study had not yet been optimised for detecting outputs from this model. Therefore, the results of the study were not affected by prior training of the tool on GPT-4 outputs. This was confirmed by a Turnitin representative following the release of the preprint of this paper (J. Thorley, personal communication, May 31, 2023).

Challenges Faced in Generating Suitable Content

Standard essays or assignments that prompted students to choose a country, organisation, or topic and complete specific tasks were relatively straightforward to generate by the research team. More challenging were tasks that required reflections on "lived experiences", such as internship experiences or personal development plans. These could be created by providing small amounts of additional details when creating the prompts entered in ChatGPT, for example, by providing information on student societies. This helped in the production of assessments that matched the context of the university and were less likely to be identified as generated by AI. Some assessment tasks proved more challenging. For example, tasks that required primary research, approval of the use of selected databases, or approval of topic choice before beginning the task were unsuitable for creation using GenAI tools and could not be developed. However, this does not mean that GenAI tools could not be used by students participating in similar assessments but that these cases were not suitable for inclusion in this study.

During the submission creation process, the research team engaged in so-called prompt engineering techniques to obtain responses that met assessment requirements, but also aimed to evade possible detection by academic staff and AI detectors. These included requests to ChatGPT to produce output at a level of complexity or language comprehension more typically seen in NNES students ('*The language level is too high, reduce it significantly*'), requests to add spelling or grammatical errors for authenticity ('*Include 2 grammatical errors in the conclusion*'), regulating output word count ('*The provided text is only 1400 words, Expand on all sections'*, and ensuring real academic references ('*Are these references real? Provide only real references*'). Despite Marche's (2022) claims about the obsolescence of traditional essays due to GenAI tools, it proved challenging to align the output of AI-generated content with the required word and content limits that could feasibly be produced by an NNES student, with repeated requests needing to be made to modify the content according to the research team's specifications. However, the output provided by ChatGPT had a tendency to revert to its typical style, resist incorporating deliberate errors, and continue to provide non-existent references–a key indicator of GenAI-created content, according to Perkins (2023). Although Sadasivan et al. (2023) and Weber-Wulff et

al. (2023) highlighted the use of paraphrasing tools as a key method for evading AI detection software, one of the aims of this study was to explore the potential of GPT-4 generated content to evade detection. Therefore, paraphrasing tools were not used, and only the output created directly by GPT-4 was included in the experimental submissions. Apart from the basic formatting of text (such as providing titles in bold) to create realistic looking submissions, this meant that the research team also chose to make no other manual adjustments to submissions to help them better fit the requirements and potentially achieve higher marks.

Ethical Considerations

The study was approved by the University Registrar, Dean, and University Human Ethics Committee (Ref: HECA-01/2023-MIKE) before commencement of the study. The academic staff members involved in the study were informed about the research, freely provided informed consent, and were offered the choice to opt out at any time. This decision was made after extensive discussions within the research team about whether it would be possible to avoid informing academic staff members about the experimental submissions and utilise a blind study methodology. However, it was ultimately determined that given the additional time requirements to mark these experimental submissions, academic staff members needed to be given an option to opt out of the study. Students' privacy was protected by creating a separate profile for submitting AI-generated content, ensuring that no student profile was linked to the test submissions. During the testing process, faculty members were requested to follow all standard procedures for reporting potential cases of academic misconduct, and all cases were centrally examined and investigated. This reduced the likelihood of any student being potentially affected by the experiment.

Results

Summary Results

Table 1 presents the results of the study. This table shows a summary of the assignment task for each of the 22 experimental submissions, the number of genuine student submissions for that assessment, the mean grade achieved by the genuine students, the mean grade of the experimental submission, the percentage above or below the mean grade obtained by the experimental submission, the detected AI percentage of the experimental submission identified by the Turnitin AI detection tool, and whether the paper was reported through formal processes as a potential breach of academic conduct regulations by the study participants. The papers are sorted in descending order by the percentage of AI-generated content identified by the Turnitin AI detection tool.

Table 2 summarises the results for all 22 assessments in which experimental submissions were created.

The findings of this study shed light on the significant disparity in the ability of academic staff to identify AI-generated content. In total, 22 academic papers were submitted as part of the investigation, of which a little over half, 12 (54.5%), were identified as potentially AI-generated by the academic staff. The mean score for genuine submissions was 54.4, whereas the AI-generated content received a mean score of 52.3, showing an average deviation of

-4.9% from the true submissions. The mean percentage of AI content detected by Turnitin AI detection was 54.8%, with 91% of the submissions being highlighted as containing some AI-generated content and potentially raising the suspicion of markers. No significant differences were found between the average scores of the detected (51.7%) and undetected papers (52.6%). This result indicated that the detection of AI-generated content did not substantially affect the grading process.

The feedback provided by academic staff members offers valuable insights into the quality of AI-generated submissions. One key aspect that emerged was the differing perceived quality of the papers generated by AI. Some markers praised the quality of the work, with one stating that the submission has "*many ideas that might be worthwhile to pursue*" and noting that the paper was "*well supported with good research and good clear thinking*" (Case 1). On the other hand, other markers found the submission to be lacking in depth or focus, with one stating that the paper "*lacks in-depth research*" (Case 7) and another noting the "*unfocused*" nature of the paper (Case 21).

