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

Sasha Nikolic[a], Carolyn Sandison[a], Rezwanul Haque[b], Scott Daniel[c], Sarah Grundy[d], Marina Belkina[e], Sarah Lyden[f], Ghulam M. Hassan[g], and Peter Neal[d]

[a]University of Wollongong, [b]University of the Sunshine Coast, [c]University of Technology Sydney, [d]University of New South Wales, [e]Western Sydney University, [f]University of Tasmania, [g]University of Western Australia,

sasha@uow.edu.au[a]

Keywords: Assessment, Academic Integrity, ChatGPT, Cheating, Education, Generative Artificial Intelligence (GenAI)

Abstract:

More than a year has passed since reports of ChatGPT-3.5's capability to pass exams sent shockwaves through education circles. These initial concerns led to a multi-institutional and multi-disciplinary study to assess the performance of Generative Artificial Intelligence (GenAI) against assessment tasks used across ten engineering subjects, showcasing the capability of GenAI. Assessment types included online quiz, numerical, oral, visual, programming and writing (experimentation, project, reflection and critical thinking, and research). Twelve months later, the study was repeated using new and updated tools ChatGPT-4, Copilot, Gemini, SciSpace and Wolfram. The updated study investigated the performance and capability differences, identifying the best tool for each assessment type. The findings show that increased performance and features can only heighten academic integrity concerns. While cheating concerns are central, opportunities to integrate GenAI to enhance teaching and learning are possible. While each GenAI tool had specific strengths and weaknesses, ChatGPT-4 was well-rounded. A GenAI Assessment Security and Opportunity Matrix is presented to provide the community practical guidance on managing assessment integrity risks and integration opportunities to enhance learning.

Introduction:

The release of ChatGPT 3 to the general public in November 2022 sent shockwaves through education institutions when its capability to disrupt traditional academic integrity safeguards and transform teaching and learning was realised (Bahroun et al., 2023). Generative Artificial Intelligence (GenAI) refers to artificial intelligence models that can create content, such as text, images, videos and music, similar to human-generated works (Chung et al., 2023). It has the capability to learn from vast amounts of data, enabling it to generate new, original outputs based on the patterns and information it has analysed (Crompton and Burke, 2024). GenAI capabilities are evolving quickly, and this study plays an important role in identifying the evolutionary changes across the last 12 months to the risks and opportunities associated with its use across higher education.

One key risk, front and centre in the minds of academics across the world, is that associated with academic integrity (Cotton et al., 2023). Academic integrity is the *expectation that teachers, students, researchers and all members of the academic community act with honesty, trust, fairness, respect and responsibility* (Tertiary Education Quality and Standards Agency, 2022). Breaching academic integrity is also known as 'academic misconduct' or 'academic dishonesty'. A key academic misconduct risk is cheating, which needs to be deterred and eliminated because it allows incompetent and dishonest professionals (those who have been wrongly accredited for learning outcomes they have not met) to work unsafely where they should not (Dawson, 2020). Within the field of engineering, incompetent and dishonest professionals can have a devasting impact on the community. For example, unsafe bridges and buildings that do not meet required standards can collapse (causing unnecessary deaths) or lead to financial ruin when rectified.

The risks to academic integrity from technological progress are not new and existed before ChatGPT emerged (Abd-Elaal et al., 2022, Rogerson and McCarthy, 2017, Dawson, 2016). In fact, in 2021, before the GenAI cheating headlines started, it was estimated that 7.9% of students purchased assignments from commercial contract cheating services, and 11.4% of students resorted to acquiring pre-written assignments from commercial file-sharing websites (Curtis et al., 2022). The identification of GenAI as a cheating risk is just evolution. Over time, the way students cheat changes as they react to the technologies available to them and the detection and deterrent policies and procedures put in place (Dawson, 2020). In terms of GenAI, this technological evolution can enable contract cheaters to work more efficiently, or students can or will soon be able to eliminate their need for such services (Nikolic et al., 2023a).

Students cheat for many reasons, a complex interconnection between teacher-related, institutional, internal and environmental factors, but a key motivator is if the reward outweighs the risks (Noorbehbahani et al., 2022). In less than a year, many students were already using GenAI for help with homework and at-home tests and quizzes (Elkhodr et al., 2023). In effect, this may come from cascading experiences where students discover that using GenAI is effective and undetected in minor tasks, providing attractiveness and confidence to use it for other activities (Purtill, 2023). This has resulted in much reflection on the types of assessments in use. For example, there are suggestions that the greater use of authentic assessment is needed to challenge the rise of GenAI, but there are signs that such approaches are not bullet-proof (Ajjawi et al., 2023).

While the focus on GenAI is predominantly on academic integrity risk, new opportunities are available. For example, integrating generative AI into engineering education can enhance teaching materials, create innovative learning spaces, lighten teacher workloads, and empower students to shape and personalise their educational journeys (Menekse, 2023). Currently, the academic community is working towards developing understanding and best practices to guide such integration (Elkhodr et al., 2023, Shanto et al., 2023).

The uncertainty of risks and opportunities associated with GenAI led this research team to undertake a comprehensive SWOT analysis in the first quarter of 2023 (Nikolic et al., 2023a). This study was both multi-disciplinary and multi-institutional, exploring the impact of ChatGPT-3.5 by pitting it up against assessment tasks across ten different subjects. In effect, the study provided guidance on academic integrity risks and assessment security options. Assessment security refers to the measures taken to secure assessments from cheating attempts (Dawson, 2020). If weaknesses of assessment delivery are understood, steps to strengthen security can be considered. The study by Nikolic et al. (2023a) concluded that assessment integrity risk was highly correlated to the assessment type and not necessarily the subject matter. It was found that against many assessment types, ChatGPT was rather strong at achieving passable grades but, in others, somewhat weak. Recommendations were made with short-term solutions to help protect integrity. However, the discussion outlined that the pace of change was immense and that ChatGPT-4 and plugins (such as Wolfram) could substantially impact performance. The suggestion was that short-term solutions would possibly last 12 – 24 months. Similar short-term recommendations were made by Raza and Hussain (2023). Therefore, the education community needs to understand the impact of technological advances on academic integrity. Since then, a range of other GenAI tools have gained much attention, including Google's product Gemini, which was recently upgraded and renamed from Bard, and Copilot from Microsoft. Amongst the diversity in choice, the education community needs guidance on identifying the most suitable GenAI for integration.

This study seeks to provide empirical evidence, 12 months from the previous benchmarking exercise, on the progress of popular generative artificial intelligence tools. Using the same ten subjects across seven Australian universities, this study will explore two research questions to address the stated limitations:

1. *Has the performance of generative artificial intelligence tools improved so that it can pass students across more assessment types?*
2. *Which of the major generative artificial intelligence tools are best suited to each assessment type?*

As the assessment types used are diverse across quantitative and qualitative methods, the results of this study can provide strong insights across higher education.

Literature Review

Generative Artificial Intelligence (GenAI) tools have evolved substantially since their inception. Chatbots, computer programs designed to communicate with humans, began with ELIZA in the 1960s (Adamopoulou and Moussiades, 2020). The next major evolution came with the integration of Natural Language Processing (NLP) and virtual personal assistants such as Siri and Alexa, which made these technologies more user-friendly and widespread by making the conversations feel more life-like (Adamopoulou and Moussiades, 2020). The development of the Generative Pre-trained Transformer (GPT) and Large Language Models (LLMs) marked an important milestone in the evolution of AI. Developed by OpenAI, GPT models utilise advanced deep learning to create content akin to human output, pre-trained on extensive text datasets for contextually accurate and coherent language generation (Hallal et al., 2023).

Starting with GPT-1 in 2018, which had been trained with a dataset of 117 million parameters, OpenAI's models have seen exponential growth. In November 2022, ChatGPT-3.5, trained with 175 billion parameters, sent shockwaves across the education sector when it was discovered that it could pass a range of tests. For example, GenAI can generate essays rated higher quality than those produced by humans (Herbold et al., 2023); pass questions involving communication skills, ethics, empathy, and professionalism in a United States Medical Licensing Examination (Brin et al., 2023); pass board-style questions provided by the Joint Commission on National Dental Examinations (Danesh et al., 2023); pass questions from introductory and advanced financial accounting

(Abeysekera, 2024); and pass a reading comprehension test from the Programme for International Student Assessment (PISA), an international student test (Vázquez-Cano et al., 2023). Reports of such success created many concerns in higher education in relation to plagiarism, copyright issues and academic dishonesty (Mai et al., 2024, Bahroun et al., 2023) and has resulted in academics being more critical of student work (Farazouli et al., 2023). The limitation of the studies was that they were very specific and did not provide insights across the broad spectrum of assessments used in higher education. Within engineering, little was known about its capability to engage with technical content.

Against this backdrop, our benchmark study (Nikolic et al., 2023a) was one of the first and most systematic approaches to gaining a holistic understanding of ChatGPT-3.5’s ability to pass different assessment types and tackle different technical content areas. It is recommended that this paper be read first to gain an appropriate understanding of the benchmarking process and capabilities of ChatGPT-3.5. The study highlighted that ChatGPT-3.5 could handle most written activities but struggled with numerical activities and often hallucinated references or information. We predicted that the upcoming ChatGPT-4 and plugins had the potential to improve performance substantially.

GPT-4 was released in March 2023 with a significantly larger training dataset (1.76 trillion parameters) and enhanced through fine-tuning and Reinforcement Learning from Human Feedback, aimed to address these deficiencies (López Espejel et al., 2023). Despite its advancements, no comprehensive study has holistically explored GPT-4's capabilities in assessment tasks, leaving unknown the extent of academic integrity risks. Furthermore, ChatGPT-4 introduced plugins (called ‘GPTs’) to access up-to-date information, perform computations, and integrate with third-party services (OpenAI, 2023). For example, ChatGPT-4 could now harness the power of Wolfram Alpha, a powerful tool for mathematics (Lingefjärd, 2024) to potentially increase its numerical capabilities. Using the Wolfram GPT, ChatGPT can formulate a query and then send it to Wolfram Alpha for computation (Wolfram, 2023). Given ChatGPT-3.5's poor computational performance (Nikolic et al., 2023a), evaluating these capabilities is crucial, especially in engineering. Similarly, GPT-3.5 struggled to engage with scholarly literature, often hallucinating or outputting old references (Nikolic et al., 2023a). SciSpace is a popular AI-powered tool to simplify research discovery and learning, containing metadata of over 200 million papers and 50 million open-access full-text PDFs (Pinzolits, 2024). A SciSpace plugin would be expected to improve research outputs, akin to how the Wolfram plugin could enhance computation. Therefore, it is of interest to understand the impact of these GPTs.

While ChatGPT garnered headlines initially, other GenAI technologies have since emerged. Microsoft's significant entry into the GenAI space came through its partnership with OpenAI's GPT-4, resulting in Bing Chat, which benefits from internet access for current information (Rudolph et al., 2023). Building on Bing Chat, Microsoft rolled out Copilot in September 2023 across Windows, Enterprise, and Microsoft 365, offering varying levels of security, data privacy, and integration (Spataro, 2023). A major advantage of Copilot for educational use is that it provides footnotes with links to sources and can provide proper academic references upon request (Rudolph et al., 2023). Given that Copilot is built on ChatGPT-4 and embedded within Microsoft Windows, which held a 68.15% market share in 2024 (Sherif, 2024), understanding its capabilities is vital as it is highly likely to be used.

Announced a day before Bing Chat, Google Bard is powered by Google's LaMDA, a language model similar to Microsoft's GPT (Rudolph et al., 2023). On the 15th of February 2024, Google introduced Gemini 1.5, Bard's successor, designed to handle multiple modalities like text, programming code, images, and video, integrating seamlessly with Google's products (Pichai and Hassabis, 2024). Given

Google's 82% market share in desktop search engines and its dominant applications like Gmail and YouTube (Bianchi, 2024), understanding Gemini's capabilities is essential.

Several studies with limited scope have tried to compare the different GenAI models. The problem with such studies is determining the best way to undertake the evaluation with different techniques providing different outcomes, for example (Chan et al., 2024, Street et al., 2024). With no comprehensive education-based assessment integrity benchmark developed, and given the substantial evolution of ChatGPT-4, the unknown capabilities of GPTs, and the widespread potential reach of Copilot and Gemini, an update of our original benchmarking study to observe and document the differences and impacts on various assessment types, is warranted.

Understanding the impacts on various assessment types is crucial to maintaining academic integrity in higher education. In Australia, universities must meet strict quality standards, including the assessment of learning outcomes (Tertiary Education Quality and Standards Agency, 2024). For professional accreditation in fields like engineering, quality assurance ensures graduates meet the necessary competencies (ABET, 2014, Engineers Australia, 2008). This ensures graduates are workforce-ready and capable of leading in innovation and safety. While academic integrity requires an institution-wide strategy (Ellis and Murdoch, 2024), assessments play an important role. Effective assessment design is key as it directly influences teaching, learning, and student motivation (Hargreaves, 1997)For this reason, a central part of the analysis will be providing updates on changes to assessment implementation and design that can secure assessment practice.

Method:

In order to answer the research questions, determining the performance improvements and identifying the best GenAI against different higher education assessment types, this study follows the same procedure as the original benchmarking study outlined in Nikolic et al. (2023a). Nine academics from seven Australian universities, all with different engineering backgrounds, collaborated to investigate how assessments in different engineering subjects stood up against ChatGPT-3.5. However, in this paper, we expand the scope by considering advances in ChatGPT and comparing against a broader suite of GenAI tools. By assessing the same subjects, this study compares the progress of ChatGPT-3.5 from the first quarter of 2023 to the first quarter of 2024 (mid-February to early March). We use this dated approach to identify versions because ChatGPT hides the build versions from the general user, and these are regularly updated. Greater insights are available at OpenAI (2024b). Additionally, the performance of ChatGPT-4 was examined to compare the impact of the newer model and larger training data. This paper also includes GPTs (Wolfram for numerical-based assessments and SciSpace for research-based assessments) in ChatGPT-4 to check against the performance predictions made in Nikolic et al. (2023a). Finally, this study compares two major competitor products, Microsoft (Copilot) and Google (Gemini). The breadth of this study will provide greater insights into the relationship between assessment integrity and GenAI, as well as current capability.

