Abstract

ChatGPT sent shockwaves throughout the education sector when it was discovered it could pass exams. It raised many questions about the authenticity of assessment and challenges in detecting plagiarism. Beyond the fear, there were hints of potential opportunities in how ChatGPT could support learning and the development of critical thinking. However, there was little empirical data to support the fears and opportunities, especially in engineering education. Most research was based on opinion, except for the key studies that had tested ChatGPT against exams. What was lacking was empirical evidence capable of offering a comprehensive understanding of its broader effects on assessments. This team undertook a multi-institution and multi-disciplinary S.W.O.T analysis to discover the true impact and available opportunities. Since publication in late May 2023, this work has obtained wide-reaching praise and has been welcomed by over 13,000 people, going well beyond the boundary of the engineering education community.

Focus and relevance:

State the questions or propositions addressed and the significance of the research to engineering education research or practice.

While under development for many years, generative Artificial Intelligence (AI) burst onto the general educational landscape in late 2022 with the release of ChatGPT. This platform and others like it continue to disrupt and transform the practice of learning and teaching in engineering and beyond. Early evidence indicated that ChatGPT could pass some assessments (Gilson et al., 2022) and evade plagiarism detection by humans or software (Else, 2023). Therefore, it was critical to understand the impact of ChatGPT on the integrity of engineering education assessment. Our team, drawn from seven different Australian universities, aimed to answer the research questions:

> *How might ChatGPT affect engineering education assessment methods?*
>
> *How might it be used to facilitate learning?*

Failure to respond to the advent of generative AI could invalidate our systems for producing competent graduate engineers. Further, it would deny our students opportunities to benefit from new AI-enabled opportunities for learning and feedback. The significance of these issues has been appreciated by the wider higher education community, with widespread interest in the work. In the four months since it was published, the journal paper has seen over 13,000 views – it is now the European Journal of Engineering Education's fifth most-viewed article of all time. Further, the relevance and significance of the work can be seen in the number of invitations to give workshops and seminars on the research both nationally (AAEE, USyd, UOW, UNSW, UTS) and internationally (U. Canterbury, NZ, iLEARN, France, UOW Dubai). Finally, members of the team have been invited to write an editorial for the ethics Special Interest Group of The European Society for Engineering Education and an opinion piece for *the Chemical Engineer*.

Context:

Situate the research within relevant bodies of knowledge and describe how it contributes to new knowledge (Note: the relevant body of knowledge should be wider than engineering education and relates to the wider context of education research, frameworks, methodologies, etc.).

At the start of 2023, engineering educators were bombarded with information related to ChatGPT. This included emails from university leaders such as Vice Chancellors, Deputy-Vice Chancellors, Deans, and Heads of School, aimed at raising awareness about the software. Invitations were also extended to attend information sessions to gain insights into the capabilities of ChatGPT, but such sessions contained very little on the known impact on current assessment practices.

Teaching and learning departments were instructed to develop resources on AI software to educate staff quickly, shared through university pages, such as Lei (2023), and social media. With little empirical evidence available, only guiding principles were made available. The rapid pace of its rollout meant there was a dearth of relevant education research, especially in engineering education. The widespread hype, confusion, and duplication of workload across institutions underscored the need for methodical research into the impact of generative AI on assessment in higher education.

The research has contributed new knowledge to engineering education, and higher education in general, through a systematic evaluation of the strengths and weaknesses of generative AI across all assessment types found in contemporary engineering education. Within the short time since the work was

published, the paper has received at least 12 citations (including a JEE editorial), has been showcased in keynote speeches by world-renowned academic integrity expert, Professor Philip Dawson in TEQSA and other seminars, and was cited three times in submissions to the Australian Government's parliamentary *Inquiry into the use of generative artificial intelligence in the Australian education system* (submissions 37, 54 and 70)

Research Design:

Describe research designs, methods, theories, and/or practices appropriate to the research performed or planned and the transportability of the processes (research validity and reliability or credibility and dependability).