Certain characteristics of the AI-generated papers, such as writing style and structure, were noted in the feedback from the academic staff. For example, one marker mentioned the "*confusing*" writing style (Case 12), while another noted the lack of proper referencing (Case 18). Similarly, another marker highlighted the "*lack of personality and visuals*" (Case 15). In other cases, markers criticised the paper for its "*very long introduction*" and "*problem with sources*" (Case 8), while another noted that the paper "*did not address the Learning Outcomes*" (Case 3). However, it is important to note that we cannot determine whether the presence of these specific features specifically influenced academic staff to make these comments, as it is plausible that the AI percentage score provided by Turnitin could have influenced their evaluations and guided their comments.

It is important to note that during this marking period, there were cases in which 'genuine' submissions by students were also reported by academic staff members as potential cases of academic misconduct due to the presumed misapplication of GenAI tools, and these judgements were supported using the Turnitin AI detection tool. Due to the confidential nature of these investigations, we cannot reveal the number of students who were identified/penalised following allegations of misconduct but wish to highlight the thorough nature of the academic misconduct process in place in the HEI where the experiment was performed. Following reports from faculty, potential cases of misconduct are investigated centrally, and proceed through several panels and discussions with students before any penalties are applied. The results of the AI detection tool were used as a starting point to investigate whether breaches of academic integrity may have occurred and not as sole evidence to penalise students or reduce marks. Other techniques used to investigate whether AI tools may have been (mis)used included viva voce exams with students, comparisons with past submitted work, reviews of draft work or previous versions of work prepared before submission, and the examination of citations and references for authenticity.

Table 1 Details of all assignment briefs in which experimental responses were created and the results obtained by both experimental submissions and genuine student submissions

Case	Summary of required task	# of genuine student submissions	Mean grade of genuine student submissions	Experimental submission grade	% grade for experimental submission above or below genuine student mean grade	Turnitin detected AI % of experimental submission	Experimental submission reported by participant as academic misconduct?
1	Organization reputation & brand management essay.	15	67	70	+4%	100%	No
2	Strengths personal reflection.	66	46	33	-28%	100%	Yes
3	E-commerce portfolio: search engine optimization, databases, communications.	67	51	35	-31%	85%	Yes
4	RACE framework digital marketing report.	97	56	51	-9%	81%	Yes
5	Credit policy essay: improvement, efficiency, data, risk management.	10	57	49	-15%	81%	Yes
6	Consumer behaviour analysis & strategy recommendations.	5	52	75	+44%	80%	Yes
7	Organizational structure analysis & recruitment recommendations.	66	36	38	+5%	79%	Yes
8	Accountancy changes due to online technology.	5	57	50	-12%	78%	Yes
9	Global supply chain competitive advantage essay.	57	54	65	19%	78%	Yes
10	Innovation-based marketing plan.	99	59	18	-69%	75%	No
11	Leadership strategies & change impact evaluation.	27	55	68	+24%	70%	No
12	Local market PESTEL analysis & marketing strategies.	7	70	50	-29%	48%	No
13	Personal Development Plan: SWOT analysis and SMART goals.	66	46	56	+23%	46%	Yes
14	Local financial markets, debt collection & regression analysis essay.	26	56	52	-7%	40%	No
15	Objective setting & budgeting business blog.	10	61	65	+6%	30%	Yes
16	Work experience reflection & recommendations.	6	NA*	NA*	NA*	18%	Yes
17	Digital marketing blogs: web presence, social media.	44	48	49	+3%	18%	No
18	Career action plan with skills gap analysis.	54	59	70	+18%	13%	No
19	Managerial work experience reflection evaluation.	6	NA*	NA*	NA*	13%	Yes
20	Company reputation & brand analysis, improvement, communication strategies.	51	60.4	53	-12%	4%	No

Table 1 (continued)

Case	Summary of required task	# of genuine student submissions	Mean grade of genuine student submissions	Experimental submission grade	% grade for experimental submission above or below genuine student mean grade	Turnitin detected AI % of experimental submission	Experimental submission reported by participant as academic misconduct?
21	Agile business essay: service operations & targets.	88	48	62	+29%	0%	No
22	Personal development plan: communication, self-management.	91	50	37	-26%	0%	No

* These two papers were identified as test student submissions by the markers and were not awarded a grade. They have been excluded from the calculation of the mean grade

Table 2 Summary of data for all 22 assessments in which experimental responses were created

Value	Figure
Total number of genuine student submissions	963
Mean number of student submissions in each marking cohort	44
Mean grade of genuine student submissions	54.4
Mean grade of experimental submissions	52.3
Mean percentage grade for experimental submissions above or below genuine student mean grade	-4.9%
Mean Turnitin detected AI % of experimental submissions	54.8%
Mean % of experimental submissions highlighted by Turnitin as containing some AI-generated content	91%
% of experimental submissions reported by academic staff through the formal academic misconduct process.	54.5%

Discussion

Interpretation of Results

The results of this study provide valuable insights into the robustness of university assessments against GPT-4 generated content. Turnitin's ability to detect 91% of the generated submissions containing AI-generated content, despite the deployment of prompt engineering techniques by the research team to evade detection, is promising. As it is likely that any detection will raise suspicion of markers assessing the paper for potential academic misconduct violations, this shows that Turnitin AI detection may be a valuable tool in supporting academic integrity. However, as only 54.8% of the total amount of AI-generated content was detected by the Turnitin AI detection tool, there is still significant room for improvement, as demonstrated by the challenges faced by academic staff in accurately detecting AI-generated papers, even with Turnitin's results available to them.