The research aims to evaluate whether the selected GenAI platforms could assist students in passing various types of assessments and the level of prompt engineering difficulty involved. The methodology focused on assigning a straightforward pass-or-fail outcome for assessments with subjective answers while noting specific grades for questions with definitive answers. For subjective content, a grade beyond pass or fail was avoided due to the bias of knowing that the output was AI-generated.

The researchers adjusted GenAI's inputs to explore the feasibility of achieving a passing grade, highlighting the importance of prompt engineering on output effectiveness. Through this process, the team assessed GenAI's ability to pass assessments and explored its potential to enhance learning, identifying new opportunities and limitations as the team gauged its capabilities. This level of detail has been omitted from the core paper. Greater insights and explanations are available in the supplementary materials.

The group utilised common methods and formats for uniform record-keeping. For instance, logged examples of both the input and output were used to facilitate a collective review of the adjustments made to the inputs. Following the data-gathering phase, the team reconvened to examine the outcomes, exchange insights, and perform the analysis necessary for this paper.

The data collection process described above was formally conducted as follows:

- For each subject, the team member responsible needed to test all assessment tasks outlined on the subject outline (the formal university documentation regarding subject structure). The team members were the subject coordinators and had designed the assessments (which were modified or replaced for future use).
- The team member would first attempt to copy and paste the question into each GenAI tool.
    - If the copy-pasted input produced an output that would pass, they would record "None" (as in no modification required to pass) and move on to the next question.
    - If the output would not produce a passing grade, the team member would reflect on the relationship between the prompt and output to re-engineer the prompt. This process was aided by the structure outlined above.
        - If the output would result in a pass, they would move to the next question, or if the reflection warranted it, further prompt engineering exploration was undertaken to explore what it would take to get a better result to aid the discussion.
        - If the output still did not result in a pass, they would continue to re-engineer the prompt. When in doubt, the gatekeeper was consulted. The adjustment cycle continued until it was determined that a passing grade was not possible, or a passing grade was achieved – in which case a judgement of Minor or Major (modifications required to pass) was recorded against that assessment item.
- When all assessments were completed, the results and examples were provided to the team. This allowed for feedback and knowledge transfer on practical, prompt engineering approaches. For each assessment, the responsible team member recorded which GenAI tool generated the strongest response with the least prompt modification.
- If the assessment task involved a random question set, the team member just needed to complete the assessment task once. That is, they did not retake the test to check different combinations of questions.
- If the assessment is not subjective (only one possible answer), a grade is provided. If the grade is subjective, only a pass or fail is recorded.

A blend of ten undergraduate and postgraduate subjects from seven universities was chosen to represent the variety of knowledge and skills required of engineers. This approach offers a comprehensive analysis of GenAI's effects on engineering education assessment, considering the similarities and differences in the types of assessments used. Specific details like subject codes and locations have been anonymised to maintain data confidentiality. Despite the variation in terminology across different universities and settings, within this study, the term ‘subject’ denotes a single study unit constituting approximately one-fourth of a full-time semester's workload, equating to an estimated 100-150 hours of student workload.

Subjects (abbreviations used for tables) investigated include:

First-Year Foundational:

- ***Engineering Physics (Physics)**:* a first-year undergraduate subject for all engineering disciplines. The subject introduces the fundamentals of engineering physics with appropriate applications in a wide range of engineering and industrial design systems.
- ***Maths**:* a first-year undergraduate subject common for all engineering disciplines focused on Calculus and Linear Algebra. A contextual focus from various engineering disciplines is used.

Technical:

- ***Introductory Programming (Prog)***: a first-year undergraduate subject for all engineering disciplines and other related fields. Using Python, students develop computational thinking to solve problems, focusing on building small programs for specialised tasks.
- ***Manufacturing Technology (Man Tech)***: a second-year undergraduate subject primarily focusing on mechanical engineering. Individually and through teamwork, students develop an understanding of modern manufacturing processes, production systems and quality management systems.
- ***Engineering Laboratory (Lab)***: a third-year undergraduate laboratory subject tailored for chemical engineers. In teams, students are presented with open-ended projects requiring them to lead, plan and execute laboratory work.
- ***Sustainable Product Engineering and Design (Design)***: first-year to final-year undergraduate, multi-disciplinary engineering and science subject. Issues concerning sustainability, safety, the engineering profession, and sustainable manufacturing are explored. Students also gain skills in information literacy.
- ***Renewable Energy and Electrical Power (Power)***: a fourth-year undergraduate and co-badged postgraduate subject for electrical engineers. Students develop knowledge in renewable energy technologies, energy storage systems, power electronics interfaces and associated control, and renewable energy system design and implementation.

Social context of engineering/professional skills:

- ***Sustainable, Environmental & Social Impacts of Technology (Impact of Tech)***: a senior-level core undergraduate subject looking at the social dimensions of emergent technologies. Students learn about ethics, engagement and consultation with stakeholders, public policy, sustainability, and other contextual considerations, and then apply these concepts to exploring an emergent technology, such as self-driving cars or facial recognition software, from a range of perspectives.
- ***Workplace Practice & Communication (WIL & Com)***: a multi-disciplinary postgraduate work-integrated learning (WIL) subject. This subject is focused on developing key employability skills required to be successful in the Australian workforce. Students refine their professional communication skills and work in a team as consultants working on real industry problems.

Research:

- ***Engineering Research (Research)***: co-taught between fourth-year undergraduate and postgraduate students (postgraduate assessment structure used in the study). The focus is on developing skills in framing a research problem, developing a research design, and designing data collection analysis and interpretation frameworks.

Limitations:

As outlined in the benchmark study (Nikolic et al., 2023a), the rate of change in this field is staggering. As such, the results presented will be outdated by the time this article is published. To ensure consistency/uniformity against various tools, the team worked together against a tight time window in February 2024. Checks and verification occurred in early March. This window was suitable to capture Google's handover from Bard to Gemini on the 15th of February. While the data will be out of date by the time of publication, this approach is expected to provide a reasonable scientific comparison across various platforms.

Every assessment task across all subjects was approached with the assumption that it could be carried out using technology enabling GenAI access, even though some tasks were originally intended for in-person completion. As such, the difficulty of these tasks might vary if they were specifically designed for another format. For instance, the challenge presented by an open-book test may differ from one intended for a closed-book scenario. Despite this constraint, an evaluation of existing practices was still conducted.

While the research analysed GenAI responses to assessments across ten subjects, we acknowledge that the breadth of subjects covered may be viewed as a limitation in terms of representativeness. Nonetheless, we believe we have addressed this issue by selecting subjects from a diverse array of the engineering curriculum. This includes first-year foundational subjects like math and physics, which are crucial for engineering, as well as technical, research, and professional skills subjects spanning different academic years and encompassing both common core and discipline-specific fields.

There is an enormous and ever-growing number of GenAI applications and plugins, and this study has focussed only on a few selected platforms. There may be other technologies that are better suited to certain assessment types and, therefore, yield different results. Further details on this are shared within the limitations section found in the discussion below.

Results and Discussion:

The assessment types were bucketed into categories of best fit by the team after an analysis of the requirements of each assessment type conducted in each subject. Upon reflection, some assessments have been switched to a different category compared to the benchmark study. The categories were defined as:

- **Online Quiz:** tasks that used an online quiz format using an e-learning platform.
- **Numerical (Assignments and Exams):** assessment tasks where the answers are numerical in nature (e.g., calculation-based) and are completed in a written format (not online).
- **Oral:** assessments comprising presentations, interviews, pitches and quality participation in discussion.
- **Visual:** visual documents (e.g., mind map) and evidence (e.g., completion certificate)
- **Programming:** assessments requiring the submission of programming code.
- **Written (Experimentation-based):** written activity associated with experimentation or laboratory work.

- **Written (Project-based):** written assessment activity associated with project work (e.g., project report)
- **Written (Reflective & Critical Thinking-based):** written assessment tasks that focused on reflective and critical thinking (e.g., reflection on student experience, strengths and weaknesses)
- **Written (Research-based):** assessments focused on research-based writing (e.g., thesis).

The results and discussion will explore each assessment type individually. As per the original study, we consider the results in terms of a battle between assessment integrity and GenAI capability. Then, the discussion will focus on the ability to upload images and documents and generate diagrams, compared to the benchmark study when this was not possible. This is followed by a discussion of assessment in the age of GenAI, analysing the results from this study compared to other recommendations.

For each assessment type an overview of the previous findings from the benchmark study are provided, together with a commentary of the findings from this study. Any observed opportunities and recommendations follow this.

It should be noted that Tables 1-9 present a simplified version of the analysis. The supplementary materials provide full details for each table.

Our framework for judging subjective assessment outputs

Some assessment types are scored subjectively, where the marker makes a subjective judgement in interpreting the submission against the marking rubric. In our own evaluation of GenAI output for such assessments, we used the following framework to categorise the outputs:

- **Pass:** At the minimum, a passing grade could be achieved (if the assessment is not subjective, the grade is provided).
- **Fail:** A passing grade was not achieved.
- **Component Pass:** While an overall fail grade was achieved, components of the assessment would receive a pass.
- **Possible Pass:** This would be a borderline pass/fail and conditional either on another assessment component or students touching up the answer (e.g. supplementing it with a real reference).

In Tables 1-9 below, the performance of different GenAI platforms has been evaluated for nine different assessment types. As most assessments have been judged subjectively, the analysis uses the above framework to evaluate performance. For online quizzes, and numerical and programming tasks, where performance can be judged objectively, the evaluations are given as percentages.

Table 1: Analysing Generative AI Performance for the Online Quiz Assessment Type

			Chat3.5 Q1 2023	Chat3.5 Q1 2024	Chat4 Q1 2024	MS Copilot Q1 2024	Gemini Q1 2024	Wolfram Q1 2024	
Subject	Assessment Name	Assess Weight	Grade	Grade	Grade	Grade	Grade	Grade	Best Tech
Power	In-semester tests	30%	63%	66%	80%	65%	71%	N/A	GPT4
Design	Summative quiz	15%	80%	88%	97%	97%	100%	N/A	Gemini
Man Tech	Theory-based understanding	15%	100%	60%	77%	53%	59%	53%	GPT4
Maths	Quizzes	10%	60%	58%	95%	48%	48%	98%	GPT4 or Wolfram
Research	Online Quizzes	10%	52%	68%	82%	68%	50%	N/A	GPT-4
Lab	Introductory Quiz	5%	15%	17%	17%	20%	7%	N/A	Copilot
WIL & Com	Moodle Quiz	5%	58%	53%	84%	84%	53%	N/A	Copilot

Table 2: Analysing Generative AI Performance for the Numerical (Assignments and Exams) Assessment Type

			Chat3.5 Q1 2023	Chat3.5 Q1 2024	Chat4 Q1 2024	MS Copilot Q1 2024	Gemini Q1 2024	Wolfram Q1 2024	
Subject	Assessment Name	Assess Weight	Grade	Grade	Grade	Grade	Grade	Grade	Best Tech
Physics	End of session exam	50%	70%	54%	78%	34%	60%	80%	GPT4 or Wolfram
Man Tech	Final Exam	50%	64%	38%	74%	48%	63%	55%	GPT4
Maths	Mid-Session Exam	35%	43%	57%	96%	48%	41%	96%	GPT4 or Wolfram
Maths	Final Exam	35%	65%	77%	100%	56%	65%	100%	GPT4
Physics	Intra-Session Exam 1	25%	66%	62%	80%	58%	72%	92%	Wolfram
Man Tech	Assignment 1	20%	0%	4%	17%	6%	18%	16%	GPT4 or Gemini
Physics	Intra-Session Exam 2	15%	36%	44%	64%	24%	12%	36%	GPT 4
Maths	Assignment 3	11%	79%	75%	88%	67%	58%	88%	GPT4 or Wolfram
Maths	Assignment 2	6%	50%	48%	76%	32%	36%	80%	GPT4 or Wolfram
Maths	Assignment 1	3%	70%	70%	90%	45%	50%	95%	GPT4 or Wolfram

Table 3: Analysing Generative AI Performance for the Oral Assessment Type

			Chat3.5 Q1 2023	Chat3.5 Q1 2024	Chat4 Q1 2024	MS Copilot Q1 2024	Gemini Q1 2024	
Subject	Assessment Name	Assess Weight	Pass/ Fail	Pass/ Fail	Pass/ Fail	Pass/ Fail	Pass/ Fail	Best Tech
Lab	Final Seminar	20%	Possible Pass	Possible Pass	Possible Pass	Possible Pass	Possible Pass	Gemini
Research	Presentation	10%	Component Pass	Possible Pass	Possible Pass	Possible Pass	Possible Pass	GPT-4
WIL & Com	Class Participation	10%	Fail	Fail	Fail	Fail	Fail	N/A
Lab	Proposal defence	9%	Fail	Fail	Fail	Fail	Fail	N/A
Power	Renewable Energy Design Project (Presentation)	7.5%	Possible Pass	Possible Pass	Possible Pass	Possible Pass	Possible Pass	Gemini
Design	Group project - Pitch	5%	Possible Pass	Possible Pass	Possible Pass	Possible Pass	Possible Pass	GPT4
Power	Individual interview and ePortfolio (interview part)	5%	Fail	Fail	Fail	Fail	Fail	N/A
WIL & Com	Presentation of the Industry project proposal	5%	Fail	Fail	Fail	Fail	Fail	N/A
WIL & Com	Reflection on Engineering Practices and Standards	5%	Possible Pass	Possible Pass	Possible Pass	Possible Pass	Possible Pass	ALL
WIL & Com	Industry Project Final Presentation	5%	Fail	Fail	Fail	Fail	Fail	N/A
Power	Renewable Energy Design Project (Individual progress Presentation)	4.5%	Possible Pass	Possible Pass	Possible Pass	Possible Pass	Possible Pass	Gemini
WIL & Com	Job Application Process	4%	Possible Pass	Possible Pass	Possible Pass	Possible Pass	Possible Pass	ALL

