ŌTAUTAHI CHRISTCHURCH AOTEAROA NEW ZEALAND
35TH AUSTRALASIAN ASSOCIATION FOR ENGINEERING EDUCATION ANNUAL CONFERENCE
8-11 DEC 2024
THE ENGINEER AND THE WORLD

Ethical AI Hacking Uncovers Vulnerabilities in Engineering Assessments

Swapneel Thite[a], Khalegh Barati[a], Morgan Harris[a], Fiacre Rougieux[a], Giordana Orsini Florez[a], Ghislain Bournival[a], Benjamin Phipps[b] and Sasha Vassar[a].
Faculty of Engineering, UNSW, Sydney, Australia[a], Academic Development, UNSW Sydney, Australia[b]
Corresponding Author Email: s.thite@unsw.edu.au

CONTEXT

Assessment design plays a vital role in shaping students' motivation to learn and ensuring the achievement of learning outcomes. There is a growing concern surrounding AI misuse among higher education academics. Current methods of AI vulnerability detection are not able to evaluate the levels of assessment vulnerability and inform assessment design.

PURPOSE

This research aims to develop a methodological framework and AI vulnerability detection scale to assist educators in detecting levels of vulnerabilities to AI misuse and inform assessment design.

METHODOLOGY

The methodology adopted an ethical hacking approach from cybersecurity into engineering assessments. Using the developed framework, vulnerability testing in one engineering assessment was conducted by blind marking a mix of AI created submission using four different AI platforms and real student submissions. Several levels of evidence were gathered during analysis to diagnose potential vulnerabilities, including marker detection of AI submissions, performance of AI against rubric criteria, and comparison with student performance.

OUTCOMES

An ethical AI hacking assessment vulnerability detection framework and assessment vulnerability scale were created. Application of the framework revealed that most AI platforms performed lower compared to students. However, one platform was statistically similar to students and all platforms could obtain a passing grade of all the course learning outcomes. In addition, it was found that most markers were able to discern an AI submission and a student one. Thus, the process of ethical hacking undertaken in this study show that knowing whether AI platforms can pass an assessment is insufficient to inform the redesign of assessment and course learning outcomes. The recommendations made to the pilot course convenor following the diagnosis were to shift from a product to a process-based learning by integrating identified vulnerable learning outcomes into existing workshops for immediate feedback, reducing reliance on a single assessment to assure student learning.

CONCLUSIONS

An AI vulnerability detection scale and Ethical AI hacking vulnerability detection framework were developed, which outlined the process of conducting ethical AI hacking to test levels of vulnerabilities in engineering assessments. This detection scale and framework analysed one engineering course. The insights gained from ethical hacking can be used to inform better assessment design using assessment for learning principles and hence reduce the risk of AI misuse in engineering assessments and assure student learning.

KEYWORDS
AI Assessment vulnerabilities; AI Assessment design; Ethical hacking; Assurance of learning

Introduction

Assessment design plays a crucial role in influencing students' motivation and their propensity to cheat (Sutherland-Smith & Dawson, 2022). In the age of artificial intelligence (AI), thoughtfully designed assessments can serve as a foundational tool to enhance assurance of learning. However, there are growing concerns regarding academic integrity surrounding AI. To address these concerns and ensure assurance of learning, it is essential to identify assessment vulnerabilities to AI misuse.

Previous research has explored assessment vulnerabilities by evaluating AI's ability to pass engineering assessments (Nikolic et al., 2023), there are still gaps in understanding and evaluating the disparities between student and AI-generated submissions. Moreover, AI detection tools like Turnitin are not fully reliable and often produce false positives (Weber-Wulff et al., 2023). These tools also fail to offer insights into the weaknesses in assessment design, which are crucial for improving the assessments themselves. Therefore, relying solely on AI detection tools is insufficient and additional analysis is needed to uncover vulnerabilities in assessment design. By identifying vulnerabilities in assessment design, academics can develop more robust assessment methods that not only deter AI misuse but also assure student learning. To address this need, this paper proposes an alternative methodology to identify vulnerabilities in assessment design to AI misuse named Ethical hacking. Ethical hacking is a methodology used in the field of cybersecurity, defined as a process employing hacking techniques to identify and remediate system vulnerabilities, which is critical in defending against malicious attacks (Raman et al., 2024). This approach and methodology can be potentially employed to increase the security of assessments within higher education (Dawson, 2020). Therefore, the research question investigated in this paper is "*How can we employ the ethical hacking methodology in engineering education to identify vulnerabilities in engineering assessment design to reduce AI misuse?*" This paper provides an "AI vulnerability detection scale" and "Ethical AI hacking vulnerability detection framework" for approaching ethical hacking of university assessments. The framework and scale were tested in an engineering course and used as a diagnostic tool for determining AI assessment vulnerability.