The nine team members from seven different Australian universities, each with different engineering backgrounds, worked together to address the research questions. The need for the research was amplified by the great institutional demand to develop a ChatGPT-based understanding of the weaknesses of current assessment practices, as well as identifying strengths and new opportunities. In designing the research, the team exploited their diversity, gathering insights from a broad range of disciplines, assessment types, and cultures, as well as spanning both undergraduate and postgraduate subjects. And yet, common elements across the subjects enabled synthesis and verification of impact in assessment types. This combination diversity and commonality in the research design enhanced both the strength and the impact of the findings as it has produced a nuanced picture of the impact ChatGPT has on a broad range of engineering education assessments. The work was the first of a kind for its approach and breadth of evaluation. As Professor Merlin Crossley (UNSW DVC Academic Quality) noted

> *"One of the best articles I've seen yet analysing carefully which assessment types ChatGPT can pass. Very thorough and scientific - for a summary go straight to the conclusions - very helpful."*

The initial phase of the work involved the authors exploring the capabilities of ChatGPT in responding to a range of assessment tasks and determining what information in the assessment prompt was important. This was completed using the free version of ChatGPT.

Team meetings followed and were used to refine the investigation and data collection process. From this, procedures and templates were formulated to create consistency across the team. A decision on which subjects would be included was made to ensure as much diversity in subject matter and assessment implementation as possible. Investigation and reporting were cascaded allowing each member to report their results to the team, scaffolding and sharing implementation to strengthen alignment. The first author played a gatekeeper role, ensuring consistency and best practice in 'prompt engineering'. This was to ensure that prompts were being applied to give ChatGPT every chance to pass the assessment tasks. The first author is also a Head of Students for undergraduate engineering, a role that requires having oversight into the structure for eleven different engineering majors. This experience provided the skills necessary to reflect on the individual efforts of each team member.

The driving force behind the research methodology was to determine if a student could use ChatGPT to pass the different assessment tasks and how difficult that would be. In grading the output, assigning a pass or fail was the primary recording method regardless of how the task might normally be assessed in practice. While trivial for closed tasks with correct answers, it was a critical part of the rigour for open ended tasks. By simply deciding whether the output was passable, the approach controlled the impact of bias from researchers knowing that the submission was AI generated.

This approach had the further advantage of allowing team members to revise and refine their inputs to see if it was possible to produce a passable output (see Figure 1). This would allow team members to best report on the impacts of input modification (that is, changes to ChatGPT's prompt). As shown in the figure, if an input produced a passable output the team member could choose to move on or further refine the input. If the output failed, then the input was generally revised until a passable outcome was achieved. However, if the input continued to fail after several revisions, the team member moved on to the next question. Sometimes it was unclear whether to continue revising and the gatekeeper would provide advice to the team member.