Table 4: Analysing Generative AI Performance for the Visual Assessment Type

			Chat3.5 Q1 2023	Chat3.5 Q1 2024	Chat4 Q1 2024	MS Copilot Q1 2024	Gemini Q1 2024	
Subject	Assessment Name	Assess Weight	Pass/ Fail	Pass/ Fail	Pass/ Fail	Pass/ Fail	Pass/ Fail	Best Tech
WIL & Com	Career Upskilling	10%	Fail	Fail	Fail	Fail	Fail	None
WIL & Com	Career Portfolio	7%	Fail	Pass	Pass	Pass	Pass	Gemini
Impact of Tech	Stakeholder persona	5%	Pass	Pass	Pass	Pass	Pass	All
Research	MindMap - Introduction Chapter	2%	Fail	Fail	Fail	Fail	Fail	N/A
WIL & Com	Career Ready Skills	1%	Fail	Fail	Fail	Fail	Fail	N/A

Table 5: Analysing Generative AI Performance for the Programming Assessment Type

			Chat3.5 Q1 2023	Chat3.5 Q1 2024	Chat4 Q1 2024	MS Copilot Q1 2024	Gemini Q1 2024	
Subject	Assessment Name	Assess Weight	Grade	Grade	Grade	Grade	Grade	Best Tech
Prog	Final Exam	55%	100%	100%	100%	100%	100%	Any
Prog	Project 2	20%	0%	0%	0%	0%	0%	N/A
Prog	Project 1	15%	30%	33%	13%	7%	10%	N/A
Prog	Weekly Labs	10%	87%	87%	94%	91%	72%	GPT4

Table 6: Analysing Generative AI Performance for the Written (Experimentation-based) Assessment Type

			Chat3.5 Q1 2023	Chat3.5 Q1 2024	Chat4 Q1 2024	MS Copilot Q1 2024	Gemini Q1 2024	
Subject	Assessment Name	Assess Weight	Pass/ Fail	Pass/ Fail	Pass/ Fail	Pass/ Fail	Pass/ Fail	Best Tech
Lab	Technical reports	45%	Component Pass	Component Pass	Possible Pass	Possible Pass	Possible Pass	GPT4
Lab	Experiment proposals	21%	Component Pass	Component Pass	Possible Pass	Possible Pass	Possible Pass	Gemini
Man Tech	Lab Report	15%	Component Pass	Component Pass	Component Pass	Component Pass	Component Pass	GPT4
Physics	Practical	10%	Component Pass	Component Pass	Component Pass	Component Pass	Component Pass	GPT4
Power	Lab work and report	10%	Component Pass	Component Pass	Component Pass	Component Pass	Component Pass	GPT4

Table 7: Analysing Generative AI Performance for the Written (Project-based) Assessment Type

			Chat3.5 Q1 2023	Chat3.5 Q1 2024	Chat4 Q1 2024	MS Copilot Q1 2024	Gemini Q1 2024	
Subject	Assessment Name	Assess Weight	Pass/ Fail	Pass/ Fail	Pass/ Fail	Pass/ Fail	Pass/ Fail	Best Tech
WIL & Com	Industry project – Final report	30%	Component Pass	Component Pass	Component Pass	Component Pass	Component Pass	Copilot
Design	Group project - Preliminary report	20%	Pass	Pass	Pass	Pass	Pass	Copilot
Design	Group project - Final report	20%	Possible Pass	Possible Pass	Pass	Pass	Pass	GPT4
Power	Renewable Energy Design Project (Report)	18%	Fail	Component Pass	Component Pass	Component Pass	Component Pass	GPT4
WIL & Com	Client Research	4%	Component Pass	Possible Pass	Possible Pass	Possible Pass	Possible Pass	Copilot

Table 8: Analysing Generative AI Performance for the Written (Reflective & Critical Thinking-based) Assessment Type

			Chat3.5 Q1 2023	Chat3.5 Q1 2024	Chat4 Q1 2024	MS Copilot Q1 2024	Gemini Q1 2024	
Subject	Assessment Name	Assess Weight	Pass/ Fail	Pass/ Fail	Pass/ Fail	Pass/ Fail	Pass/ Fail	Best Tech
Impact of Tech	Weekly Worksheets	30%	Component Pass	Component Pass	Pass	Component Pass	Component Pass	GPT-4
Power	Individual interview and ePortfolio (portfolio part)	25%	Possible Pass	Possible Pass	Possible Pass	Possible Pass	Possible Pass	Gemini
Design	Class participation	10%	Possible Pass	Pass	Pass	Pass	Pass	GPT4
Design	Team evaluations	10%	Pass	Pass	Pass	Pass	Pass	GPT4
Impact of Tech	Evaluation & recommendation	10%	Pass	Pass	Pass	Pass	Pass	All
Impact of Tech	Stakeholder consultation	6%	Pass	Pass	Pass	Pass	Pass	Gemini
Impact of Tech	Incorporation of feedback	6%	Fail	Possible Pass	Possible Pass	Possible Pass	Possible Pass	GPT4
Design	Class participation	5%	Pass	Pass	Pass	Pass	Pass	GPT4
Design	Peer evaluation	5%	Pass	Pass	Pass	Pass	Pass	GPT4
Research	Critique	5%	Fail	Fail	Fail	Fail	Fail	N/A
WIL & Com	Career Ready Skills	4%	Pass	Pass	Pass	Pass	Pass	ALL
WIL & Com	Job Application Process	4%	Pass	Pass	Pass	Pass	Pass	Gemini
Impact of Tech	Stakeholder mapping	3%	Fail	Fail	Pass	Pass	Fail	GPT4
WIL & Com	Job Application Process	1%	Pass	Pass	Pass	Pass	Pass	GPT4
WIL & Com	Client Research	1%	Pass	Pass	Pass	Pass	Pass	Gemini

Table 9: Analysing Generative AI Performance for the Written (Research-based) Assessment Type

			Chat3.5 Q1 2023	Chat3.5 Q1 2024	Chat4 Q1 2024	MS Copilot Q1 2024	Gemini Q1 2024	GPT4 + SciSpace Q124	
Subject	Assessment Name	Assess Weight	Pass/ Fail	Pass/ Fail	Pass/ Fail	Pass/ Fail	Pass/ Fail	Pass/ Fail	Best Tech
Research	Final Report	41%	Fail	Possible Pass	Possible Pass	Fail	Possible Pass	Possible Pass	GPT4 + SciSpace
Impact of Tech	Policy initiatives	18%	Pass	Pass	Pass	Pass	Pass	Pass	GPT4 + SciSpace
Research	Introduction Chapter	18%	Fail	Fail	Possible Pass	Fail	Possible Pass	Possible Pass	GPT4 + SciSpace
Design	Group project - Technical research task	10%	Fail	Possible Pass	Pass	Pass	Pass	Pass	GPT4 + SciSpace
Research	Final Report	10%	Fail	Pass	Pass	Pass	Pass	N/A	Gemini
Impact of Tech	Summary of topic	6%	Possible Pass	Pass	Pass	Pass	Fail	N/A	GPT3.5
Impact of Tech	Stakeholder stake	6%	Possible Pass	Fail	Pass	Fail	Pass	N/A	GPT4 or Gemini
Impact of Tech	Stakeholder's stake	5%	Fail	Fail	Pass	Pass	Pass	N/A	GPT4
Impact of Tech	Topic summary	5%	Possible Pass	Pass	Pass	Pass	Pass	N/A	GPT3.5 or GPT4
Research	Progress Check	5%	Fail	Fail	Component Pass	Component Pass	Component Pass	Component Pass	GPT4 + SciSpace

Online Quiz

Table 1 outlines the results associated with the Online Quiz. Seven subjects used this assessment type. The results show that the various GenAI tools can help students cheat. The engineering laboratory subject had the only quiz that none of the GenAI platforms could pass, frequently erring in both knowledge-based and verification questions. Therefore, regardless of platform, GenAI is the winner for this assessment type.

Benchmark Study: The online quiz was identified as high risk with a pass mark obtainable for almost every activity. ChatGPT-3.5 was exceptionally good at questions focussing on theoretical knowledge and simple calculation questions. It could not work with figures and tables, and descriptions needed to be translated into words for the prompt. Therefore, a key short-term recommendation of the study was to use figures and tables to make it more difficult and to slow students down (especially if it was a timed quiz).

Q1 2024: A year later the performance of ChatGPT-3.5 remained relatively the same. There were instances where it performed better and worse. As expected, ChatGPT-4 performed substantially better than ChatGPT-3.5 with an average mark improvement of 24%. While ChatGPT-3.5 generally produced marks in the 50 – 70% range, ChatGPT-4 often produced marks in the 80 – 85% range. Depending on the content, Copilot performed at a similar level to ChatGPT-4 or ChatGPT-3.5, but overall, the outcomes were closest to ChatGPT-3.5. The results for Gemini were similar to that of Copilot. ChatGPT-4 was rated the best technology for 7 of the 9 assessment activities. This was due to its strength in getting most questions correct and its ability to upload and analyse images and documents (see 'Images and Documents' below). The advantage of Copilot was its ability to provide references to its output, allowing students to fact-check information easily. When it came to questions that required a particular context outside of general knowledge, most platforms could recognise this and provide a warning, but Gemini tended to excel on this front. Where calculations were a focus of the quiz, the Wolfram plugin was tested. Some slight improvements were noticed, but due to ChatGPT-4's standalone performance, those differences were not substantial.

Opportunity and Recommendation: While image recognition has made a huge leap, complex-to-read figures and tables remain a weak spot for GenAI. Therefore, for at least the next 12 months, adding complex figures and tables can strengthen assessment security. Beyond that, high-stake quizzes aimed to ensure competency, need to be supervised or supported by proven proctoring software that can detect contract cheating or GenAI use, including access through browser extensions. For unsupervised activities, online quizzes would be best suited to formative low-stakes assessments where even if GenAI use were to occur, the activity in itself becomes the learning activity. For example, online quizzes could create the motivation for the student to come across a concept at least once.

Numerical (Assignments and Exams)

Table 2 outlines the results associated with the numerical assignment and exam assessment type. This category refers to non-online quiz questions that are focused on calculations. These questions are generally more complex than those found in online quizzes and require substantial working. A total of three subjects used this assessment type. The results show that the various GenAI platforms have the capability to help students cheat. Success is dependent on the GenAI used and the content matter. Therefore, GenAI is the winner for this assessment type, especially if ChatGPT-4 is used with or without the Wolfram plugin.

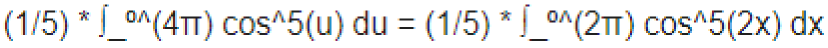
Benchmark Study: The numerical assessment type was a borderline case in which the final answer was generally wrong, but most of the working was correct, providing enough partial marks to get students over the line. ChatGPT-3.5 capability was also mixed, determined by the content area. As with quizzes, assessment security was strengthened through diagrams and figures. Testing with Wolfram Alpha was also conducted, and it was acknowledged that such a plug-in could make a substantial difference in mathematical questions.

Q1 2024: The performance outcome of ChatGPT-3.5 remained similar. One noticeable improvement was that it had become more adept at interpreting questions without specialised syntax from the user. For example, "Find the eigenvalues and eigenvectors for the 3x3 matrix 1,1,2,0,2,3,0,0,3" would be correctly interpreted.

The improvement associated with GPT-4 was substantial, with the average result improving 41% on average compared to ChatGPT-3.5 in Q1 2023. Where the assessment focus was not on questions requiring interpretation of figures, ChatGPT-4 performance could be classified as amazing, especially for pure mathematics, with working and answers mostly correct. ChatGPT-4 solved the issue of poor calculations by deferring to Python. At times, an approximation to the correct answer would be given, but when asked for an "exact" answer, ChatGPT-4 could combine its mathematical knowledge with the correct numerical answer and produce a model answer. The main points for which ChatGPT-4 did not score full marks were the lack of working provided for long answer questions (even when asked for more detail) and misinterpretation of subtle theoretical questions. Due to ChatGPT-4's standalone performance, the Wolfram plugin only provided marginal improvement. If anything, ChatGPT-4 alone tended to provide more working and better-detailed information than when the Wolfram plugin was used. Furthermore, ChatGPT-4 and Wolfram only required minor prompting skills for success, while the other technologies required more prompting and content understanding. ChatGPT-4's ability to produce quality graphs and work with uploaded images was noteworthy, but improvement was needed to become an assessment integrity risk (see 'Images and Documents' below).

It is worth highlighting that only ChatGPT-4 was able to pass the Physics Data Analysis test, albeit with multiple mistakes and partial marks given for working. In this type of assessment each question consists of multiple steps, requires data extraction from the table, and the use of procedures prescribed in the laboratory manuals. The passing grade can be improved by formatting the prompts and typing the specific equations that should be used in the question. However, this requires much work from the student.