The similar average scores between the detected and undetected papers imply that the detection of AI-generated content did not significantly influence grading. The lower grades obtained by submissions where the task had additional requirements highlight the role of assessment design in mitigating the potential impact of AI-generated content on grading outcomes. Incorporating assessment tasks that require the application of specific frameworks or the pre-approval of topics by the assessor and datasets may serve as additional barriers to AI-generated submission. For example, requiring a student to use a specific dataset that cannot be easily imported into an online platform could theoretically reduce the likelihood of using GenAI to produce a written analysis.

It is important to emphasise the differences between the effectiveness of Turnitin in identifying AI-generated content and the faculty's ability to use this information and make a judgement as to whether this constitutes an academic misconduct violation. Although Turnitin correctly identified 91% of the papers as containing AI-generated content, faculty members formally reported only 54.5% of the papers as potential cases of academic misconduct. This may have occurred because of the underestimation of the percentage of AI-generated content detected by Turnitin in many cases or because faculty members felt there was not enough conclusive evidence to report AI use. The mean percentage of content identified by the Turnitin AI detection tool was only 54.8%,

despite 100% of the content of all papers being generated using AI tools. This is in direct contrast to the claimed results that a maximum of 15% of AI written text is missed in AI reports (Turnitin.com, n.d.-a), indicating that prompt engineering is an effective method for evading AI detection tools. This disparity may have led faculty members to believe that cases with a small amount of AI-detected content were either false positives (as can sometimes be seen in Turnitin similarity detection tools), or that a small amount of AI tool usage was acceptable. For example, one test submission received a Turnitin AI score of 13%, but the academic staff member still marked it as authentic, possibly because of the perceived relatively low percentage of AI-generated content (Case 13). However, some papers with very high AI scores (for example, Cases 1, 10, and 11) were not reported as potential academic misconduct violations. Debriefing with the faculty post-experiment highlighted that these cases were perceived as genuine.

This highlights the need for additional training to be provided to faculty members to enable them not only to rely on the raw results of the Turnitin AI detection tool but also to recognise the identifiers of potential AI tool usage that were experienced by the research team in the development of submissions. Some of these identifiers include overly complex language, papers not meeting length requirements, core texts and content discussed in class not being covered, and falsified references. A further consideration is that shorter and non-continuous texts are less likely to be accurately detected by the software (Turnitin.com, n.d.-a) and are, therefore, not flagged as AI content. Consequently, the use of non-continuous text (e.g. bullet points) may be an adversarial strategy that can be employed by well-informed students to circumvent AI detection tools. Adopting such an approach, combined with other interventions as mentioned above, severely limits the effectiveness of both AI tools and faculty observation in identifying AI-generated texts.

In three cases (15, 16, and 19), the marker not only reported the cases as potential academic misconduct offences but was also able to correctly identify the experimental submissions. The following feedback was provided for case 15:

> '*Dear Research Team, the lack of personality and visuals was the giveaway for this paper. I had also given a company [meaning a specific business was assigned as a focal point] to each group. However, in a large cohort this could slip through the net'.*

This assessment had only 10 genuine submissions, suggesting that with a smaller set of assessment submissions, markers may be more aware of the specific circumstances of each student or group of students' writing capabilities and their progress on the assessment task and would be more able to detect the usage of AI tools in submitted work. As a result, it is likely that AI detection tools will be of less benefit in smaller cohorts, in which the instructor has a greater deal of insight into student ability. However, larger cohorts which include hundreds of students and may be taught by a team of teachers will lead to less familiarity with the cohort and its expected attainment, meaning that detectors are likely to be more useful. One method which may help reduce the likelihood of AI-generated content being used in larger cohorts is to integrate group-based assignments. With multiple group members participating in the production of work, there is an additional incentive to self-police the disallowed use of GenAI tools by individuals who might otherwise use them for their individual work. The feedback provided by the assessor suggests thematically that a clue to the AI-generated text is that of a 'lack of personality', which might be interpreted as the

writing not having any demonstrable errors, mistakes, and lacking an idiolect or individual style of writing. Through the use of prompting in AI software, it is possible to ask LLMs to write in specific styles, include mistakes, and write in a 'non-generic' manner or mimic a style by providing training data. This also demonstrates an area for further development of assessor understanding.