Literature Review

The digital transformation, spurred by the release of OpenAI's ChatGPT model in November 2022 has resulted in a disruption to engineering education, and more broadly the education landscape. Tools like ChatGPT, based on transformer architecture models, leverage the structured nature of human language to generate text that closely mimics human writing using advanced natural language processing (NL) models. While generative AI tools have the potential to support student learning, there are still concerns about ethical and effective use of this technology (Nikolic et al., 2023). Two broad approaches to addressing AI and academic integrity have emerged in the literature (Moorhouse et al., 2023). The first, a proactive strategy, focuses on designing assessments that are less susceptible to AI-assisted cheating (e.g. oral assessments and interviews). The second, a reactive approach, involves detecting AI-generated content. These strategies are complementary, not mutually exclusive. Regarding the former, it is advisable to try to probe assessment tasks for potential holes in their security using an ethical hacking method from the field of cybersecurity (Dawson, 2020).

Nikolic et al. (2023) tested a range of engineering assessments. The focus of their study was to know if ChatGPT 4.0 could pass different assessment tasks. A pass or fail approach was used as much as possible (a grade was calculated when a single answer was the output, e.g., multiple-choice question). This was to minimise the bias of knowing that the work had been AI-generated. The research team was both generating the output through AI and marking it and were free to modify their prompts to try to achieve a passing output whilst marking. This ensures that the best

quality output, with respect to the marking criteria, is generated. Other studies have explored vulnerability detection using blind marking. In a computer science course, markers received 10 student scripts and 5 ChatGPT-generated scripts (Richards et al., 2024). The AI-generated scripts were compiled from ChatGPT 4.0 where obvious AI artifacts were removed, and the scripts were properly structured. The 15 anonymous scripts that each marker received were randomised and checked for plagiarism (but not AI detection) as part of the evaluation process. Markers were not aware of the experiment but were instructed to flag anything unusual. Additionally, Chaudhry et al. (2023) compared the grade of AI-generated submissions with that of students to gauge AI's capabilities and test assessment vulnerability. Similarly, Scarfe et al. (2024) conducted a Turing Test to determine whether humans could distinguish between AI-generated and human-produced work, focusing on testing the integrity of university examinations. The markers were unaware of the AI involvement, and ethics approval was not sought, as the study was classified as a quality assurance exercise within the university. However, markers were instructed to flag anything unusual. This blind marking does help remove biases. However, it has been suggested that a more ethical approach would be to make markers aware that some assignments are AI-generated (Yeadon et al., 2023). Albeit they may not know which assessment tasks have been AI-generated. Furthermore, this study primarily examined the markers' interpretations of whether submissions were human or AI-generated, supported by histograms comparing the marks of AI and student submissions. While detecting vulnerabilities is a valuable first step, a deeper evaluation of the results is necessary to inform improved assessment design to reduce AI misuse and assure student learning.

Assessment AI Hacking Experiment

The methodology employed in this ethical Assessment AI Hacking Experiment was rooted in the penetration testing or ethical hacking approach within the cybersecurity domain (Raman et al., 2024). This method has been recognized for its potential application in educational contexts to enhance assessment security (Dawson, 2020). The steps taken in this experiment are discussed in detail in the sections below.