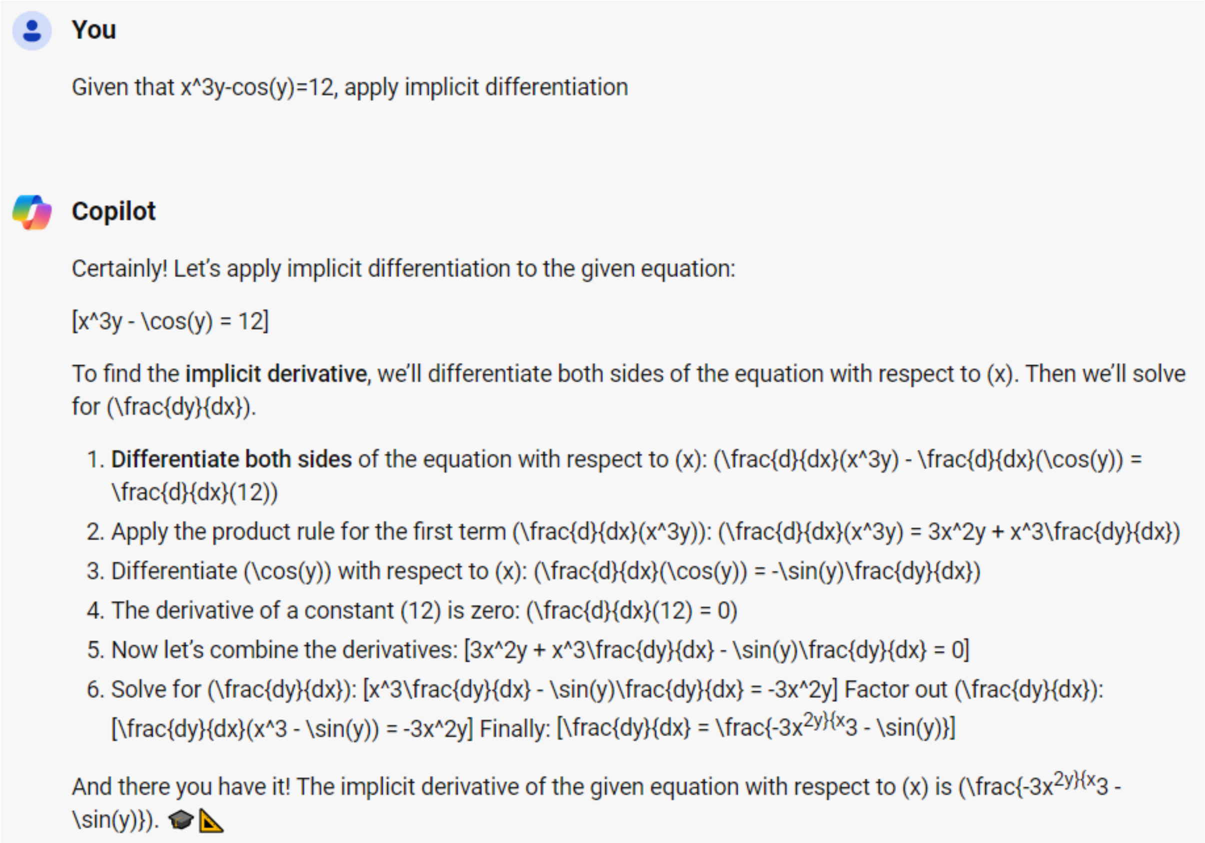
Copilot and Gemini struggled to complete numerical tasks, with performances similar to ChatGPT-3.5. Both Gemini and CoPilot made many calculation errors, however, in contrast to ChatGPT-3.5 which would happily recalculate and end up with a different answer, Gemini and CoPilot more often gave the same incorrect answer. When questioned, Gemini would often respond with, "I am still under development, and I appreciate your patience and guidance in helping me learn and improve." On the other hand, CoPilot was not so polite, remaining silent. A further complication for the user of these two platforms, is in the presentation of the mathematics. Gemini made an attempt to present mathematical expressions such as fractions and square roots clearly, but struggled with more complex expressions, such as the integral shown in equation 1.



(1/5) * ∫_⁰^(4π) cos^5(u) du = (1/5) * ∫_⁰^(2π) cos^5(2x) dx

Equation 1: Example of an expression Gemini struggled to represent

CoPilot only presented simple mathematics clearly, opting instead to provide LaTeX commands such as those shown in Figure 1. In this study, only answers that were readily identifiable were taken as correct. If the syntax was too complicated and the platform would not rewrite it or compile the LaTeX code, the answer was taken to be incorrect as a student user would struggle to interpret the answer correctly.



You

Given that x^3y-cos(y)=12, apply implicit differentiation

Copilot

Certainly! Let's apply implicit differentiation to the given equation:

[x^3y - \cos(y) = 12]

To find the **implicit derivative**, we'll differentiate both sides of the equation with respect to (x). Then we'll solve for (\frac{dy}{dx}).

1. **Differentiate both sides** of the equation with respect to (x): (\frac{d}{dx}(x^3y) - \frac{d}{dx}(\cos(y)) = \frac{d}{dx}(12))
2. Apply the product rule for the first term (\frac{d}{dx}(x^3y)): (\frac{d}{dx}(x^3y) = 3x^2y + x^3\frac{dy}{dx})
3. Differentiate (\cos(y)) with respect to (x): (\frac{d}{dx}(\cos(y)) = -\sin(y)\frac{dy}{dx})
4. The derivative of a constant (12) is zero: (\frac{d}{dx}(12) = 0)
5. Now let's combine the derivatives: [3x^2y + x^3\frac{dy}{dx} - \sin(y)\frac{dy}{dx} = 0]
6. Solve for (\frac{dy}{dx}): [x^3\frac{dy}{dx} - \sin(y)\frac{dy}{dx} = -3x^2y] Factor out (\frac{dy}{dx}): [\frac{dy}{dx}(x^3 - \sin(y)) = -3x^2y] Finally: [\frac{dy}{dx} = \frac{-3x$^{2y\}\{x}$3 - \sin(y)}]

And there you have it! The implicit derivative of the given equation with respect to (x) is (\frac{-3x$^{2y\}\{x}$3 - \sin(y)}). 🎓📐

Figure 1: An example of Copilot struggling to represent mathematical equations

Opportunity and Recommendation: As for online quizzes, image recognition has not reached the stage where complex-to-read figures and tables become an assessment security risk. Integrating question types with complex images and tables will provide some level of security for at least the next twelve months. Beyond that, high-stakes unsupervised assessment practices become a high assessment integrity risk. Especially if ChatGPT-4 or Wolfram was used, students could perform extremely well with its use. Supervised exams remain a secure option, and despite their limitations, remain a popular assessment method in engineering (Gratchev et al., 2024). Advancement in GenAI accuracy provides a growing opportunity for students to use it for tutoring purposes, as considered by (Sánchez-Ruiz et al., 2023) and (Nikolic et al., 2023a). It can be difficult and expensive for students to find a tutor to help them develop their mathematical capabilities. With increased accuracy, users are less likely to learn incorrect information, a significant issue with ChatGPT-3.5.

Oral

Table 3 outlines the results associated with the oral assessment type. This category refers to oral presentations and interviews. A total of five subjects used this assessment type. The results show that the various GenAI platforms have the capability to help students cheat, but as the students need to present the work themselves, there is no clear-cut opportunity. Interview-styled assessments remain strongest. Regardless of platform, GenAI can be used to provide support to the student to help develop scripts to memorise or read. For this reason, there is no clear winner for this assessment type. For this to change, a greater shift to interview assessments would be required.

Benchmark Study: It was found that ChatGPT-3.5 could not take the place of students in oral assessments. If the focus of the oral assessment was on presentation skills, ChatGPT-3.5 could provide students with a script that they could work with and memorise. Advancements in the soon-to-be-released Microsoft Copilot to convert text to full presentations with guiding scripts were also highlighted. When confirming understanding, it was acknowledged that interview-styled assessments would have better assessment security.

Q1 2024: Twelve months on, the results for ChatGPT-3.5 remain the same, and all platforms perform on similar grounds in which they can support a script the student can memorise for a presentation. This scenario is acceptable if the learning objective is simply presentation competency and not knowledge attainment. For some tasks, it was found that Gemini provided the best outline for slides, images and speaker notes for presentations. In contrast, when work was already produced by a student in a document, they could simply upload it into ChatGPT-4, and it was efficiently summarised into speech notes. Therefore, the best platform came down to the needs of the rubric and the starting point for the oral presentation. The version of Microsoft Copilot used did not provide the facility to convert documents into a ready-made presentation, something that would differentiate Copilot from the others if such a version was used.

Opportunity and Recommendation: Over the last twelve months, there has been growing awareness of the assessment integrity provided by oral assessment (Bearman et al., 2023, Newell, 2023). In particular, at a growing number of the author's universities, this has resulted in a trend of oral vivas becoming a key safeguard when academic integrity breaches are identified. There has been some discussion on the impact that oral assessment may have on neurodivergent students, such as those suffering from dyslexia (Borsotti et al., 2024), or students with anxiety (Iannone and Simpson, 2015), but there may be cohorts of students who face disadvantage that may benefit from oral assessment (Shanmugam et al., 2024, Nogues and Dorneles, 2023)This highlights the importance of administrative processes for those with disabilities, regardless of assessment type. Interestingly, the viability of online interviews (not supervised) as a reliable assessment may not be long-term. This is because GenAI could be used to create deepfakes or set up in a way that allows real-time voice recognition, conversion to prompt, and immediate output to a prompt to support answers.

Visual

Table 4 outlines the results associated with the visual assessment type. This category refers to assessments that require visual inspection of documents (e.g. mind map) and evidence (e.g. completion certificate). A total of three subjects used this assessment type. The results show that the various GenAI platforms have some capability, but the current limitations support assessment security. Therefore, this assessment type can be associated with supporting assessment integrity.

Benchmark Study: If the diagram or image was required to be generated in a specific format, ChatGPT-3.5 would not be able to complete the task to achieve a pass grade. At best, it could be used as a support tool for the activities. It could not generate a screen grab of a portfolio, build and demonstrate a personalised working website, or design a research-based mind map.

Q1 2024: Most of the limitations remain in place. If the requirements of the visual evidence are strict or specific enough, GenAI is not of much help. For example, it cannot produce a screen grab of evidence within a specific e-portfolio used within an institution, a picture of the student completing an activity, or a picture of a completion certificate or award. Therefore, for authentic learning experiences, requiring supporting visual evidence for cross-checking can strengthen assessment security. One of the activities that GenAI could now pass was building a portfolio-style, personal website. This would require major prompting, and probably take longer and be more difficult to do than using a web-based website builder, but if coding was the requirement, it provided great suggestions on developing the structure and content and could then turn that into code and content.

Opportunity and Recommendation: Image and video editing with the support of GenAI is rapidly improving and becoming rather powerful. While there may be some copyright and ethical concerns regarding GenAI generated content (Fenwick and Jurcys, 2023) it is becoming clear it will be able to accomplish more and more over time, especially in creating deepfakes (Shoaib et al., 2023). Short term, the recommendation carries over from the benchmark study. Consider how visual evidence can be used individually, or to support other assessment tasks, especially with authentic learning opportunities. For example, consider the feasibility of visual evidence of a student undertaking experimentation to support a laboratory report, or consider a very specifically formatted and structured mind-map that needs to support the structure of a research piece. While the evidence may eventually be doctored (consider, for example, the increasing realism of deepfakes), at the time of writing, it would at least make cheating more difficult.

Programming

Table 5 outlines the results associated with the programming assessment type. This category refers to assessments that require students to submit, analyse or write code in any programming language. Of all assessment types tested, this is the only type to be associated with only one subject, potentially limiting the generalisability of the insights. The results show that the various GenAI tools have some great coding capabilities, but if the assessment is designed correctly, it is possible to ensure assessment integrity. If we exclude the specific CSV use case and focus on general programming, this assessment type can be associated with a win for GenAI.

Benchmark Study: ChatGPT-3.5 was highly capable of completing entry-level code, answering questions, and providing explanations to help students learn how to code or work through error messages. However, it struggled with project assessments that were designed even before GenAI risks emerged. This involved writing a program that placed many design restrictions as a part of the scope and required extensive manipulation of a CSV file.

Q1 2024: Performance twelve months later mirrored that of the benchmark study. This was surprising because demonstrations of ChatGPT-4's coding ability, in particular, showed how easy it could be to write entire programs (Ouh et al., 2023). The design of the project, including the limitations placed on the design, made it very difficult for GenAI tools to create a working program. Multiple authors tried, and failed, and it was concluded that if it was possible, the student would have had the ability to write the code themselves directly if they wanted to.

Opportunity and Recommendation: With the results the same as per the benchmark study, the recommendations remain the same. The explanations, the assistance of working through ideas and translating those ideas to code, and the ability to support error correction, make coding a primary opportunity to use GenAI tools to integrate learning (Bente et al., 2024). This integration can be supported through supervised and well-designed non-supervised assessment.

Written (Experimentation-based)

Table 6 outlines the results associated with the written experimentation-based assessment type. This category refers to assessments that require students to submit, analyse or write some form of report or summary on an experimental experience. Four subjects used this assessment type. The results show that while experimentation cannot be conducted by GenAI, the concentration of written assessment is an academic integrity risk. Therefore, this assessment type can be associated with a tied outcome.

Benchmark Study: ChatGPT-3.5 could not take the place of the student in experimental work, but the assessment integrity weakness came from the concentration of assessments being conducted as written work. If students could get the results from other students or from previous years, GenAI's contribution would have been relatively straightforward. Undertaking experimental work is a strength, and assessment alternatives to take advantage of that need greater consideration.

Q1 2024: The results from 2023 largely remained the same. However, the risk profile is increasing as the platforms build the capability to work with uploaded documents and being able to execute the code they write. When supplied with data (e.g. sources from other students), ChatGPT-4 was able to produce tables and figures that could be copied directly into a final report. The other observation was the varied constraints and strengths of each tool, using a variety of tools in concert could produce a better result than using a single tool. However, achieving this standard requires significant prompt engineering for which the obstacles are falling. The user experience is also improving.