In some cases, faculty members still provided relatively high marks to papers which had high AI scores and were correctly identified as being generated by AI, such as a paper with a Turnitin AI score of 80% that received a mark of 75 out of 100 (Case 6). Under the UK grading system, this equates to a first-class result, which is the highest possible classification of results. This suggests that AI-generated content may still be perceived as valuable or well-structured, potentially raising concerns about the reliability and integrity of assessments if AI detection tools are not used, and ultimately undermining and minimising the cognitive ability of the students participating in higher education.

A notable number of outliers with very low attained scores were observed, particularly in assessments that required the use of frameworks not specified in the assessment paper or the pre-approval of topics and data sets (Cases 2, 3, 7, 10, and 22); however, in some of these cases there was no indication of AI tool usage by either Turnitin AI detect, or the individual faculty member. Closer explorations of the feedback provided in case 22 revealed that some academic staff members expected specific usage of tools from core texts: '*You have not used any of the SWOT from your text book'*, or that additional exemplars were provided to students in class and on the Learning Management system but not used in the test submission: '*This PDP is meant to be in paragraphs as demonstrated in the exemplar on Canvas [The LMS system used in the institution]'.* This suggests that even though GenAI tools available to students may support them in the development of work used for submission, without engagement with course materials or attendance in class, the final marks achieved may be very low.

The results of this study highlight the potential of AI detection tools in identifying AI-generated content but also underscore the challenges faced by academic staff in detecting such content. There remains a gap in both faculty members' ability to recognise and respond to such content, as well as in the accuracy of the software in identifying GenAI text which has been produced using adversarial methods such as prompt engineering. The results obtained from the current experiment demonstrate that these tools are not infallible and must be combined with appropriate training to support faculty in developing assessments which are robust to the use of GenAI tools, as well as spotting the signs of these tools being used. Providing additional training and support to faculty members will help develop a more comprehensive understanding of AI-generated content indicators and improve the overall integrity of the assessment process.

Implications for AI Detection Software Development

The results of this study also have implications for the ongoing development of AI detection software. While Turnitin demonstrated a broadly accurate detection rate of GenAI content, continuous improvements and updates to AI detection algorithms will be necessary to keep pace with the evolving capabilities of AI models such as GPT-4. Collaboration between

academia and AI detection software developers is essential for ensuring the effectiveness of these tools in maintaining academic integrity.

However, it is important to recognise that this 'arms race' scenario between GenAI tools and AI detectors is not sustainable. Although Turnitin claims that its detector mostly retains the ability to identify GenAI content after paraphrasing (Turnitin.com, n.d.-a), the results of this study show problems with this claim. In addition, Sadasivan et al. (2023) demonstrated that the accurate detection of AI-generated text by software can almost always be thwarted by adversarial methods such as the use of Automated Paraphrasing Tools (APTs) or cross-translation into different languages. Students will inevitably find ways around any such tools used to detect AI output, so it is likely to be more effective to focus on the adjustment of assessment strategies to either discourage AI-generated submissions or change assessment strategies entirely to encourage or accept their use rather than relying solely on detection software.

Additionally, AI detection software can sometimes produce false positives, particularly when faced with work written by non-native English speakers (NNES) (Liang et al., 2023). It has been demonstrated that students with lower levels of English proficiency are more likely to commit academic misconduct violations (Perkins et al., 2018), Fröhling & Zubiaga, (2021) highlight the ethical challenges of deploying AI detectors, which may potentially discriminate against EFL students by incorrectly identifying human-created text as machine-written. This is a particular concern in higher education institutions with a high concentration of non-native English-speaking students, as the use of AI detection software can inadvertently undermine the academic integrity of their work. Considering these challenges, it is crucial for higher education institutions to strike a balance between utilising AI detection software and adapting assessment strategies to account for the evolving landscape of AI-generated content. By working in tandem with AI detection software developers and incorporating more robust assessment methods, higher education institutions can uphold academic integrity while recognising the potential benefits and limitations of AI tools in the academic environment.

Implications for Assessment Design and Academic Integrity

The findings of this study have several implications for assessment design and academic integrity in higher education. First, incorporating assessment tasks that require unique frameworks, pre-approved topics, or datasets can reduce the likelihood of successful AI-generated submissions. By designing assessments that are either more challenging for AI tools to complete or where their use could be easily identified, institutions can encourage students to engage more deeply with the subject matter and rely less on AI-generated content.

A move towards overall assessment strategies involving a significant number of in-class written or oral examinations, presentations, or invigilated assessments may also be an option that some HEIs may choose to use. An alternative approach would be to recognise that a certain amount of cognitive offloading using GenAI tools could be considered acceptable for students. For example, institutions can increase the depth of detail required in assignments, knowing that students may rely on AI tools to help generate content. This approach acknowledges the reality of the presence of AI tools in the academic environment and encourages students to use them in a responsible and productive manner (Kumar, 2023).

Designing assessments that explicitly require the use of AI tools and include marks for discussing and critically evaluating the prompt engineering that occurs to generate an effective output would take this step further. As graduates are likely to use these tools in their future careers, institutions should prepare students by training them in responsible and effective use of AI tools. By incorporating AI tools into the assessment process, HEIs can ensure that students are ready to leverage the benefits of AI in their professional lives, while maintaining a strong commitment to academic integrity.