Courses and Assessments Selection

Upon receiving the ethics approval, a recruitment advert was distributed across the faculty of engineering for academics to participate in this research project. The inclusion criteria for participation in this study were for academics: (i) to have concern of AI misuse within their courses and/or (ii) to wish to develop AI relevant assessments and/or (iii) to wish to test their assessments for any vulnerabilities to AI. A total of four pilot courses expressed interest to participate, out of which one pilot course was selected. Although the pilot course in this study also involved group submissions, teaching assistants were not assigned to specific groups. Furthermore, blind marking was manually set up. From consultations with the course conveners, written form assessments were selected as the pilot conveners deemed them to be more vulnerable to generative AI misuse. Therefore, this study focused on a written technical report in a post-graduate engineering course. There was a total of 105 submissions, with 57 student group submissions and 48 AI submissions, out of which 12 submissions were made for each of the four AI platforms considered for this study.

Course Conveners' Consultation

Course convenor consultations were conducted to gather information and resources regarding the assessments criteria to generate appropriate AI submissions and evaluate the data in a manner that informed assessment design. Course convenors were asked to provide the course learning outcomes (CLO), assessment instructions and weightage of each CLO in relation to the criterion assessed. The number of markers and the process of setting up blind marking and the deidentification process was also discussed. The assessment instructions for the group technical report required students to think critically and evaluate data of energy consumption of their own homes.

AI Platform Selection

Four AI platforms were selected for this experiment. ChatGPT, Gemini and Copilot were selected based on their general availability and popularity among students (Đerić et al., 2024). Claude was also selected as it was (at the time of the experiment) ranked second on the LMSYS Chatbot Arena leaderboard (Chiang & Angelopoulos, 2024), only appearing below GPT 4o. This provides a reasonable cross-section of platforms in current use as well as the bleeding edge of AI innovation. The specific models being used were GPT-4o-2024-05-13 for ChatGPT, Gemini-Advanced-0514 for Gemini, and Claude-3.5-Sonnet for Claude. Microsoft does not release model details for Copilot. For this experiment, ChatGPT and Claude were both provided with the full contents of all supplied course resources in PDF format. Both of these platforms support persistent indexed document stores, which can be used for retrieval-augmented generation (Lewis et al., 2020). Gemini and Copilot only received the assessment task instructions and marking rubric as PDF files.

AI-Powered Human Operators

While it may be that future AI models are sufficiently powerful to directly assume the role of a student in a course; at present, they require human operators. Therefore, casual staff were recruited to perform this task. The AI operators were each required to follow specific instructions to ensure that no domain-specific knowledge could influence the resulting assessment submission. These instructions were based on a similar standard system message and initial prompt but allowed for some tailoring for individual models to achieve consistent and useful results. The operators were given training guidelines based on these general principles: (i) do not explain the concept of the course or the assessment to the AI platform beyond the instructions that are provided with the assessment task itself; (ii) do not re-word assessment instructions; (iii) even when asking the AI to complete the assessment in parts, always give the instructions for each part exactly as written; (iv) do not read the AI output in detail; (v) editing to conform to expected structure (for example, fixing headings) is permissible; and (vi) where the AI generates data in the form of a table, chart or diagram, use the indicated tool to generate that content. In addition, to ensure that markers could not distinguish genuine submissions from AI-generated submissions, submissions were provided to markers with all identifying marks removed, including names, groups and title pages. Files were then assigned a random number and emailed to markers to ensure anonymity.

Result Analysis

The analysis of the results from the group technical report was conducted using “R”, a programming language for statistical computing and data visualizations. It entailed comparing the marks and feedback of students’ assessments with those generated by AI, and how markers assessed each of these. A component of the analysis involved comparing the scores for each criterion in the rubric. This was performed by conducting a two tailed t-test between the student group and the AI group. An arc-sin of the square-root of the percentage (as a fraction) transformation was applied before conducting the t-test to remove heterogeneity of variance (Underwood, 1996). To gain more insights into the performance of each of the groups, a test result analysis was performed using the approach of TU Delft (Harting et al., 2023). This included Cronbach’s alpha value, the item-rest correlation, and the correlation coefficient. Furthermore, the score of the students was also assessed in terms of learning outcomes by converting the marking criterion-based marks to learning outcome-based marks for each of the sections of the report using the weightage provided by the course conveners to observe to what level the submissions have attained the learning outcomes. Finally, the written feedback given by the markers was analysed using text analysis. Word clouds and word correlation plots were used to get insights on the most common mistakes.