Opportunity and Recommendation: The recommendations remain the same as per the benchmark study. Experimentation provides a tremendous opportunity for students to learn by doing, off limits to GenAI. This should encourage a greater refocus of learning opportunities to the laboratory/experimentation. However, as reported in three recent studies (Nikolic et al., 2021, Nikolic et al., 2023c, Nikolic et al., 2023b) the focus is heavily skewed towards cognitive learning objectives, and written activities ignore the holistic learning capability available from experimentation. If the wider set of experimentation skills and assessment types are considered, assessment integrity can be improved. Laboratory innovations can support this (Dunne and Nikolic, 2021). In the meantime, it is suggested that research be conducted to explore the holistic set of experimental assessment options to improve choice, competency measurement and academic integrity.

Written (Project-based)

Table 7 outlines the results associated with the written project-based assessment type. This category refers to assessments associated with project work, such as reports, engagement plans, scoping requirements, and solutions. A total of three subjects used this assessment type. The results show that while the comprehensive spectrum of project-based activities cannot be completed by GenAI, there are components that can. Even so, this assessment type can still be considered secure.

Benchmark Study: ChatGPT-3.5 could not take the place of students working on a project in a team-based environment. However, written project-based assessments were more successful than other written types of assessments. ChatGPT-3.5 could provide help with components of work, in particular, report types of assessment. ChatGPT-3.5 is particularly useful to help students get started on projects. It was found to be a useful tool for recommendations on brainstorming ideas, engineering standards and regulations. Its biggest weakness was hallucinations where some references and websites were non-existent or inaccurate. Students still needed to do further research to validate and justify the output. Overall, it yielded possible passes if it was prompted with the right information for some report options, making it a somewhat secure assessment type.

Q1:2024: The results from 2023 largely remained the same; that is, ChatGPT-3.5 continued to provide generic, non-specific answers and, more importantly, still did not have the capability to access journal papers published within the last 12 months if required. However, like the experiment-based lab report results, the capability of the newer GenAI's to gather information has improved. Each technology provided a unique perspective and different links to ideas and resources. The written report type assessments were designed so that students needed to work with a citation less than 12 months. GPT-4 explicitly stated it could not do this, but this limitation could be overcome if using the SciSpace plugin. Copilot was able to provide some references within this timeframe. Furthermore, if a student was brainstorming and developing understanding, Copilot provided many good suggestive prompts to help a student move forward.

Opportunity and Recommendation: The recommendation would remain the same as the benchmark study. While all the GenAI tools would enable students to at least pass components, students would need to undertake further detailed research and analysis. Higher-order cognitive thinking would be required to determine the best solutions. Combining the use of all the new GenAI's would be the most beneficial approach for students to gather a stronger perspective of the project and possibilities. As there are many challenges faced with implementing project work (Miao et al., 2024, Lee et al., 2016), GenAI tools can provide a starting point for students to build on. Research to understand best practices for facilitating these educational integrations is recommended.

Written (Reflective & Critical Thinking-based)

Table 8 outlines the results associated with the written reflective and critical thinking-based assessment types. This category refers to assessments that are associated with reflecting on experiences and critical thinking, including ePortfolios, reflections on in-class experiences, evaluations of team members, quantitative and qualitative feedback and reflection on and responses to feedback. A total of five subjects used this assessment type across fifteen different assessment tasks. The results show that GPT-4 and Gemini typically performed as the best technology on these types of tasks. The results also show that this type of assessment is not secure, but for many activities, an understanding of quality was needed.

Benchmark study: ChatGPT-3.5 was mostly successful in generating passable content, particularly if the student provided context with the prompting or built upon the generated content. Responses to basic prompts were generally found to be rather generic, requiring students to be aware of what was missing and what additional information they would need to add to the prompt. ChatGPT-3.5 performed well at helping students to structure reflective and critical thinking writing but inhibits the development of deep reflective skills for the student.

Q1 2024: Twelve months on and the performance of ChatGPT-3.5 had mostly remained constant, with two assessments seeing some improvement. Generally, major prompt engineering was required to move towards passing grades, particularly in integrating information from real experiences on which to reflect or in students modifying provided outputs to contextualise with their own personal experiences. This is consistent for critical thinking and writing case studies where students are required to reflect based on their classroom teaching and learning experiences (for example, through practical workshops attended, industry site visits or industry guest seminars) or providing self-reflection and team feedback (for example, peer evaluation in the context of professional development such as leadership, teamwork, etc.). Gemini explicitly recognises its own gap and encourages users "to use their own learning," while GPT-4 provides students with a framework on reflection to personalise their work. The ability to upload files and images in ChatGPT-4 and Copilot provided advantages in some of these critical thinking tasks.

Opportunity and Recommendations: Assessment security can be increased through tasks that contain video, which cannot currently be analysed by any of the GenAI tools. However, as technologies further develop, this advantage may be short-lived. Many of the GenAI tools recognised weaknesses in the prompts and, in some cases, guided the user to provide better information, resulting in a better output. Specific GenAI-based video editing and generation tools outside the scope of this study are already available.

Written (Research-based)

Table 9 outlines the results associated with the written research-based assessment type. This category refers to assessments that require students to research information, requiring higher-order cognitive skills such as analysis, synthesis and evaluation. A total of three subjects used this assessment type. The results show that substantial change has occurred in twelve months. Various GenAI tools have the capability to help students cheat, but it still requires the students to understand quality and expectations. Yet GenAI is the winner for this assessment type, especially when using ChatGPT-4.

Benchmark Study: It was found that ChatGPT-3.5 was great at fact-finding but was limited in its ability to reference real, current, and relevant literature. Its tendency to hallucinate references was a major concern but also a giveaway for educators suspicious of students' misusing GenAI output. Text length limitations meant users could work with only a small chunk of information at a time. As a result, it failed many critical research tasks. However, many opportunities were discovered, including helping students to understand research, idea generation and structure, research questions and methods, and improve writing quality.

Q1 2024: Twelve months later, ChatGPT-3.5 has performed slightly better. The biggest differences are with the new generation of GenAI tools. Each excelled differently, depending on the task at hand. For these new platforms, hallucination has substantially reduced, with these tools able to access and cite real current research. Writing a quality traditional literature review is possible, but major prompting skills and expectation awareness are required. However, on social media platforms such as YouTube, there is a growing library of guided tutorials that step users through how to do this and how to avoid plagiarism detection and hallucinations, especially if using SciSpace (note: the research team has decided not to share any example links to avoid promotion of such resources). ChatGPT-4's ability to work with uploaded PDFs (such as those found with SciSpace) made for a seamless experience in decoding research. SciSpace provided the opportunity to efficiently find relevant papers (especially recent ones beyond ChatGPT-4's training date) that meet specific criteria, as well as smart summaries and powerful paraphrasing tools (by going outside of the ChatGPT-4 plugin and using the direct platform). Nevertheless, in some instances, SciSpace was not required to get a high-scoring response as GPT-4 unaided could cite and discuss real relevant research. It is important to note that SciSpace is only one of many powerful tools that interface with real and recent scientific literature.

Opportunity and Recommendation: If we consider research in two parts, one being the analysis and summary of previous information to be scaffolded upon, the other, research experimentation, then research experimentation is a secure component. For the literature component, GenAI has made substantial gains. Traditional literature searching and analysis provided an immersive experience for developing holistic understanding and comprehension skills. However, for someone experienced with those skills, and seeking to develop their publishing record, GenAI is the tool to increase efficiency. This potentially allows researchers to focus more on the experimentation, rather than the literature.

The use of GenAI as a tool in this way brings about an ethical debate. Two of the authors are associate and deputy editors for high-ranking journals, and they have experienced attempts by authors to submit GenAI-written articles. This is not a standalone case. There have been numerous papers published with GenAI prompting accidentally incorporated into the body of published papers. At the time of writing, two examples of such papers with the accidental prompting visible and no ethics declaration of GenAI use include Zhang et al. (2024) and Bader et al. (2024). The question becomes, how did the quality review process allow this to happen? This is not an isolated case. Undertaking a search for "*as of my last knowledge update*", a key output prompt in ChatGPT, Google Scholar finds over 100 academic studies. Studies already suggest that GenAI is useful for research in that it can enhance efficiency and quality by speeding up writing, developing outlines, adding details, and improving writing style (Huang and Tan, 2023, Alshami et al., 2023). If the academic community is finding the efficiency of GenAI too good to pass up, it can be well assumed that students will too. The question for the community is how do we integrate and acknowledge this? How do we ensure learning and quality? How do we balance this?

Currently, assessment security can be strengthened by requiring students to write their literature components in a strictly structured format. All platforms struggled with the structured components used within the research subject. It required some clever prompting to get the GenAI platforms to produce something close to acceptable.

If cheating is put aside, and GenAI is put to ethical use, many of the opportunities discovered in the benchmark study have only gotten better. The ability of GenAI to provide structure, brainstorming, relevant references and understanding, has improved. If a student wants to learn and not cheat, GenAI can be a powerful supporting tool.

Images and Documents

In the benchmark study, a short-term recommendation was the use of images and tables to slow down GenAI capability. A warning was in place that image recognition was coming and that such approaches would not last long. As discovered in the data analysis, substantial progress on this front has been made. While not perfect, the trajectory is on track, and current limitations will eventually disappear. In this section, progress is investigated.

Uploading Images

For simple images requiring an explanation of what the image is depicting, all the new GenAI's perform reasonably well. Mostly accurate descriptions were provided. Complications arose for Copilot if facial expression needed analysis as Copilot applies a privacy blur. Gemini, on the other hand, would not analyse any image that contained people. When it came to engineering context requiring mathematical problem solving, ChatGPT-4 was easily the best. However, there were limits to how complex the image could get before the answer would be wrong.

As an example, consider the circuits shown in Figure 2. When it came to finding the total circuit current for circuit a, it could read the image correctly and undertake the correct analysis to obtain the correct results. However, the slight increase in difficulty for circuit b, by simply adding an additional power source, led to mistakes being made. Minor mistakes, but mistakes are highly likely to be eliminated in 12 months' time.

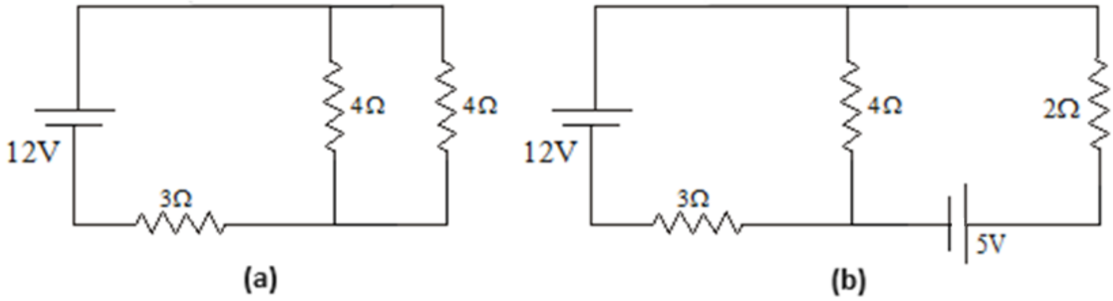


Figure 2. Example of circuit diagrams examined (a) GenAI can easily find the total current and (b) GenAI made minor mistakes in finding the total current

A second example is shown in Figure 3. For this activity, the shape factor needed to be calculated. When asked to determine the shape factor directly, most of the GenAI returned incorrect answers consistently. However, ChatGPT-4 offered a somewhat promising solution. When broken into its components to find out why, it was found that ChatGPT-4 used an incorrect equation. After correcting the equation, the answer remained incorrect. Further dissection was conducted, and it was observed that ChatGPT-4 calculated the volumes for both the shapes correctly, however, made some minor interpretation errors in calculating the surface areas, which is why the answer provided by ChatGPT-4 was incorrect. The authors believe that accurately determining the volume of these complex shapes is a significant accomplishment, showcasing future potential for such questions to be correct.

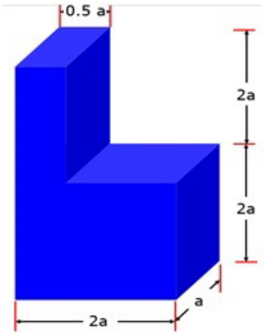


Figure 3. GenAI was tested to find out the shape factor of this object

A third example is shown in Figure 4. Once again, ChatGPT-4 demonstrated a promising initial approach but ultimately fell short of achieving the desired outcome. The question depicted in Figure 4 required calculating the resultant force from two forces shown on a diagram using the method of components. The correct process involves splitting each force into horizontal and vertical components and adding them as vectors. GPT-4 could read data from the diagram, such as force magnitudes and angles. However, it didn't correctly apply the vectors' direction, which is critical for this question. For example, it calculated the x-component of the 80N force as 80cos(30), while the 30-degree angle is opposite the x-component. Instead, 80sin(30) should be used. For the 40N force, it calculated the x-component correctly using 40cos53 but didn't consider the negative direction of the force to the left.