We recognise that many HEIs may not be willing or able to take such an approach, and would therefore focus more on the detection of GenAI content. In this case, HEIs must also focus on providing training and resources to faculty members to improve their ability to identify AI-generated content and train and support students to use these tools in an appropriate manner. As academic misconduct training has been identified as a successful approach to mitigating instances of academic misconduct (Perkins et al., 2020), this will foster a more comprehensive approach to academic integrity that is not solely reliant on AI detection tools.

Limitations and Future Research

Despite providing key insights, this study has certain limitations. The research was conducted within a single Business School, which could limit how widely the findings can be applied to other fields or institutions. Moreover, the small group of academic staff that evaluated the AI-produced papers may not fully capture the range of possible responses that markers may have to the AI-generated content. The possibility of faculty reusing comments from comment banks, a common practice to save time and maintain consistency, was also not assessed in this study and could have influenced the responses provided.

The novelty of the Turnitin AI detection software could also have affected the results. This was the first time the academic staff had used this tool, and although all participants were used to marking with Turnitin, the AI score provided is a distinct concept from the commonly used similarity index where, unlike the AI index, a low level may be overlooked or considered normal by markers. Such unfamiliarity could have potentially skewed their perception or evaluation of which papers were AI generated, underlining the need for further training and familiarisation with these novel tools in future studies. This study focuses on the content created by one LLM: GPT-4. As other LLM models, such as Bard (Google), LLaMA (Meta), and Claude (Anthropic) are becoming more popular, future research would benefit from including these models to ensure a broader perspective. Finally, given the small scale of this study, we should be careful when extending these results to broader contexts.

Extending the scope of the study to include larger participant groups, broader ranges of assessment types, and multiple institutions and disciplines would further enhance the generalisability of the findings. Additional research could also explore the impact of specific faculty training programs on the ability to detect AI-generated content and determine the most effective strategies for training and supporting students in the ethical use of these tools. As more advanced AI models continue to emerge, future studies should evaluate the efficacy of evolving AI detection tools in identifying the content generated by these sophisticated models.

Conclusion

The rapid evolution of AI technologies and their adoption in academia raises crucial questions regarding their implications for academic integrity and assessment processes. This study focuses on the capacity of Turnitin's AI detection software and academic faculty to identify AI-generated content and provides initial insights into this complex landscape.

While the findings demonstrate the potential of Turnitin's AI detection tool in supporting academic staff in detecting AI-generated content, the relatively low detection accuracy of the participants underscores the need for further training and awareness. The lack of significant differences between the mean scores achieved by the AI-generated submissions and the student submissions further highlights the potential implications for the integrity of university assessments, especially if AI detection tools are not utilised. Additionally, the existence of potential strategies for defeating AI detection, such as the use of paraphrasing tools and prompt engineering techniques to adjust the output of LLMs, further emphasises the significance of this challenge. The results demonstrated that the Turnitin AI detection tool was not particularly robust to the use of these adversarial techniques, and there will likely be a continued 'arms race' (Fröhling & Zubiaga, 2021); Roe & Perkins, 2022) between the techniques employed to evade detection and the tools designed to detect GenAI content.

The social dimensions of AI- and LLM-generated text and detection should not be underestimated. An area that has not been touched on in this research is the viability and accessibility of detection tools in higher education worldwide. Although there has been explosive growth in the AI field, access to technology in learning environments globally remains inequitable across socioeconomic divisions and geographic locations (Reimers et al., 2020). The disparity in digital access and AI exposure, often termed the 'digital divide', can lead to further educational inequalities (Cullen, 2001). Moreover, the financial burden of implementing AI detection, particularly in developing countries and underprivileged communities, can be cost prohibitive. Turnitin has made clear that the AI detection feature will require additional subscription costs from January 2024 (Turnitin.com, n.d.-a), suggesting that even when such tools are effective, there are limitations to their ease of implementation and scalability, and it is likely that HEIs with less access to resources will face the greatest risk of GPT-generated content undermining academic integrity.

This study also raises questions regarding the ongoing development of AI detection software and its effectiveness. As advanced AI models and prompt engineering techniques continue to evolve, there is a pressing need for the constant improvement and updating of detection algorithms. However, it is important to acknowledge that detection software alone is insufficient to maintain academic integrity. The future of academia should focus on striking a balance between harnessing AI technology responsibly and maintaining the sanctity of its academic integrity. The development of assessment tasks that explicitly involve the use of AI tools coupled with comprehensive training programmes for faculty and students could shape the future of assessments in Higher Education. Not only will this promote the responsible use of AI tools, but it will also equip students with skills that will be invaluable in their future careers, while upholding the integrity of the assessment process.

While this study is an initial step towards understanding the interplay between AI-generated content and academic assessment, it has also opened the door to further research. As AI continues to evolve, our focus should also be on understanding its continued and significant impact on education. We encourage HEIs, academic staff members, and software developers to continue to engage in this area of research, contributing to the development of robust strategies that both utilise the potential of AI and preserve the principles of academic integrity. Ultimately, this study serves as a call for academic institutions and faculty to adapt to the rapidly changing landscape of AI technology. It has been highlighted that HEIs do not quickly adapt their policy approaches to AI tools (Perkins & Roe, 2023). As AI continues to permeate various facets of our lives, it is crucial that the academic community remains abreast of these advancements and their implications for the integrity of our assessment processes and overall quality of education.