Results and Discussion

Assessment vulnerability tests offer the potential to help academics better design and use assessments that assure learning. However, this ethical hacking process requires careful planning and guidance. Therefore, this study introduces the "Ethical AI Hacking Assessment Vulnerability Detection Decisions Framework" (Figure 1**Error! Reference source not found.**) to support academics intending to undertake vulnerability detection in an ethical manner.

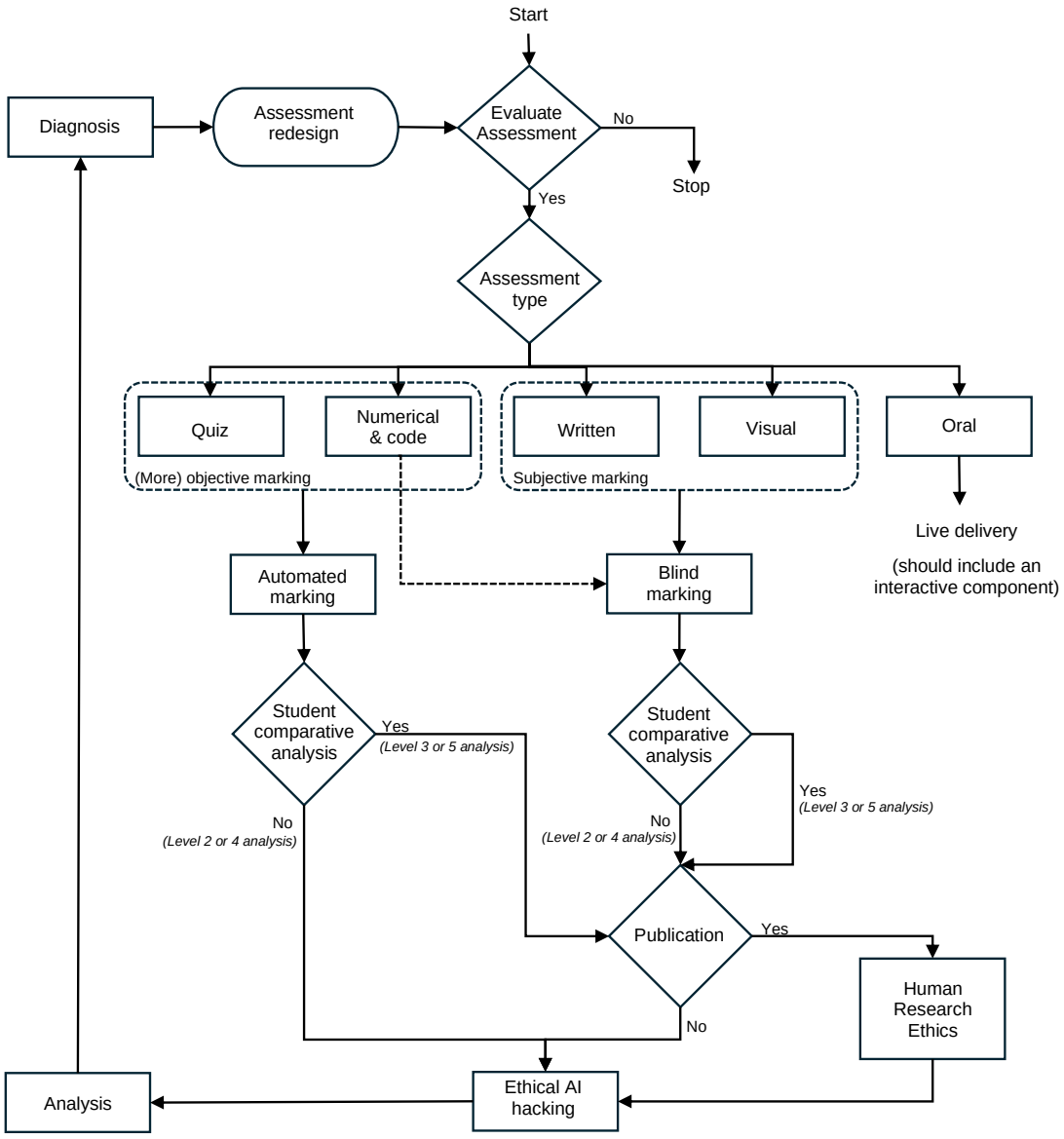


Figure 1. Flow diagram to evaluate assessment vulnerabilities using an "Ethical AI Hacking Assessment Vulnerability Detection Decisions Framework".