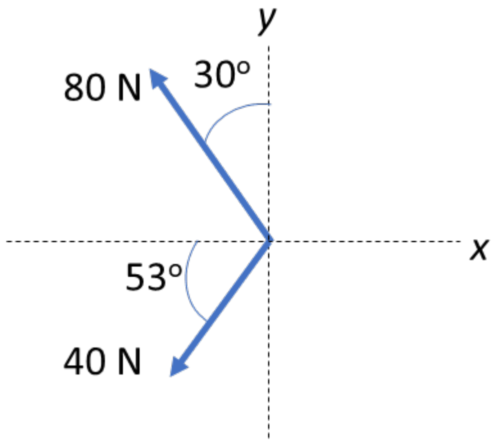


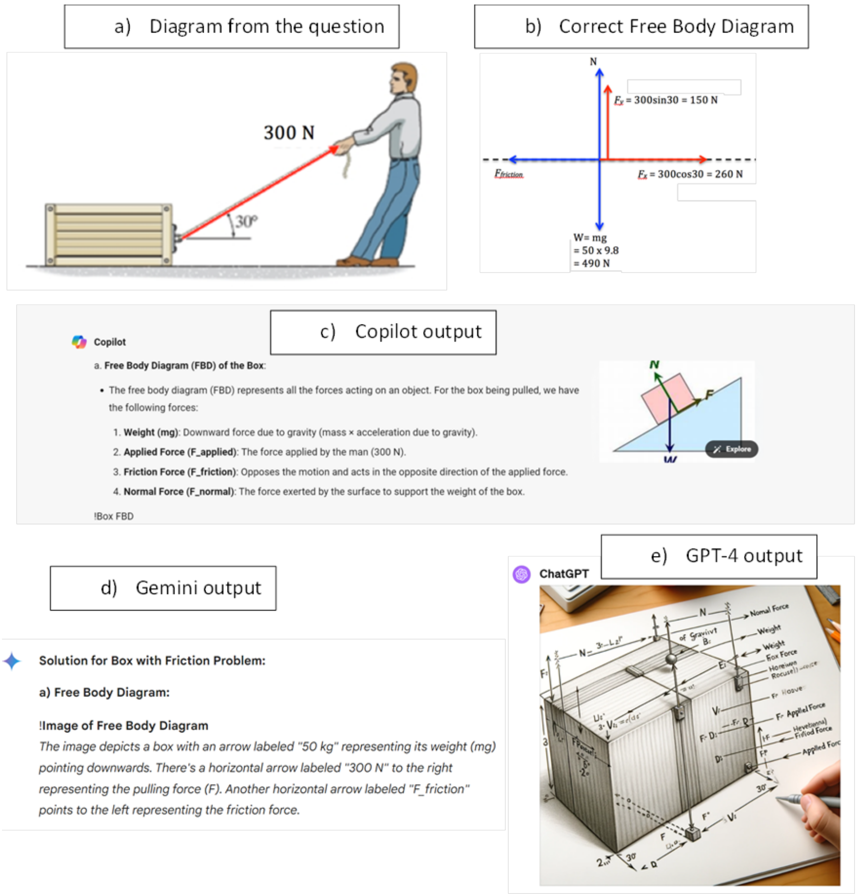
Figure 4. Example of a Physics question that GenAI was partially able to answer correctly. GenAI was able to correctly find the values but made some miscalculations in terms of vector direction

Uploading Documents

ChatGPT-4 also excelled in uploading documents, going beyond images and providing versatility in type. Its ability to upload and analyse PDF documents was especially beneficial. For example, journal articles or other information could be uploaded, and precise summaries could be achieved instantly. Alternatively, it is possible to analyse a document against a specific framework. As another use case, a CV could be uploaded with a job advertisement, and responses to selection criteria that merged a student's evidence with the job requirements would be easily created. Together, these examples highlight that non-trained data could be uploaded to be analysed separately or used together with trained data. Therefore, this approach can create a workaround for some limitations identified where performance was not ideal when dealing with recent or non-trained data. As the number of file formats and types available to upload increases, the capability of GenAI to have an impact on assessment security will increase. The problem at large, however, is centred on the copyright, privacy and ethical uses of data and what can or should not be uploaded (Daniel and Nikolic, 2023).

Generating images

In engineering, drawing diagrams is a common task. In our study, we tested the ability of AI tools to generate images for physics questions that required students to draw a diagram as part of the exam answer. Students were provided with a simple diagram of a man pulling a box (Figure 5). The first part of the question was to draw a Free Body Diagram, which is commonly used in mechanics problems to visualise forces. While this task is relatively simple, none of the AI tools were able to draw the Free Body Diagram correctly. Gemini provided written instructions about the arrows that should be included in the diagram, but it still missed one force. Copilot provided an image from Wikipedia of a Free Body Diagram that was not applicable to the question. GPT-4 generated a futuristic image that looked similar but did not have anything in common with the mechanics of a Free Body Diagram.



Copilot

a. **Free Body Diagram (FBD) of the Box**:

- The free body diagram (FBD) represents all the forces acting on an object. For the box being pulled, we have the following forces:
  1. **Weight (mg)**: Downward force due to gravity (mass × acceleration due to gravity).
  2. **Applied Force (F_applied)**: The force applied by the man (300 N).
  3. **Friction Force (F_friction)**: Opposes the motion and acts in the opposite direction of the applied force.
  4. **Normal Force (F_normal)**: The force exerted by the surface to support the weight of the box.

!Box FBD

Solution for Box with Friction Problem:

a) Free Body Diagram:

!Image of Free Body Diagram

The image depicts a box with an arrow labeled "50 kg" representing its weight (mg) pointing downwards. There's a horizontal arrow labeled "300 N" to the right representing the pulling force (F). Another horizontal arrow labeled "F_friction" points to the left representing the friction force.

Figure 5. Example of physics question diagram (a), correct Free Body Diagram (b) and generated AI outputs: Copilot (c), Gemini (d), GPT-4 (d)

In the Manufacturing Technology course, the capability of GenAI to develop diagrams was also tested. Students were provided with some data and asked to develop different control charts. Only Gemini and GPT-4 were able to produce control charts amongst all the GenAI platforms that were tested in this study. GPT-4 developed a perfect chart (Figure 6) which accurately represented all the data.

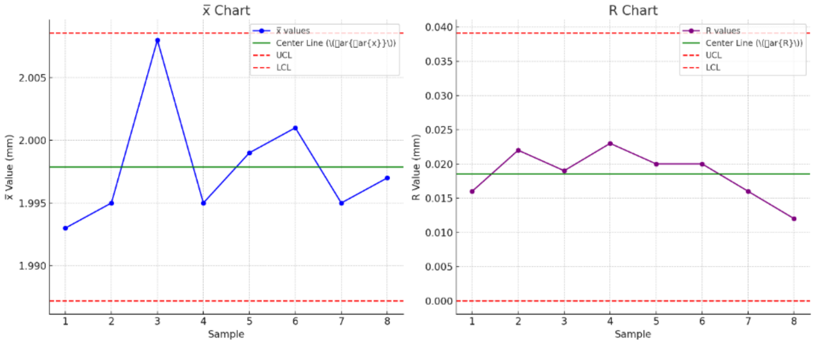


Figure 6. Sample charts produced by GPT-4 in response to a Manufacturing Technology question with given data.

ChatGPT-4o

ChatGPT-4o was released on May 23, 2024, while this paper was undergoing review. ChatGPT-4o is a slightly updated version that was trained as a single new model end-to-end across text, vision, and audio, meaning that all inputs and outputs are processed by the same neural network, resulting in similar performance on traditional tasks, but improved performance on multilingual, audio, and vision capabilities (OpenAI, 2024a). Due to this possible change in image performance, we retested some of the image questions. On the test sample, it could now understand and calculate Figure 2b correctly, but still struggled with other complex images. For Figure 3, the calculation remains exactly the same as the version of GPT-4 we tested in February 2024. While getting the answers wrong, it was observed that GPT-4o was getting more accurate in translating the images and setting them up into appropriate equations. If a student had some understanding, they could identify the interpretation errors and correct the equations to get the correct results. Improvements in this area will only continue. It is recommended that further research is carried out to investigate the multimodal opportunities available.

Platform Choices, Limitations and Ethics

This study explored various GenAI tools but did not include all possibilities, particularly paid versions like Microsoft Copilot Pro, Google's Gemini Pro, and Anthropic's Claude AI. The main reason was workload constraints. Each author, acting as a subject coordinator, handled the evaluations without research assistants, limiting the capacity to analyse multiple platforms. Only OpenAI's ChatGPT-4 was selected as a paid version due to its popularity and large user base. However, during the review period, ChatGPT-4o was released with limited free access. The free versions of ChatGPT-3.5, Copilot, and Gemini were tested, reflecting the starting point for most students due to financial constraints and providing a benchmark of lower capabilities. While Copilot Pro and Gemini Pro likely offer better performance, this limitation should be considered in the results.

Another minor limitation is the blurred boundaries between different assessment types. Simplifying the analysis into distinct categories aids comparison but doesn't always fit neatly. For instance, tasks like generating a stakeholder persona involve research, creativity, and graphical skills. Similarly, creating a visual stakeholder map requires prior research. Readers should align their own tasks with the assessment types discussed in the paper to gain relevant insights.

It could be considered that the potential variance in our interpretation of the GenAI outputs is a limitation. However, seeing as we were each evaluating GenAI outputs for assessments in the same subjects that we coordinate at our respective institutions, we are, in fact, the best-placed people to make these interpretations as we are making them every day with our own students.

GenAI tools enhance content understanding through extended prompting. Using the same chat thread for each subject could evoke stronger responses due to accumulated context, impacting performance. Ethical considerations highlighted by Daniel and Nikolic (2023) emphasise equity and accessibility, ensuring no student is disadvantaged. While free platforms offer some benefits, the results show a clear advantage in using paid products. Access to different platforms can create inequities, particularly concerning the allowable use and integration of GenAI in educational settings.

The GenAI Assessment Security & Opportunity Matrix

To provide a practical resource for the academic community, the outcomes from this study have been synthesised in Table 10. This table adapts an engineering risk register methodology [TEQSA, 2024b] to summarise the impact of GenAI on different assessment types, and then propose short and long-term assessment security options, as well as the opportunities to embrace integration for learning. Greater detail on how the risk and options were developed is given in the Supplementary Materials.

Table 10. The Generative AI Assessment Security & Opportunity Matrix

Assessment Type	GenAI Impact	Risk	Short-Term Security Options	Long-Term Security Options	Opportunity	Use Case
Class Participation	GenAI cannot take the place of a student, but it can serve as a tool to support their learning activities.	Low	Integrate GenAI into class participation.	Integrate GenAI into class participation.	Class participation allows for active engagement between the teaching staff, students and learning content. GenAI can be used in structured activities for critical thinking and exploration.	Use subjectively to consider competencies, including communication, participation, knowledge and affective skills.
Comprehension	GenAI has the capability to handle multiple digital formats, analyse information and provide the comprehension required.	High	To confirm text-based comprehension, it is best to use non-graded activities or activities conducted in a supervised environment. Comprehension of audio/video content is a possible short-term opportunity.	Required to be completed in supervised environments, as the ability to analyse larger, more complex text and audiovisual sources will only increase.	GenAI is a perfect tool for self-evaluation to confirm understanding of what they read, heard or said, providing immediate feedback. Such feedback can aid learning.	Use to develop comprehension and critical-thinking skills
Essay	GenAI can be used to write the essay. Can be used for almost any topic with the right prompting.	Very High	It is best used for a learning-driven, non-graded activity. If assessment is needed, it must be in a supervised environment. Consider mandating a unique structure.	It is best used for a learning-driven, non-graded activity. If assessment is needed, it must be in a supervised environment. Consider mandating a unique structure.	GenAI can help students develop ideas, structure, and evidence and provide feedback. Such feedback can aid learning.	Use to develop critical-thinking, communication, analysis, synthesis and evaluation skills.
Evaluation (Self or Peer)	GenAI tools offered solid, concise feedback and reflections as a starting point for personalisation, provide frameworks guiding users on how to refine their feedback for copy/paste.	High	Pair evaluations with evidence (e.g. image of team tools, conversations, outcomes etc.) that require a degree of critical thinking for the student to decide what is best to include.	Consider making such evaluations a part of the process rather than a marked assessment.	Use GenAI as a tool to help students better structure and consider the feedback they provide. Doing so can help them structure their thoughts and gain awareness of the feedback they need to learn to provide.	Reflections can be tied to critical-thinking, emotional intelligence, meta-cognition and self-assessment.
Exam or Quiz (In person)	GenAI can be used if concealed technologies are used. For example, hidden video feed and microphone to relay with GenAI, short-term via another person, long-term directly.	Medium	Ensure the identification of the participant. Watch for suspicious behaviour.	Ensure the identification of the participant. Watch for suspicious behaviour.	For complex content, GenAI can be used as a tool, much in the same way as calculators are used today after students have an understanding of what is being calculated.	Use to confirm problem-solving, knowledge, comprehension, application, analysis, synthesis and evaluation competencies.

Exam or Quiz (Online)	Potential to copy/paste questions or answer directly using a browser extension. High passing rate in answering most forms of quiz questions/context. Risks have increased with the ability to upload files and images for analysis. Deepfakes are of growing concern.	Very High	Image analysis is still at an introductory level. Therefore, use questions that incorporate the analysis of complex figures and tables. Most GenAI tools have limitations when images include people.	Over time all approaches will be vulnerable. Quizzes should only be used for low-stake assessments or questions that embed GenAI as part of the solution.	GenAI can provide support in generating online quiz questions, especially multiple choice. Questions can be altered so that using GenAI becomes a part of the exam.	**Low-Security:** Use to help students remember, understand and apply knowledge **High-Security:** High-stakes online exams can still be completed in a supervised computer-based environment.
Experimentation based	GenAI can be used to produce laboratory-styled reports. It is easiest when data is available or if paraphrasing an existing report is required.	Medium	Without data collection, major prompt engineering would be required. Therefore, do not repeat experiments.	Experimentation offers many opportunities for greater security, including practical examination, observation, interview, and visual evidence such as videos to demonstrate competencies (although such videos may also become compromised with the rise of deepfakes).	In industry, using technology to support error correction/resolving faults is commonplace. Such activities are also very supportive of learning. Some components of experimentation can be altered to focus on fault-finding, using GenAI as a tool to suggest solutions.	Non-written assessment forms allow for greater assessment of psychomotor and affective competencies available to be extracted from experimentation.
Numerical (Assignments and Exams)	Many free versions perform poorly, but GPT-4 has become somewhat reliable. Students can copy/paste answers from GenAI with detailed explanations.	Medium	Stronger on some topics than others, some specific testing can reveal the risk ratio. Include images and tables to make questions more complex and difficult to prompt.	Unsupervised assessments will provide little evidence of competency. If the pace of improvement continues, all mathematical problem-solving, including the use of images and tables, will be at high risk. Supervised assessments will be required or assessments that integrate GenAI use.	Support for math and problem-solving has often been difficult and expensive to obtain. Using GenAI as part of the learning process to help develop understanding can be of great benefit.	**Low-Security:** Non/low graded activities, used to develop computational and problem-solving skills. **High-Security:** When computational and problem-solving capability competencies need confirmation.
Oral Interview	GenAI can provide students with a range of preparation questions for practice. These answers can be memorised. For online, demonstrations of GenAI receiving real-time questions and providing immediate talking points have been demonstrated. Use of deepfakes is a growing possibility.	Low	The likelihood of real-time online use or deepfakes is low. Therefore, F2F or online can be used.	For online interviews, GenAI will eventually be able to intercept and provide support, and deepfake technology will become reliable and accessible. Longer-term, direct impersonation could be possible. F2F would be the most secure option.	Students can use GenAI to practice directly. ChatGPT-4o will soon provide direct human-to-GenAI voice communication. Such practice would help enable learning.	Interviews provide an opportunity to check for oral communication competencies and understanding. The randomness of questions and the ability to delve deeper when required makes this a secure assessment type.