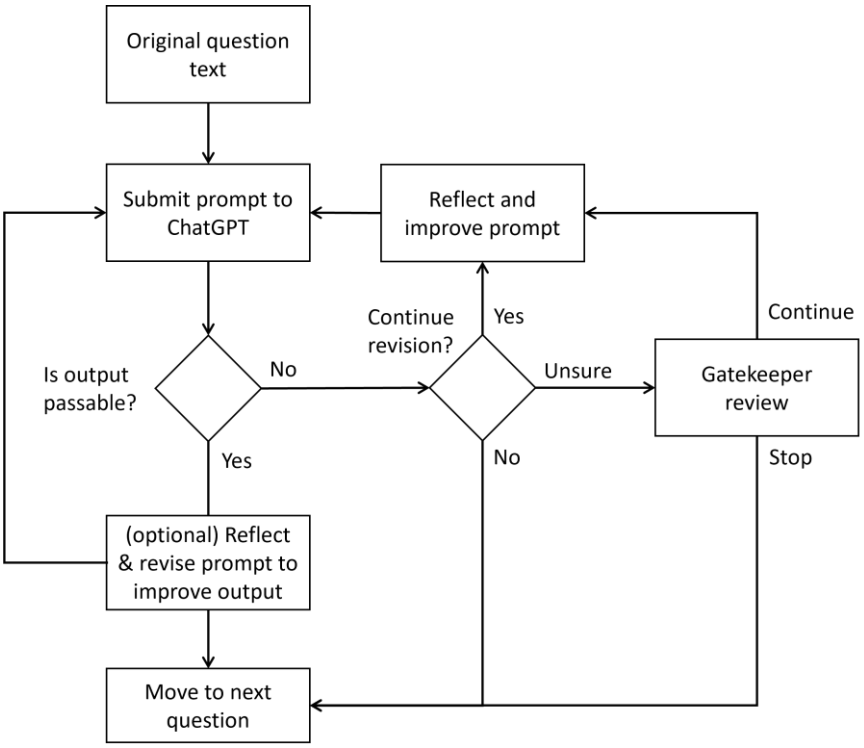


Figure 1 – Decision tree for testing and refining input to ChatGPT for each assessment task

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.

Once a team member had tested all assessments in their course, the results and examples were provided to the team. This allowed for feedback and knowledge transfer on practical prompt engineering approaches. The next team member would then repeat the process on their subject.

The reflective process of interacting with changes to the input and output, created a connection to the second research question (i.e. "how might it be used to facilitate learning?"). It was discovered during the investigation stage, that as the team uncovered what ChatGPT could or could not do, reflectively the team was also identifying opportunities.

The outcome of this process was the classification of each assessment in each course according to the level of modification required. The classification levels are given in Table 1. Further, the assessments were categorised by type (see Table 2). The team looked for the best fit of each assessment task after an analysis of the requirements of each assessment type conducted in each subject.

Table 1 – Classifications for the amount of adjustment required to achieve a passable prompt

Classification	Description
N/A	Modification not being applicable (e.g., ChatGPT could not take the place of a student for an oral presentation)
Fail	Even with modification, the output resulted in work that would fail.
None	Achieved an output that passed simply by copying the assessment task information or question directly into ChatGPT.
Minor	Minor modification to the input needed to get a pass grade. In general, this involved inputting specific parameters or instructions that provided more detailed guidance of the structure and content required in the output. For example, simply rewording the question to use structured prompt engineering (Lei, 2023a).
Major	Substantial modifications required to achieve a passing grade. In essence, this requires a student that has a solid understanding of what the output needs to be and has the technical understanding to experiment with the input to achieve the required result.

The team agreed to use shared procedures and templates to ensure consistent documentation. Examples of the input and output were required to be recorded to allow team analysis of the modification levels applied to the input. At the end of the data collection period, the team reconnected to analyse the results, share their findings, and undertake analysis.

Table 2 – Categorisation of assessment types according to task requirements and format

Assessment type	Description
Online Quizzes	Tasks that used an online quiz format using an e-learning platform.
Numerical (Assignments and Exams)	Assessments tasks where the answers are numerical in nature (e.g. calculation based) and are completed in a written format (not online).
Code Submission	Assessments requiring the submission of programming code.
Oral	Assessments comprising presentations, interviews, pitches and quality participation in discussion.
Visual	Visual documents (e.g. mind map) and evidence (e.g. certificate)
Written (Experiment)	Written activity associated with experimentation or laboratory work.
Written (Project-based)	Written assessment 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).

Results

Present original ideas or results of general significance supported by clear reasoning and compelling evidence.

The first component of the research study consisted of a benchmarking activity to set a baseline of comparison for measuring the advancement of ChatGPT into the future, and also allowing others undertaking similar subjects to compare. The second component of the research study was to investigate the impact ChatGPT has on assessment implementations. An analysis was made of the impact on different assessment types by looking at each assessment across all subjects, noting that the findings would be quickly obsolete due to the rapidly changing nature of the technology. The two data analysis sections have been undertaken using the framework of a SWOT (Strengths, Weaknesses, Opportunities

and Threats). SWOT is a popular framework used in academic peer-reviewed literature and used by consultants, trainers and educators around the world (Helms and Nixon, 2010). This methodical approach has proven transferable and was used in a recent "Generative AI and Assessment 101" workshop at UNSW aimed at those who had little experience of using ChatGPT.