The framework begins by assessing an intention to evaluate the assessment. If this is the case, it is necessary to understand the best approach to undertake for each assessment type and how marking is done for each of these. The assessment types within this framework are loosely based on those outlined by Nikolic et al. (2023). Like Nikolic et al. (2023), this framework excludes oral assessments since AI cannot deliver presentations on behalf of students. Blind marking is recommended when marking cannot be automated. A comparative analysis can be used depending on the intent of the project. Five-level AI vulnerability detection scale (**Error!**

Reference source not found.) is suggested to determine if AI should be benchmarked against students. Following this, determining whether this vulnerability testing will be used for quality assurance purposes or publication is necessary to determine whether ethics approval is required. Once this is defined, ethical hacking can then be performed, as well as the analysis of the results using the five-level AI vulnerability detection scale. This analysis will lead to diagnosing how vulnerable the assessment is and will reveal specific aspects of an assessment that need redesign. Educators can focus on redesigning the assessment or exploring alternative methods to assure learning, while considering the specific constraints and opportunities within the course.

The evaluation operates across five levels as reported in **Error! Reference source not found.**. The first involves visual detection by markers or by detection software. The second examines the percentage of AI submissions that pass the assessment. The third level includes a comparison with students. These first three levels are established benchmarks from previous studies (Chaudhry et al., 2023; Nikolic et al., 2023; Scarfe et al., 2024). Uniquely, this study introduces additional types of analysis for level 3. Finally, level 4 evaluates how well AI submissions meet the learning

Table 1. Levels of evidence used in the analysis to diagnose vulnerabilities in assessment design.

Evidence	Description	Analysis	Diagnosis (Interpretation of analysis)
Level 1	Detection of AI submission by AI detection software and/or marker	• Potential AI usage percentage[1] • Visual detection	• Determine potential of AI usage by students
Level 2	AI submissions pass the assessment	• Testing whether the overall mark achieves a passing mark	• Determine assessment vulnerability by using the overall passing mark as a benchmark
Level 3	Comparison of AI performance to student performance	• Analyse the performance of AI compared with students using a t-test. • Feedback analysis	• Level 2 diagnostic • Incentive for students to use AI if AI can do better than a student
Level 4	Level 2 process applied to the Course Learning Outcomes (CLOs) and/or rubric marking criteria within the assessment	• Level 2 applied to each CLO and/or marking criterion • Correlation matrix (correlation coefficient between CLOs/rubric criteria) • Item-rest correlation • Cronbach's alpha value • Feedback analysis	• Level of vulnerability for (1) the CLO's the assessment is intended to measure and (2) whether AI could pass each rubric criterion in an assessment. • Statistical reliability and validity of the assessment or part thereof • Aspects of the assessment need to be redesigned or removed
Level 5	Level 3 process applied to the CLOs and/or rubric marking criteria within the assessment	• Level 3 (applied to each CLO and/or marking criterion) • Level 4	• Level of vulnerability across the different levels of achievement for each CLO and rubric criterion • Determine assessment redesign needed to ensure students achievement levels of learning outcomes are accurately represented in the assessment marks