Oral Presentations	GenAI can help students develop a detailed and professional script. This can be prompt engineered or based on existing work via a document upload. It also has the potential to help develop presentation slides.	Medium	GenAI cannot undertake the presentation for the student but can provide a memorisable script.	A presentation would be a security risk if the key outcome was to check student understanding. However, giving more weight to the ensuing impromptu Q&A may offset this.	Learning to present is a skill in itself. Let GenAI help students refine their speeches and optimise their presentations. Feedback on students' presentation skills has generally been limited.	**Low Security:** Use where developing presentation skills is the core learning outcome and/or building confidence in public speaking is a critical objective. **High-Security:** Use when testing for understanding is needed, complement with an interview.
Portfolio	GenAI can produce some, but not all evidence. It can write the reflective pieces associated with the evidence.	Medium	Require a diverse range of evidence, including photos of experiences, certification, reference letters, etc. The focus needs to be on the critical-thinking connection between the requirements, evidence and writing.	Deepfakes could potentially falsify most evidence and create a connection between requirements and evidence. Security may be a long-term issue.	GenAI can help students identify which evidence is best suited to different competencies. It can also help them reflect and map out a long-term evidence collection plan.	The portfolio can be used to help develop critical-thinking, reflection, creativity, professional and discipline competencies.
Programming	GenAI can generate solutions or approaches to solve them.	High	Ensure that there are external files that are required to be analysed for writing the code for the problem.	Over time, GenAI will improve and be able to generate code which can be improved by better prompting. The focus of the assessment needs to shift towards "computational thinking" rather than coding only.	GenAI will be a tool for generating the initial code, which can be improved by prompting or tailoring it. The focus will shift to computational thinking, i.e., solving problems using computing.	Programming can be linked to computational thinking to assess critical-thinking, analytical and problem-solving skills.
Project Reports	GenAI is well suited to components of project work such as identifying WHS risks, relevant standards, writing and paraphrasing.	Medium	Embed the GenAI strengths into a more holistic assessment rubric that requires greater critical thinking, synthesis of ideas and evidence beyond GenAI capability. Solution expectations can increase.	This will remain a balancing act between identifying the components that GenAI can do well and those that are beyond its reach.	Project work can benefit from GenAI integration if its strengths and weaknesses are celebrated and taken advantage of. Use GenAI to help with brainstorming, feedback, and identifying solutions, but raise the goal post of expectations because students will be capable of more.	Project work can be tied with higher-order cognitive skills (analysing, evaluating, creating), as well as psychomotor and affective skills. An opportunity to connect a variety of assessment tasks to consider a multitude of learning competencies

Reflections	GenAI can use any reflection framework to generate passable work. Newer GenAI tools recognise poor prompting and suggest the extra context needed for a better output.	High	Connect the reflections with a portfolio of evidence that GenAI can't produce. This may include photos, videos and certificates	Deepfakes of portfolio evidence may make reflections an important part of the learning process, but not the assessment process. Alternatively, they can be assessed in a high-security assessment environment.	Use GenAI as a tool to help students better appreciate how they can use reflective frameworks to improve their understanding of the benefits of their learning activities.	Reflections can be tied to critical-thinking, emotional intelligence, meta-cognition and self-assessment.
Research Writing	Heavily dependent on the GenAI used, modern GenAI tools are capable of finding relevant and recent research, paraphrasing, summarising, synthesising and producing a suitable end product. The end product is reliant on the student following a process, but how-to guides are available.	High	It takes some high-end prompt engineering to get the output to shape to a specific structure. Therefore, consider defining a specific structure that is unique and contrary to the expected writing structure used by GenAI.	GenAI will improve at identifying and understanding unique structures and incorporating them into the output without the need for advanced prompt engineering. Supplementary assessments such as interviews and oral vivas will aid reassurance. The focus needs to shift to the experimentation associated with research.	GenAI can assist students throughout the entire research process, including brainstorming, identifying suitable articles, understanding supporting articles and data analysis, extracting and summarising key information, synthesising information, paraphrasing, conducting all writing, and providing feedback. Such feedback can improve learning.	Research supports a diverse range of skills, including communication, critical thinking, data analysis, and project and time management. The writing component will become a measure of evaluative judgment.
Video	Video can be created and manipulated by GenAI.	Medium	Ensure that the student plays some role in the video.	Over time deepfakes and creative generations of scripts provided by students will change the security profile. The focus of assessment will need to gain a greater focus on the creative elements (the evaluative judgment of the script and storytelling).	Digital communication is a key 21st-century skill. GenAI can unleash a student's creative potential to convey information in the most effective and entertaining way. The evaluative judgment of the storytelling will be of focus.	Video can be used to assess communication, critical-thinking and creativity

Conclusion:

This study attempted to answer two research questions. The first was, "*Has the performance of generative artificial intelligence tools improved so that it can pass students across more assessment types?*". In our benchmark study, GenAI could more likely than not achieve a passing grade for the online quiz, numerical, programming and written (reflection & critical thinking) assessment types. While visual, written (project-based) and written (research) assessment types were found secure. The assessment types, oral and written (experimentation), were classified as a tie due to the selected assessment designs and GenAI's ability to assist with specific components. This study has shown that the performance of GenAI has indeed improved when considering the performance of ChatGPT-4 over ChatGPT-3.5. On average, for the online quiz, performance increased 24% and for numerical, 41%. This represents a change from a pass to a credit. What new capability will ChatGPT-5 bring? For the given assessment tasks, programming performance was much the same, primarily due to the specific CSV-based project continuing to cause problems for GenAI. For written (reflection & critical thinking), the tools became even more user-friendly, providing a helping hand to provide the much-needed context. The assessment security held by written (research) was undone by greater literature access and lower hallucination rates. The performance of visual, written (project-based), oral and written (experimentation) assessment types was steady. As shown through multiple examples, the early capability of the GenAI tools (especially ChatGPT-4) to upload images and documents and generate diagrams has been a substantial stepping stone, transforming the opportunity to more easily engage and benefit from GenAI.

The second research question was, "*Which of the major generative artificial intelligence tools are best suited to each assessment type?*". From the limited sample of technologies tested, this study has shown that there is no one-stop solution. Each GenAI tool excelled in different ways as documented across the data tables. Further, there can be great benefits to using multiple GenAI tools, especially for project work, where diverse insights can support brainstorming and solution generation. However, ChatGPT-4 is a reliable tool for most engineering applications, especially when used with GPTs (plugins) that help it do more. This shows the possible advantage given to those who can afford to use paid GenAI's compared to those who cannot and the potential for inequity that this brings. We note that just prior to publication, free access to the newly released ChatGPT-4o has become available, allowing all students access to high performance.

Beyond the risks to academic integrity, the study once again provided recommendations on how GenAI tools can support teaching and learning. Feedback, brainstorming, project and critical thinking-based activities are supportive. Tutor-based support, especially for problem-solving, shows tremendous promise, something that tends to be difficult for students to receive without paying for expensive face-to-face tutorial sessions. The time is right to encourage innovative teaching and learning assessments and practices that can work alongside or integrate GenAI. Doing so may allow for a greater appreciation of the diverse activities, competencies and graduate attributes associated with engineering education (Crossin et al., 2023).

Masked in the numbers is the fact that the more complex the task, the greater the need for critical thinking capability, the more a student needs to understand what a quality output should look like, and the subsequently higher order prompt engineering is required to get there (Nikolic et al., 2023a, Nweke et al., 2023, Bearman et al., 2024). That is, just because GenAI can provide the correct output, there is no guarantee that the student understands what that should be or could get GenAI to produce it.

The introduction discussed how cheating is not new. Regardless of the assessment medium, some students will always find a strategy to cheat in some way if they are determined enough. In recent years, contract cheating or file-sharing services have supported at least 10% of students (Curtis et al., 2022), and there are technological solutions that work around proctoring software (Burgess et al., 2022, Bergmans et al., 2021). Technologically skilled students (e.g. engineering students) could have the capability to work around such software (Dawson, 2024). This confirms that assessment practices have been unsecure for some time, but possibly low-impacting enough to be handled with care. With contract cheating posing a transaction risk, for example, the possibility of users being subject to future blackmail events (Groch, 2024), its attractiveness to the masses has always been limited. However, as GenAI becomes mainstream, it can create a comfort zone level of trust between students and technology (especially as it gets more reliable and integrated into everyday products), where the risk-reward balance may tempt many students towards gentle and then more serious academic integrity breaches. The more intertwined technology gets with our lives, many breaches may become unintentional. Already demonstrated in published studies is that some PhD students and academics use GenAI in ways that some may not be considered appropriate. In our benchmark study in 2023, we stated that we had 12 – 24 months before GenAI's capability created a serious threat. Twelve months have passed, and this paper has shown that GenAI capability has leapt ahead. While the precise advancements of the next year remain uncertain, GenAI is constantly evolving. Therefore, organisations that have not yet undergone a security audit are strongly encouraged to consider initiating one now to ensure their current safeguards remain effective.

Now is also the time for education institutions to holistically consider their options, such as those presented in the educational integrity enforcement pyramid presented by Ellis and Murdoch (2024). Those who try to ignore it will likely face significant future challenges.

Acknowledgement

This work is an initiative of the Australasian Artificial Intelligence in Engineering Education Centre (AAIEEC), a special interest group of the Australasian Association of Engineering Education (AAEE). The 2023 AAEE Engineering Education Australian Council of Engineering Deans Aligned Grant provided licences for access to the paid version of ChatGPT-4.

GenAI was used throughout this study as outlined in the method to generate the data for analysis. Furthermore, a limited use of GenAI was made for brainstorming and helping to succinctly reframe some sentences or ideas produced by the authors.

References:

ABD-ELAAL, E.-S., GAMAGE, S. H. P. W. & MILLS, J. E. 2022. Assisting academics to identify computer generated writing. *European Journal of Engineering Education,* 47**,** 725-745.

ABET. 2014. *Why Accreditation Matters* [Online]. Available: http://www.abet.org/why-accreditation-matters/ [Accessed Dec 23 2014].

ABEYSEKERA, I. 2024. ChatGPT and academia on accounting assessments. *Journal of Open Innovation: Technology, Market, and Complexity,* 10**,** 100213.

ADAMOPOULOU, E. & MOUSSIADES, L. An overview of chatbot technology. Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II 16, 2020. Springer, 373-383.

AJJAWI, R., TAI, J., DOLLINGER, M., DAWSON, P., BOUD, D. & BEARMAN, M. 2023. From authentic assessment to authenticity in assessment: broadening perspectives. *Assessment & Evaluation in Higher Education***,** 1-12.

ALSHAMI, A., ELSAYED, M., ALI, E., ELTOUKHY, A. E. & ZAYED, T. 2023. Harnessing the power of ChatGPT for automating systematic review process: Methodology, case study, limitations, and future directions. *Systems,* 11**,** 351.

BADER, R., IMAM, A., ALNEES, M., ADLER, N., ZUGAYAR, D., DAN, A. & KHALAILEH, A. 2024. Successful management of an Iatrogenic portal vein and hepatic artery injury in a 4-month-old female patient: A case report and literature review. *Radiology Case Reports,* 19**,** 2106-2111.

BAHROUN, Z., ANANE, C., AHMED, V. & ZACCA, A. 2023. Transforming education: A comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis. *Sustainability,* 15**,** 12983.

BEARMAN, M., NIEMINEN, J. H. & AJJAWI, R. 2023. Designing assessment in a digital world: an organising framework. *Assessment & Evaluation in Higher Education,* 48**,** 291-304.

BEARMAN, M., TAI, J., DAWSON, P., BOUD, D. & AJJAWI, R. 2024. Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education***,** 1-13.

BENTE, S., RANDALL, N. & WÄCKERLE, D. A Conceptual Framework to Transform Coding Education in Times of Generative AI. Software Engineering im Unterricht der Hochschulen 2024, 2024. Gesellschaft für Informatik eV, 93-104.

BERGMANS, L., BOUALI, N., LUTTIKHUIS, M. & RENSINK, A. On the efficacy of online proctoring using proctorio. 13th International Conference on Computer Supported Education, CSEDU 2021, 2021. Scitepress, 279-290.

BIANCHI, T. 2024. *Global market share of leading desktop search engines 2015-2024* [Online]. Available: https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/ [Accessed 03/06/2024].