In a subject-by-subject analysis, ChatGPT passed three, failed five, and two subjects were too close to call. When it came to assessment types, ChatGPT passed four, failed three, and two types were tied. The study highlighted areas where ChatGPT was strong and areas where it was weak. However, all weaknesses related to upcoming technologies that if incorporated into future versions of ChatGPT would shift the balance in its favour. The overarching message from the study was that trying to beat it was futile – instead, to move forward, the community needs to live with it and use it as a tool.

The results of this work were immediately recognised both within engineering education and beyond. Professor Fiona Saunders (Associate Editor of EJEE, publisher of the paper) commented,

> *Important benchmark paper on #chatgpt3's ability to pass different types of UG Engineering Assessment. Well done to the authors…*

While Professor Philip Dawson commented,

> *Revisiting this paper today, and it's just great to see the depth they went into in trying out a bunch of different assessment types and engineering subdisciplines. Would be great to see other disciplines try a similar approach.*

This highlights that our research is highly innovative. Across all university disciplines, we were the first to undertake such a substantial investigation. Being a first mover, has led to the positive outcomes achieved to date, and the opportunities to come.

Summary

The research team addressed a significant and relevant problem to engineering education, namely an evaluation of the impact of generative AI for assessment validity across a variety of assessment types from a broad range of courses. The research was a first-of-a-kind study for engineering education and higher education more broadly and set a standard for future assessments across disciplines. The research design incorporated a systematic, reflective and peer-reviewed approach that enabled common assessment standards and transferable outcomes for educators in engineering and beyond. The significance of the work has been demonstrated by the quickly accumulated record of very large readership, citations, and requests for further dissemination through seminars, workshops and editorials. Further, the team have been praised for their work by leaders and experts in the broader higher education community.

References

ELSE, H. 2023. Abstracts written by ChatGPT fool scientists. Nature (London), 613, 423-423.
GILSON, A., SAFRANEK, C., HUANG, T., SOCRATES, V., CHI, L., TAYLOR, R. A. & CHARTASH, D. 2022. How Does ChatGPT Perform on the Medical Licensing Exams? The Implications of Large Language Models for Medical Education and Knowledge Assessment. medRxiv, 2022.12.23.22283901.
HELMS, M. M. & NIXON, J. 2010. Exploring SWOT analysis–where are we now? A review of academic research from the last decade. Journal of strategy and management, 3, 215-251.
LEI, L. 2023. How to Design Essay Assignments if Students can Access ChatGPT or Other Similar Tools [Online]. Available: https://tl.hku.hk/2023/02/how-to-design-essay-assignments-if-students-can-access-chatgpt-or-other-similar-tools/ [Accessed 18/02].

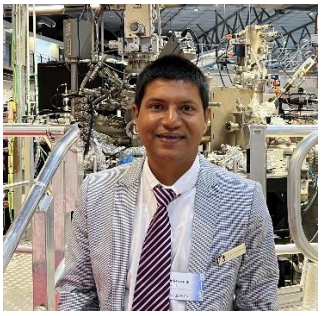
Biography

Sasha Nikolic

Sasha Nikolic received a B.E. degree in telecommunications and a PhD in engineering education from the University of Wollongong, Australia, in 2001 and 2017, respectively. He is a Senior Lecturer of Engineering Education at the University of Wollongong. His interest is developing career-ready graduates involving research in teaching laboratories, industry engagement, work-integrated learning, knowledge management, communication, and reflection. Dr Nikolic has been recognised with many awards, including an Australian Award for University Teaching Citation in 2012 and 2019. He is a member of the executive committee of AAEE and an Associate Editor for AJEE and EJEE.