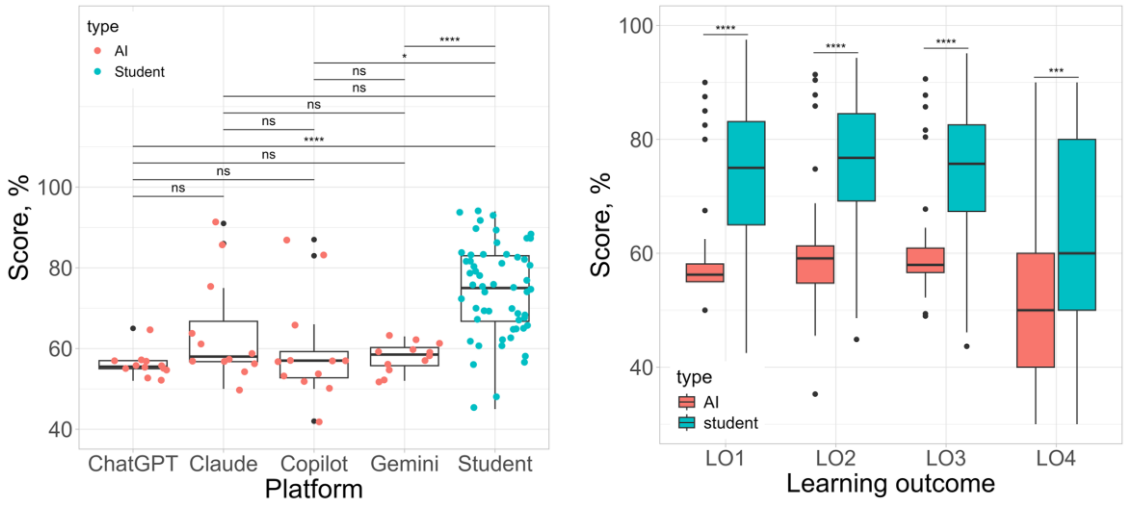
[1] Student submission cannot be assumed to be student-generated since some may have used AI.

outcomes and/or rubric criteria. This may include benchmarking against student submissions (level 5). Levels 3, 4 and 5 can better inform specific aspects of assessment and course design to improve assurance of learning and secure assessments to Gen AI.

Following the framework, level 1 and 5 evidence were considered for this pilot. Only a few level 5 evidence can be presented. Overall, blind markers showed good ability to evaluate whether the submissions were AI or not with only 4 out of 105 submissions not identified by markers as generated by AI. An interesting observation is that the misclassified 4 AI submissions all received high marks above 85%. This study highlights the effectiveness of an ethical hacking approach as compared to previous studies (Nikolic et al., 2023) in reducing biases and more accurately simulating a real-world scenario where AI submissions are created by students, and teaching assistants mark the submissions.

The results from the level 5 analysis presented intriguing observations. The Cronbach's alpha for the student and AI marks were 0.88 and 0.84, respectively. These high alpha values suggest minimal measurement error, meaning that the marks accurately reflect the true performance of the respective submissions. The minimum item-rest correlation for the marks obtained by student and AI submissions for the learning outcome related rubric criteria was 0.66 and 0.63, respectively, indicating that each marking criterion was able to distinguish low and high performing submissions.

Another important finding of this study was that only one AI submission failed the entire assessment, and all four AI platforms tested performed similarly, as in most cases, no statistically significant difference was found between these four platforms using t-tests ($p<0.05$) (Figure 2a). However, there is a moderate statistically significant difference between student and Copilot marks and no statistically significant difference between student's and Claude marks. Overall, these results indicate a high assessment vulnerability to AI, as all but one AI submission passed, and, statistically, Claude could perform as well as a student in this experiment. These findings are consistent with the results obtained within a test conducted within a final examination in a mechanical engineering course, where the AI platforms such as ChatGPT4o and Claude performed better than engineering academic staff (Tian et al., 2024).



(a) (b)

Figure 2. Boxplot of assessment scores for students and AI submissions of a technical report assessment based (a) on the platform used, and (b) reported for the different learning outcomes.