Acknowledgements The authors are very grateful for the support of the academic staff who participated in the study and the ongoing support of the university Examinations Office team who provided technical assistance throughout the project.

Author Contributions Mike Perkins conceived and designed the study. Data collection and analysis were performed by the authors Mike Perkins, Darius Postma, James McGaughran, and Don Hickerson. The first draft of the manuscript was written by Mike Perkins, and all authors subsequently revised the manuscript. All authors have read and approved the final manuscript.

Funding No funding was received for this study.

Data Availability The full commentaries left by academic staff members when marking the test submissions cannot be shared because of the confidentiality of the assessment results and associated comments. All other data generated or analysed during this study are included in this published article.

Declarations

Competing interests Mike Perkins, Darius Postma, James McGaughran, and Don Hickerson are currently employed by the university where the study took place. Jasper Roe was previously employed at the same university. This study was not connected to or funded by Turnitin.

Ethics Approval and Consent to Participate The study was approved by the institution's Human Ethics Committee before commencement, and all participants provided informed consent with the option to opt out of the study at any point in time.

LLM Usage This study used Generative AI tools to produce draft text, and revise wording throughout the production of the manuscript. Multiple versions of ChatGPT over different time periods were used, with all versions using the underlying GPT-4 Large Language Model. The authors reviewed, edited, and take responsibility for all outputs of the tools used in this study.

References

Abd-Elaal, E. S., Gamage, S. H., & Mills, J. E. (2022). Assisting academics to identify computer generated writing. *European Journal of Engineering Education*, 1–21. https://doi.org/10.1080/03043797.2022.2046709.

Azaria, A., & Mitchell, T. (2023). The Internal State of an LLM knows when its lying. *arXiv*. https://doi.org/10.48550/arXiv.2304.13734. arXiv:2304.13734.

Baidoo-Anu, D., & Owusu Ansah, L. (2023). *Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning* (SSRN Scholarly Paper 4337484). https://doi.org/10.2139/ssrn.4337484

Biderman, S., & Raff, E. (2022). Fooling MOSS Detection with Pretrained Language models (arXiv:2201.07406). *arXiv*. https://doi.org/10.48550/arXiv.2201.07406.

Bowman, S. R. (2023). *Eight things to know about large language models* (arXiv:2304.00612). arXiv. https://doi.org/10.48550/arXiv.2304.00612

Bretag, T., Harper, R., Burton, M., Ellis, C., Newton, P., Rozenberg, P., Saddiqui, S., & van Haeringen, K. (2019). Contract cheating: A survey of Australian university students. *Studies in Higher Education*, *44*(11), 1837–1856.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html?utm_medium=email&utm_source=transaction.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4 (arXiv:2303.12712). *arXiv*. https://doi.org/10.48550/arXiv.2303.12712.

Campello de Souza, B., Serrano de Andrade Neto, A., & Roazzi, A. (2023). *Are the New AIs Smart Enough to Steal Your Job? IQ Scores for ChatGPT, Microsoft Bing, Google Bard and Quora Poe* (SSRN Scholarly Paper 4412505). https://doi.org/10.2139/ssrn.4412505.

Cassidy, C. (2023, April 16). Australian universities split on using new tool to detect AI plagiarism. *The Guardian*. https://www.theguardian.com/australia-news/2023/apr/16/australian-universities-split-on-using-new-tool-to-detect-ai-plagiarism.

Chakraborty, S., Bedi, A. S., Zhu, S., An, B., Manocha, D., & Huang, F. (2023). *On the Possibilities of AI-Generated Text Detection* (arXiv:2304.04736). arXiv. https://doi.org/10.48550/arXiv.2304.04736.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All That's `Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 7282–7296. https://doi.org/10.18653/v1/2021.acl-long.565.

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 1–12. https://doi.org/10.1080/14703297.2023.2190148.

Crawford, J., Cowling, M., & Allen, K. A. (2023). Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). *Journal of University Teaching & Learning Practice*, *20*(3), 02.

Cullen, R. (2001). Addressing the digital divide. *Online Information Review*, *25*(5), 311–320. https://doi.org/10.1108/14684520110410517.

Dawson, P. (2020). Cognitive Offloading and Assessment. In M. Bearman, P. Dawson, R. Ajjawi, J. Tai, & D. Boud (Eds.), *Re-imagining University Assessment in a Digital World* (pp. 37–48). Springer International Publishing. https://doi.org/10.1007/978-3-030-41956-1_4.

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2021). Some students plagiarism tricks, and tips for effective check. *International Journal for Educational Integrity*, *17*(1), 1–12. https://doi.org/10.1007/s40979-021-00082-w.

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*(1), https://doi.org/10.1007/s40979-023-00140-5.