Scott Daniel

Scott Daniel is a Senior Lecturer in Humanitarian Engineering at the University of Technology Sydney, and serves as Deputy Editor at the Australasian Journal of Engineering Education and on the Editorial Boards of the European Journal of Engineering Education, the African Journal of Teacher Education and Development, and the Journal of Humanitarian Engineering. Scott uses qualitative methodologies to explore different facets of engineering education, particularly humanitarian engineering. He won the 2019 Australasian Association for Engineering Education Award for Research Design for his work with Andrea Mazzurco on the assessment of socio-technical thinking and co-design expertise in humanitarian engineering.



Rezwanul Haque

Dr. Rezwanul Haque is a Senior Lecturer specialising in Manufacturing Technology at the University of the Sunshine Coast. As an inaugural member of the AAEE Academy, he has contributed significantly to the academic community. In 2019, Dr. Haque served as an Academic Lead at the School of Science and Technology, overseeing the launch of two new Engineering programs and reviewing existing ones. His dedication to learning and teaching earned him the prestigious Senior Fellow status at the Higher Education Academy (UK) in the same year. His research focuses on Engineering Education and material characterisation through neutron diffraction.

Marina Belkina

Dr. Marina Belkina is Lecturer and First Year Experience Coordinator at Western Sydney University. She has taught various subjects and courses (Foundation, Diploma, first and second years of Bachelor's Degree, online Associate Degree). She has implemented numerous projects to support learning, including: Creating the YouTube channel Engineering by Steps, Leading the development of HD videos for the first-year engineering courses, Developing iBook for physics, creating 3D lectures and aminations for Engineering Materials, and conducting research focused on exploring student's barriers to Higher Education.

Ghulam M. Hassan

Dr. Ghulam Mubashar Hassan is Senior Lecturer in Department of Computer Science and Software Engineering at The University of Western Australia (UWA). He received his PhD from UWA. He completed MS and BS from Oklahoma State University, USA and University of Engineering and Technology (UET) Peshawar, Pakistan, respectively. His research interests are multidisciplinary problems, including engineering education, artificial intelligence, machine learning and optimisation in different fields of engineering and education. He is the recipient of multiple teaching excellence awards and is awarded AAEE Engineering Education Research Design Award 2021

Sarah Grundy

Sarah Grundy is an education-focused lecturer at the School of Chemical Engineering, The University of New South Wales. Sarah predominantly teaches design subjects at all levels (undergraduate to postgraduate). Sarah has over 15 years of experience in Research & Development, manufacturing, and project management in industry. Sarah's passion is to develop students to be credible engineers and make their impact in whatever industry through authentic learning practices.

Sarah Lyden

Sarah completed her BSc-BE (Hons) at the University of Tasmania in 2011. From 2012 to 2015 she was a PhD candidate with the School of Engineering and ICT at the University of Tasmania. From March 2015 to February 2018 Sarah was employed as the API Lecturer in the field of power systems and renewable energy. Since 2018, Sarah has been employed as Lecturer in the School of Engineering. Sarah has been a member of the School of Engineering and ICT's STEM education and outreach team.

Peter Neal

Peter holds BE(Hons) and PhD in Chemical Engineering from UNSW. Peter Neal is an Education Focused lecturer and Academic Governance Coordinator with UNSW Chemical Engineering. He has been teaching undergraduate and postgraduate engineering students for more than 10 years, and has a key focus on developing his students' capacity in design, enquiry, and professional skills. As the first Education Focused academic in the School of Chemical Engineering and as a PVCESE Education Focused Champion, he works to develop the educational capacity of his colleagues in his School, the Faculty of Engineering and across UNSW.

Caz Sandison

Carolyn (Caz) is an Associate Professor in mathematics and the Head of Students for Mathematics and Physics at the University of Wollongong. An award-winning lecturer, she has a keen interest in mathematics education, spending lots of time with students and working out more effective ways of teaching and assessing, while thinking about Topological Groups in her spare time. Caz is keenly interested in how to better recruit, train and retain maths and science teachers through collaboration between discipline experts (mathematicians and scientists) and educationalist experts.