In addition, the distribution of the scores reported as a percentage for each learning outcome showed statistically significant differences between students and AI against each learning outcome (Figure 2b). Overall, AI submissions obtained more than 50% of the learning outcomes

and sometimes AI scored better than most of the students for this assessment, particularly for learning outcome 4 on presentation and formatting, which were mediated by human operators, indicating high vulnerability and potential issues with assurance of learning based on evaluative judgement. Similarly, Nikolic et al. (2024) conducted AI vulnerability tests across platforms like ChatGPT4, Copilot, and Gemini, finding that most written assessments were successfully passed by AI-generated submissions, further suggesting that learning outcomes in these assessments are not reliably assured. Therefore, it can be argued that measuring many learning outcomes in a single assessment can compromise assurance of learning if the assessment becomes susceptible to AI-generated work. Vulnerability tests where AI and student marks are compared based on rubric criteria and learning outcomes distributions can help inform educators of which rubric criteria and learning outcomes are most vulnerable within their assessments. In addition, a correlational analysis of the student and AI marks based on rubric criteria could help provide deeper insights into the elements of assessment design which are vulnerable to AI. Educators can then make decisions towards how those elements need to be assured in other areas of assessment and improve assurance of learning in any single assessment by distributing critical learning outcomes across diverse assessment types. This multi-faceted approach would strengthen the reliability of assessments in measuring authentic student performance, ensuring that learning objectives are met even if some tasks are more vulnerable to AI.

After a review of the pilot course, it was found that all four learning outcomes in the assessment could be achieved through AI submissions and were equally vulnerable. As a result, it was recommended to reduce reliance on the final product and shift towards process-based learning, incorporating assessment for learning principles. The team identified opportunities to embed learning outcomes within the course and suggested integrating small activities into existing workshops or tutorials. This would allow teaching assistants to provide immediate feedback, reducing the dependence on technical reports for assessing learning. Additionally, a comparison of the correlational indices between student and AI rubric scores showed that students outperformed AI in visual data representation, which impacted their scores in the methodology, recommendations, and limitations criteria. Consequently, it was recommended to include diverse visual data representation in the assessment instructions.

The ethical hacking approach in this study presents a few limitations. A major challenge is the need for ethics approval if the results are to be published. Additionally, setting up blind marking manually and using human operators to generate AI submissions adds layers of administrative complexity and funding requirements. As AI technologies continue to evolve rapidly, this methodology will need to be streamlined and updated to support regular, periodic assessments that can address emerging vulnerabilities efficiently. To address these limitations, future work should aim to integrate this methodology into the university's quality assurance framework, potentially reducing or even eliminating the need for repeated ethics approvals. Automating the blind marking process would further reduce administrative and financial burdens and increase efficiency. More importantly, mitigating the risks of AI misuse in assessments will require innovative solutions that assure learning and involve students in their learning process. Future work could focus on providing educators with support for redesigning assessments or ideating alternative assessments by leveraging AI as a learning tool, adopting programmatic assessment strategies, and shifting from product-based to process-driven approaches or redistributing learning outcomes across a diverse array of assessments within a course to safeguard both assessment integrity and the quality of learning outcomes.

Conclusion

This study addressed the research question, “How can we employ the ethical hacking methodology in engineering education to identify vulnerabilities in engineering assessment design to reduce AI misuse?” An “Ethical AI Hacking Assessment Vulnerability Detection Framework” and an "AI Vulnerability Detection Scale” were developed which are used to detect the level of assessment vulnerabilities and inform assessment redesign. The framework was tested through an AI hacking experiment, revealing that the technical group report tested was

vulnerable to AI. Notably, one AI platform performed as well as a student in the experiment. A key limitation of this approach is the need for ethics approval and the administrative complexity of manually setting up blind marking. Future work should focus on automating these processes to reduce administrative burdens and explore innovative solutions for redesigning assessments to further mitigate AI misuse risks and assure student learning.

References

Chaudhry, I. S., Sarwary, S. A. M., El Refae, G. A., & Chabchoub, H. (2023). Time to revisit existing student's performance evaluation approach in higher education sector in a new era of ChatGPT - A case study. *Cogent Education*, *10*(1), 2210461. https://doi.org/https://doi.org/10.1080/2331186X.2023.2210461

Chiang, W.-L., & Angelopoulos, A. (2024). *Chatbot Arena LLM leaderboard: Community-driven evaluation for best LLM and AI chatbots*. Retrieved 1 July from https://lmarena.ai/

Dawson, P. (2020). *Defending assessment security in a digital world: Preventing e-cheating and supporting academic integrity in higher education*. Routledge. https://doi.org/https://doi.org/10.4324/9780429324178

Đerić, E., Frank, D., & Malenica, M. (2024, 20-24 May). *Comparison and quatification of GAI tools use among different academic population segments* 47th MIPRO ICT and Electronics Convention (MIPRO), Opatija, Croatia.