Foltynek, T., Bjelobaba, S., Glendinning, I., Khan, Z. R., Santos, R., Pavletic, P., & Kravjar, J. (2023). ENAI recommendations on the ethical use of artificial intelligence in education. *International Journal for Educational Integrity, 19*(1), 1–4. https://doi.org/10.1007/s40979-023-00133-4.

Fröhling, L., & Zubiaga, A. (2021). Feature-based detection of automated language models: Tackling GPT-2, GPT-3 and Grover. *PeerJ Computer Science*, *7*, e443. https://doi.org/10.7717/peerj-cs.443.

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

GPTZero. (n.d.-a). *GPTZero FAQ*. Retrieved 28 (May 2023). from https://app.gptzero.me/app/faq.

GPTZero. (n.d.-b). *Home*. GPTZero. Retrieved 28 (May 2023). from https://gptzero.me/.

Guerrero-Dib, J. G., Portales, L., & Heredia-Escorza, Y. (2020). Impact of academic integrity on workplace ethical behaviour. *International Journal for Educational Integrity*, *16*(1), https://doi.org/10.1007/s40979-020-0051-3.

Gunser, V. E., Gottschling, S., Brucker, B., Richter, S., & Gerjets, P. (2021). Can users distinguish narrative texts written by an artificial intelligence writing tool from purely human text? *International Conference on Human-Computer Interaction*, 520–527. https://doi.org/10.1007/978-3-030-78635-9_67

Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020). *Automatic detection of generated text is easiest when humans are fooled* (arXiv:1911.00650). arXiv. https://doi.org/10.48550/arXiv.1911.00650

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274.

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*, 106553. https://doi.org/10.1016/j.chb.2020.106553.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). *A watermark for large language models* (arXiv:2301.10226). arXiv. http://arxiv.org/abs/2301.10226

Koubaa, A. (2023). *GPT-4 vs. GPT-3.5: A concise showdown*. TechRxiv. https://doi.org/10.36227/techrxiv.22312330.v2.

Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). *Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense* (arXiv:2303.13408). arXiv. https://doi.org/10.48550/arXiv.2303.13408.

Kumar, R. (2023). Faculty members' use of artificial intelligence to grade student papers: A case of implications. *International Journal for Educational Integrity, 19*(1), 9. https://doi.org/10.1007/s40979-023-00130-7.

Kumar, R., Mindzak, M., Eaton, S. E., & Morrison, R. (2022). *AI & AI: Exploring the contemporary intersections of artificial intelligence and academic integrity*. Canadian Society for the Study of Higher Education Annual Conference, Online. Werklund School of Education. https://tinyurl.com/ycknz8fd.

Lancaster, T. (2023). Artificial intelligence, text generation tools and ChatGPT – does digital watermarking offer a solution? *International Journal for Educational Integrity*, *19*(1), https://doi.org/10.1007/s40979-023-00131-6.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). *GPT detectors are biased against non-native English writers* (arXiv:2304.02819). arXiv. http://arxiv.org/abs/2304.02819.

Malinka, K., Perešíni, M., Firc, A., Hujňák, O., & Januš, F. (2023). *On the educational impact of ChatGPT: Is artificial intelligence ready to obtain a university degree?* (arXiv:2303.11146). arXiv. https://doi.org/10.48550/arXiv.2303.11146

Marche, S. (2022, December 6). *The college essay is dead*. The Atlantic. https://www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-writing-college-student-essays/672371/.

Microsoft (2023). *Confirmed: The new Bing runs on OpenAI's GPT-4 | Bing Search Blog*. https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI%E2%80%99s-GPT-4.

Netus AI. (n.d.). *NetusAI Paraphrasing Tool | Undetectable AI Paraphraser*. Netus AI Paraphrasing Tool. Retrieved 28 (May 2023). from https://netus.ai/.

Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, *2*, 100033. https://doi.org/10.1016/j.caeai.2021.100033.

OpenAI (2023a). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. http://arxiv.org/abs/2303.08774.

OpenAI (2023b, January 31). *New AI classifier for indicating AI-written text*. https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.

Originality.AI. (2023, April 9). *AI Content Detection Accuracy – GPTZero vs Writer vs Open AI vs CopyLeaks vs Originality.AI – Detecting Chat GPT AI Content Accuracy—Originality.AI*. https://originality.ai/ai-content-detection-accuracy/.

Perkins, M. (2023). Academic Integrity considerations of AI large Language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching & Learning Practice*, *20*(2), https://doi.org/10.53761/1.20.02.07.

Perkins, M., Gezgin, U. B., & Roe, J. (2018). Understanding the relationship between Language ability and plagiarism in non-native English speaking business students. *Journal of Academic Ethics*, *16*(4), https://doi.org/10.1007/s10805-018-9311-8.

Perkins, M., Gezgin, U. B., & Roe, J. (2020). Reducing plagiarism through academic Misconduct education. *International Journal for Educational Integrity*, *16*(1), 3. https://doi.org/10.1007/s40979-020-00052-8.

Perkins, M., & Roe, J. (2023). Decoding Academic Integrity policies: A Corpus Linguistics Investigation of AI and other Technological threats. *Higher Education Policy*. https://doi.org/10.1057/s41307-023-00323-2.