BORSOTTI, V., BEGEL, A. & BJORN, P. 2024. Neurodiversity and the Accessible University: Exploring Organizational Barriers, Access Labor and Opportunities for Change.

BRIN, D., SORIN, V., VAID, A., SOROUSH, A., GLICKSBERG, B. S., CHARNEY, A. W., NADKARNI, G. & KLANG, E. 2023. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Scientific Reports,* 13**,** 16492.

BURGESS, B., GINSBERG, A., FELTEN, E. W. & COHNEY, S. Watching the watchers: bias and vulnerability in remote proctoring software. 31st USENIX Security Symposium (USENIX Security 22), 2022. 571-588.

CHAN, C., JIAYANG, C., YIM, Y., DENG, Z., FAN, W., LI, H., LIU, X., ZHANG, H., WANG, W. & SONG, Y. 2024. NegotiationToM: A Benchmark for Stress-testing Machine Theory of Mind on Negotiation Surrounding. *arXiv preprint arXiv:2404.13627*.

CHUNG, H.-H., CHUNG, F.-L., LIN, S.-M. & LAN, Y.-J. 2023. Tools and Approaches of Generative Artificial Intelligence Used in Education. *31st International Conference on Computers in Education*. Japan.

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.

CROMPTON, H. & BURKE, D. 2024. The Educational Affordances and Challenges of ChatGPT: State of the Field. *TechTrends,* 68**,** 380-392.

CROSSIN, E., RICHARDS, J. I., DART, S. & NASWALL, K. 2023. A taxonomy of common engineering activities and competencies. *Australasian Journal of Engineering Education,* 28**,** 181-193.

CURTIS, G. J., MCNEILL, M., SLADE, C., TREMAYNE, K., HARPER, R., RUNDLE, K. & GREENAWAY, R. 2022. Moving beyond self-reports to estimate the prevalence of commercial contract cheating: an Australian study. *Studies in Higher Education,* 47**,** 1844-1856.

DANESH, A., PAZOUKI, H., DANESH, K., DANESH, F. & DANESH, A. 2023. The performance of artificial intelligence language models in board-style dental knowledge assessment: A preliminary study on ChatGPT. *The Journal of the American Dental Association,* 154**,** 970-974.

DANIEL, S. & NIKOLIC, S. 2023. *Benchmarking AI tools and assessing integrity: Assessment integrity in the AI age* [Online]. Editorial, SEFI Ethics SIG. Available: https://www.sefi.be/2023/10/14/benchmarking-ai-tools-and-assessing-integrity-assessment-integrity-in-the-ai-age/ [Accessed].

DAWSON, P. 2016. Five ways to hack and cheat with bring-your-own-device electronic examinations. *British Journal of Educational Technology,* 47**,** 592-600.

DAWSON, P. 2020. *Defending assessment security in a digital world: Preventing e-cheating and supporting academic integrity in higher education*, Routledge.

DAWSON, P. 2024. Remote Proctoring: Understanding the Debate. *Second Handbook of Academic Integrity.* Springer.

DUNNE, I. & NIKOLIC, S. Autonomous Assessment of a Laboratory Exam for the Digital Hardware Curriculum. 2021 IEEE International Conference on Engineering, Technology & Education (TALE), 2021. IEEE, 829-833.

ELKHODR, M., GIDE, E., WU, R. & DARWISH, O. 2023. ICT students' perceptions towards ChatGPT: An experimental reflective lab analysis. *STEM Education,* 3**,** 70-88.

ELLIS, C. & MURDOCH, K. 2024. The educational integrity enforcement pyramid: a new framework for challenging and responding to student cheating. *Assessment & Evaluation in Higher Education***,** 1-11.

ENGINEERS AUSTRALIA 2008. G02 Accreditation Criteria Guidelines. *Education Programs at the level of Professional Engineer.*

FARAZOULI, A., CERRATTO-PARGMAN, T., BOLANDER-LAKSOV, K. & MCGRATH, C. 2023. Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices. *Assessment & Evaluation in Higher Education***,** 1-13.

FENWICK, M. & JURCYS, P. 2023. Originality and the future of copyright in an age of generative AI. *Computer Law & Security Review,* 51**,** 105892.

GRATCHEV, I., HOWELL, S. & STEGEN, S. 2024. Academics' perception of final examinations in engineering education. *Australasian Journal of Engineering Education***,** 1-10.

GROCH, S. 2024. These students cheated on a test and got away with it. Then the blackmail started. *The Sydney Morning Hearld*, 29/03.

HALLAL, K., HAMDAN, R. & TLAIS, S. 2023. Exploring the potential of AI-Chatbots in organic chemistry: An assessment of ChatGPT and Bard. *Computers and Education: Artificial Intelligence,* 5**,** 100170.

HARGREAVES, D. J. 1997. Student Learning and Assessment Are Inextricably Linked. *European Journal of Engineering Education,* 22**,** 401-409.

HERBOLD, S., HAUTLI-JANISZ, A., HEUER, U., KIKTEVA, Z. & TRAUTSCH, A. 2023. A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports,* 13**,** 18617.

HUANG, J. & TAN, M. 2023. The role of ChatGPT in scientific communication: writing better scientific review articles. *American journal of cancer research,* 13**,** 1148.

IANNONE, P. & SIMPSON, A. 2015. Students' views of oral performance assessment in mathematics: straddling the 'assessment of' and 'assessment for' learning divide. *Assessment & Evaluation in Higher Education,* 40**,** 971-987.

LEE, M. J. W., NIKOLIC, S., VIAL, P. J., RITZ, C., LI, W. & GOLDFINCH, T. 2016. Enhancing project-based learning through student and industry engagement in a video-augmented 3-D virtual trade fair. *IEEE Transactions on Education,* 59**,** 290 - 298.

LINGEFJÄRD, T. 2024. Empowering mathematics education through programming. *Journal of Mathematics and Science Teacher,* 4.

LÓPEZ ESPEJEL, J., ETTIFOURI, E. H., YAHAYA ALASSAN, M. S., CHOUHAM, E. M. & DAHHANE, W. 2023. GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing Journal,* 5**,** 100032.

MAI, D. T. T., DA, C. V. & HANH, N. V. 2024. The use of ChatGPT in teaching and learning: a systematic review through SWOT analysis approach. *Frontiers in Education,* 9.

MENEKSE, M. 2023. Envisioning the future of learning and teaching engineering in the artificial intelligence era: Opportunities and challenges. *Journal of Engineering Education,* 112**,** 578-582.

MIAO, G., RANARAJA, I., GRUNDY, S., BROWN, N., BELKINA, M. & GOLDFINCH, T. 2024. Project-based learning in Australian & New Zealand universities: current practice and challenges. *Australasian Journal of Engineering Education***,** 1-13.

NEWELL, S. J. 2023. Employing the interactive oral to mitigate threats to academic integrity from ChatGPT. *Scholarship of Teaching and Learning in Psychology*.

NIKOLIC, S., DANIEL, S., HAQUE, R., BELKINA, M., HASSAN, G. M., GRUNDY, S., LYDEN, S., NEAL, P. & SANDISON, C. 2023a. 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**,** 559-614.

NIKOLIC, S., GRUNDY, S., HAQUE, R., LAL, S., HASSAN, G. M., DANIEL, S., BELKINA, M., LYDEN, S. & SUESSE, T. F. 2023b. A ranking comparison of the traditional, online and mixed laboratory mode learning objectives in engineering: Uncovering different priorities. *STEM Education,* 3**,** 331-349.

NIKOLIC, S., ROS, M., JOVANOVIC, K. & STANISAVLJEVIC, Z. 2021. Remote, Simulation or Traditional Engineering Teaching Laboratory: A Systematic Literature Review of Assessment Implementations to Measure Student Achievement or Learning. *European Journal of Engineering Education,* 46**,** 1141-1162.

NIKOLIC, S., SUESSE, T. F., GRUNDY, S., HAQUE, R., LYDEN, S., HASSAN, G. M., DANIEL, S., BELKINA, M. & LAL, S. 2023c. Laboratory learning objectives: ranking objectives across the cognitive, psychomotor and affective domains within engineering. *European Journal of Engineering Education,* 48**,** 559-614.

NOGUES, C. P. & DORNELES, B. V. 2023. School interventions for students with ADHD and learning difficulties: a mini review. *Global Journal of Intellectual & Developmental Disabilities. Irvine, CA. Vol. 11, n. 2 (Jan./2023), 555806, p. 1-2*.

NOORBEHBAHANI, F., MOHAMMADI, A. & AMINAZADEH, M. 2022. A systematic review of research on cheating in online exams from 2010 to 2021. *Education and Information Technologies,* 27**,** 8413-8460.

NWEKE, M. C., BANNER, M. & CHAIB, M. An Investigation Into ChatGPT Generated Assessments: Can We Tell the Difference? The Barcelona Conference on Education 2023: Official Conference Proceedings, 2023. The International Academic Forum (IAFOR), 1-6.

OPENAI. 2023. *ChatGPT plugins* [Online]. OpenAI. Available: https://openai.com/blog/chatgpt-plugins [Accessed 25/02 2024].

OPENAI. 2024a. *Hello GPT-4o* [Online]. OpenAI. Available: https://openai.com/index/hello-gpt-4o/ [Accessed 03/04/2024].

OPENAI. 2024b. *Models* [Online]. Available: https://platform.openai.com/docs/models [Accessed 03/06/2024].

OUH, E. L., GAN, B. K. S., JIN SHIM, K. & WLODKOWSKI, S. ChatGPT, Can You Generate Solutions for my Coding Exercises? An Evaluation on its Effectiveness in an undergraduate Java Programming Course. Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1, 2023. 54-60.

PICHAI, S. & HASSABIS, D. 2024. *Our next-generation model: Gemini 1.5* [Online]. Google. Available: https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#sundar-note [Accessed 25/02 2024].

PINZOLITS, R. 2024. AI in academia: An overview of selected tools and their areas of application. *MAP Education and Humanities,* 4**,** 37-50.

PURTILL, J. 2023. ChatGPT was tipped to cause widespread cheating. Here's what students say happened. *ABC Australia*, 22/11/2023.

RAZA, M. R. & HUSSAIN, W. Preserving Academic Integrity in Teaching with ChatGPT: Practical Strategies. 2023 IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2023. IEEE, 158-162.

ROGERSON, A. M. & MCCARTHY, G. 2017. Using Internet based paraphrasing tools: Original work, patchwriting or facilitated plagiarism? *International Journal for Educational Integrity,* 13**,** 1-15.

RUDOLPH, J., TAN, S. & TAN, S. 2023. War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning and Teaching,* 6.

SÁNCHEZ-RUIZ, L. M., MOLL-LÓPEZ, S., NUÑEZ-PÉREZ, A., MORAÑO-FERNÁNDEZ, J. A. & VEGA-FLEITAS, E. 2023. ChatGPT challenges blended learning methodologies in engineering education: A case study in mathematics. *Applied Sciences,* 13**,** 6039.

SHANMUGAM, S., VELOO, A. & YUSOFF, Y. A. B. J. 2024. Examining Utility of Oral-Administered Test Accommodation in Assessing Aboriginal Pupils' Mathematics Performance using Score Comparability. *International Journal of Science and Mathematics Education***,** 1-24.

SHANTO, S. S., AHMED, Z. & JONY, A. I. 2023. PAIGE: A generative AI-based framework for promoting assignment integrity in higher education. *STEM Education,* 3**,** 288-305.

SHERIF, A. 2024. *Global market share held by computer operating systems 2012-2024* [Online]. Available: https://www.statista.com/statistics/268237/global-market-share-held-by-operating-systems-since-2009/ [Accessed 03/06/2024].

SHOAIB, M. R., WANG, Z., AHVANOOEY, M. T. & ZHAO, J. Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models. 2023 International Conference on Computer and Applications (ICCA), 28-30 Nov. 2023 2023. 1-7.

SPATARO, J. 2023. *Announcing Microsoft 365 Copilot general availability and Microsoft 365 Chat* [Online]. Microsoft. Available: https://www.microsoft.com/en-us/microsoft-365/blog/2023/09/21/announcing-microsoft-365-copilot-general-availability-and-microsoft-365-chat/ [Accessed 25/02 2024].

STREET, W., SIY, J. O., KEELING, G., BARANES, A., BARNETT, B., MCKIBBEN, M., KANYERE, T., LENTZ, A. & DUNBAR, R. I. 2024. LLMs achieve adult human performance on higher-order theory of mind tasks. *arXiv preprint arXiv:2405.18870*.

TERTIARY EDUCATION QUALITY AND STANDARDS AGENCY. 2022. *What is academic integrity?* [Online]. Available: https://www.teqsa.gov.au/students/understanding-academic-integrity/what-academic-integrity [Accessed 28/02 2024].

TERTIARY EDUCATION QUALITY AND STANDARDS AGENCY. 2024. *Higher Education Standards Framework (Threshold Standards) 2021* [Online]. Available: https://www.teqsa.gov.au/how-we-regulate/higher-education-standards-framework-2021 [Accessed 28/02 2024].

VÁZQUEZ-CANO, E., RAMÍREZ-HURTADO, J. M., SÁEZ-LÓPEZ, J. M. & LÓPEZ-MENESES, E. 2023. ChatGPT: The brightest student in the class. *Thinking Skills and Creativity,* 49**,** 101380.

WOLFRAM, S. 2023. *ChatGPT Gets Its "Wolfram Superpowers"!* [Online]. Available: https://writings.stephenwolfram.com/2023/03/chatgpt-gets-its-wolfram-superpowers/ [Accessed 23/03].

ZHANG, M., WU, L., YANG, T., ZHU, B. & LIU, Y. 2024. The three-dimensional porous mesh structure of Cu-based metal-organic-framework - aramid cellulose separator enhances the electrochemical performance of lithium metal anode batteries. *Surfaces and Interfaces,* 46**,** 104081.