Harting, L., Sies, P., van Bergen Bravenboer, B., & Blok, J. (2023). *TU Delft assessment manual: Teaching and learning services*. https://www.tudelft.nl/en/teaching-support/didactics/assess/guidelines/tu-delft-assessment-manual

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020, 6-12 December). *Retrieval-augmented generation for knowledge-intensive NLP tasks* NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada.

Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open*, *5*, 100151. https://doi.org/https://doi.org/10.1016/j.caeo.2023.100151

Nikolic, S., Daniel, S., Haque, R., Belkina, M., Hassan, G. M., Grundy, S., Lyden, S., Neal, P., & Sandison, C. (2023). ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. *European Journal of Engineering Education*, *48*(4), 559-614. https://doi.org/https://doi.org/10.1080/03043797.2023.2213169

Nikolic, S., Sandison, C., Haque, R., Daniel, S., Grundy, S., Belkina, M., Lyden, S., Hassan, G. M., & Neal, P. (2024). ChatGPT, Copilot, Gemini, SciSpace and Wolfram versus higher education assessments: an updated multi-institutional study of the academic integrity impacts of Generative Artificial Intelligence (GenAI) on assessment, teaching and learning in engineering. *Australasian Journal of Engineering Education*, 1-28. https://doi.org/https://doi.org/10.1080/22054952.2024.2372154

Raman, R., Calyam, P., & Krishnashree, A. (2024). ChatGPT or bard: Who is a better certified ethical hacker? *Computers & Security*, *140*, 103804. https://doi.org/https://doi.org/10.1016/j.cose.2024.103804

Richards, M., Waugh, K., Slaymaker, M., Petre, M., Woodthorpe, J., & Gooch, D. (2024). Bob or Bot: Exploring ChatGPT's answers to university computer science assessment. *ACM Transactions on Computing Education*, *24*(1), 1-32. https://doi.org/https://doi.org/10.1145/3633287

Scarfe, P., Watcham, K., Clarke, A., & Roesch, E. (2024). A real-world test of artificial intelligence infiltration of a university examinations system: A “Turing Test” case study. *PLoS ONE*, *19*(6), e0305354. https://doi.org/https://doi.org/10.1371/journal.pone.0305354

Sutherland-Smith, W., & Dawson, P. (2022). Higher education assessment design. In S. E. Eaton, G. J. Curtis, B. M. Stoesz, J. Clare, K. Rundle, & J. Seeland (Eds.), *Contract cheating in higher education: Global perspectives on theory, practice, and policy* (pp. 91-105). Palgrave Macmillan. https://doi.org/https://doi.org/10.1007/978-3-031-12680-2

Tian, J., Hou, J., Wu, Z., Shu, P., Liu, Z., Xiang, Y., Gu, B., Filla, N., Li, Y., Liu, N., Chen, X., Tang, K., Liu, T., & Wang, X. (2024). Asessing large language models in mechanical engineering education: A study on mechanics-focused conceptual understanding. https://doi.org/https://doi.org/10.48550/arXiv.2401.12983

Underwood, A. J. (1996). *Experiments in ecology: Their logical design and interpretation using analysis of variance*. Cambridge University Press. https://doi.org/https://doi.org/10.1017/CBO9780511806407

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. *International Journal for Education Integrity*, *19*(26). https://doi.org/https://doi.org/10.1007/s40979-023-00146-z

Yeadon, W., Inyang, O.-O., Mizouri, A., Peach, A., & Testrow, C. P. (2023). The death of the short-form physics essay in the coming AI revolution. *Physics Education*, *58*(3), 035027. https://doi.org/https://doi.org/10.1088/1361-6552/acc5cf

Copyright statement