Pichai, S. (2023, February 6). *An important next step on our AI journey*. Google. https://blog.google/technology/ai/bard-google-ai-search-updates/.

Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for Education and Research: Opportunities, threats, and strategies. *Applied Sciences*, *13*(9), https://doi.org/10.3390/app13095783. Article 9.

Reimers, F., Schleicher, A., Saavedra, J., & Tuominen, S. (2020). *Supporting the continuation of teaching and learning during the COVID-19 Pandemic* (pp. 1–38). OECD. https://globaled.gse.harvard.edu/files/geii/files/supporting_the_continuation_of_teaching.pdf.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002.

Rodgers, C. M., Ellingson, S. R., & Chatterjee, P. (2023). Open Data and transparency in artificial intelligence and machine learning: A new era of research. *F1000Research*, *12*, 387. https://doi.org/10.12688/f1000research.133019.1.

Roe, J. (2022). Reconceptualizing academic dishonesty as a struggle for intersubjective recognition: A new theoretical model. *Humanities and Social Sciences Communications*, *9*(1). https://doi.org/10.1057/s41599-022-01182-9

Roe, J., & Perkins, M. (2022). What are automated paraphrasing tools and how do we address them? A review of a growing threat to academic integrity. *International Journal for Educational Integrity*, *18*(1). https://doi.org/10.1007/s40979-022-00109-w

Roe, J., Renandya, W., & Jacobs, G. (2023). A review of AI-Powered writing tools and their implications for Academic Integrity in the Language Classroom. *Journal of English and Applied Linguistics*, *2*(1). https://doi.org/10.59588/2961-3094.1035

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6*(1). https://doi.org/10.37074/jalt.2023.6.1.9

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). *Can AI-generated text be reliably detected?* (arXiv:2303.11156). arXiv. https://doi.org/10.48550/arXiv.2303.11156

Sohail, S. S., Madsen, D., Himeur, Y., & Ashraf, M. (2023). *Using ChatGPT to navigate ambivalent and contradictory research findings on artificial intelligence* (SSRN Scholarly Paper 4413913). https://doi.org/10.2139/ssrn.4413913

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., & Wang, J. (2019). *Release strategies and the social impacts of language models* (arXiv:1908.09203). arXiv. https://doi.org/10.48550/arXiv.1908.09203

Sparrow, J. (2022, November 18). 'Full-on robot writing': The artificial intelligence challenge facing universities. *The Guardian*. https://www.theguardian.com/australia-news/2022/nov/19/full-on-robot-writing-the-artificial-intelligence-challenge-facing-universities.

Strzelecki, A. (2023). To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. *Interactive Learning Environments*, *0*(0), 1–14. https://doi.org/10.1080/10494820.2023.2209881.

Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning and Teaching*, *6*(1).

Turnitin.com. (2021, January 21). *A new path and purpose for Turnitin*. https://www.turnitin.com/blog/a-new-path-and-purpose-for-turnitin

Turnitin.com. (2023, April 4). *The launch of Turnitin's AI writing detector and the road ahead*. https://www.turnitin.com/blog/the-launch-of-turnitins-ai-writing-detector-and-the-road-ahead.

Turnitin.com. (n.d.-a). *AI writing detection frequently asked questions*. Retrieved 28 (May 2023). from https://www.turnitin.com/products/features/ai-writing-detection/faq

Turnitin.com. (n.d.-b). *Turnitin for universities*. Retrieved 16 (August 2023). from https://www.turnitin.com/regions/uk/university

Uzun, L. (2023). ChatGPT and Academic Integrity concerns: Detecting Artificial Intelligence Generated Content. *Language Education and Technology*, *3*(1), Article 1. http://www.langedutech.com/letjournal/index.php/let/article/view/49.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). *Testing of detection tools for AI-generated text* (arXiv:2306.15666). arXiv. https://doi.org/10.48550/arXiv.2306.15666

Zhang, S. J., Florin, S., Lee, A. N., Niknafs, E., Marginean, A., Wang, A., Tyser, K., Chin, Z., Hicke, Y., Singh, N., Udell, M., Kim, Y., Buonassisi, T., Solar-Lezama, A., & Drori, I. (2023). *Exploring the MIT mathematics and EECS curriculum using large language models* (arXiv:2306.08997). arXiv. https://doi.org/10.48550/arXiv.2306.08997

Authors and Affiliations

Mike Perkins[1] · Jasper Roe[2] · Darius Postma[3] · James McGaughran[1] · Don Hickerson[1]

✉ Mike Perkins
mgperkins@gmail.com

Jasper Roe
jasper.roe@jcu.edu.au

Darius Postma
Darius.p@buv.edu.vn

James McGaughran
James.mg@buv.edu.vn

Don Hickerson
Don.h@buv.edu.vn

[1] British University Vietnam (School of Business), Hanoi, Vietnam

[2] James Cook University Singapore, Singapore, Singapore

[3] British University Vietnam (School of Hospitality and Tourism), Hanoi, Vietnam