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

Analyse, evaluate, create: Using ChatGPT to develop student's critical analysis skills for 'what good looks like'.

Zachery Quince, Hannah Seligmann and Andrew Maxwell
University of Southern Queensland (UniSQ), School of Engineering
Corresponding Author Email: zach.quince@unisq.edu.au

ABSTRACT

CONTEXT

A student's ability to critically analyse text is crucial for engineering students both during, and beyond their studies. Building on this ability to analyse, is then the ability to evaluate and create, completing Krathwohl's revision of Bloom's taxonomy (Krathwohl, 2002). Current students are suffering from a lack of professional skills when graduating (Crosthwaite, 2021; Crosthwaite et al., 2018), demonstrating the inability to utilise the higher levels of Bloom's taxonomy even in the most basic examples. Furthermore, with the prevalence of generative artificial intelligence (GenAI), it is more important than ever to develop these critical evaluation skills early in their studies.

PURPOSE

This paper describes an innovative assessment design that utilises ChatGPT to generate an essay text that students are required to critique. The idea stems from students' requests for an essay exemplar to assist them in completing this particular assessment in past years. When investigating the click-through rate of previously provided examples, it showed that these were not typically opened by students. As such, embedding the exemplar essay as part of the assessment yields multiple benefits including exposure to critical analysis, evaluation, and providing feedback.

APPROACH

The design of the assessment is constituted of two parts. The first was to read, mark and critique an AI generated essay for topical, formatting, and grammatical errors; the second took this to a higher level by writing their own essay. The premise is that the students would develop their own written essays at a comparable, or higher level, to the exemplar essay ChatGPT had generated. This forms the research question if students are required to read and mark an essay, do they then write to that level? This assessment item was included in a second-level course at UniSQ in 2023 that focused on technology and society. The cohort comprised of 64 students most of which in their first or second year of their studies.

OUTCOMES

This study showed that the premise of critically analysing an essay that is embedded as part of an assessment allowed for the development of the higher levels of Bloom's taxonomy. Students on average, received a mark that was higher than the mark the GenAI-generated essay received. In three out of the four sections that were marked, the students' essays received a higher grade than the ChatGPT-generated essay by more than 5%. Tangential outcomes were also observed showing the development of students' knowledge of ChatGPT (GenAI) and it's place in both industry and study. During this process, it was noted that the ChatGPT-generated essay passed both the academics' and students' marking process and that both the students and academics marked to the same overall level.

CONCLUSIONS

Overall, the project demonstrated students were able to develop critical analysis skills while evaluating and creating to a comparable or higher level than what they had previously evaluated. Through the assessment the students were exposed to GenAI and shown its positives and negatives. The results showed that embedding an analysis style assessment into the curriculum can help develop critical skills in a more efficient and applicable manner than directly teaching it.

KEYWORDS

Critical analysis, GenAI & innovative assessment

Introduction

The ability to analyse, evaluate and create is critical for engineering students to develop during their studies (Meda & Swart, 2018). It is particularly important as students have different educational backgrounds (Quince & Phythian, 2023) and some students may not have previous practice in this area. From analysing technical documentation to evaluating engineering designs and creating their own, more often than not students are still struggling with this in their final year of studies (Ahern et al., 2019). Elements of personal and professional creativity are often required to help this development, allowing the synthesis of concepts and extending one's evaluation capabilities. Cropley and Cropley (Cropley & Cropley, 2005) identified the benefits of creativity in education, showing that in groups of students with high IQs, those with the additional capability of high-level creativity will outperform other students consistently at university. Often divergent creativity, i.e. spontaneous, is relied on however training in convergent, or on demand creativity, is also strongly warranted to support evaluation and design skills (Maxwell et al., 2023). As such, it is critical that analysis, evaluation and creativity skills are capacity built throughout their degree. These three skills link directly to Krathwohl's revision of Bloom's taxonomy (Krathwohl, 2002). Whilst Bloom's taxonomy is not necessarily explaining these skills over a prolonged period, it does require time to continually build upon, such that students are graduate ready. This aligns with the Engineers Australia (EA) stage 1 competency statements, namely element 3 – personal and professional skills (Australia, 2019) which currently is being undertaught in first-year engineering education (Quince et al., 2023).

For students to capacity build it is important to understand what capacity is being targeted. As such the marking rubric for assessment is crucially important for students to understand, especially in first year. Students are not experts in the constructive alignment between the program and course learning outcomes, and the assessment rubric, nor should they but they should be able to understand and apply a rubric to a piece of work. This not only capacity builds the students towards understanding their own learning but also their capacity to analyse and evaluate. With the emergence of Chatbot technology (such as ChatGPT) it is expected that the higher levels of Bloom's taxonomy will need to be applied to inputs for such technology to get the best outcome.

Student uptake of this technology has been so widespread that mainstream news articles are presenting stories about it and, consequentially, academic integrity modules within institutions are being re-written to include artificial intelligence. With such an uptake, there is increasing concern regarding the ethical impacts of this technology. Specifically, the potential increase of cheating, plagiarism and academic misconduct within the academic setting, as well as the potential spread of inaccurate data, misinformation, or bias (Zhou et al., 2023).

On the positive side, studies have demonstrated the many potential uses for chatbot technology within the education sector. From the student perspective, use of ChatGPT has been linked to increased student motivation and engagement, (Xia et al., 2022) improved outcomes for students with learning challenges, (Zheng et al., 2021) an accessible tool for grammatical text revision and refinement (Fang et al., 2023) and a revolutionary tool for teaching and learning mathematics, (Wardat et al., 2023) From the teacher perspective, ChatGPT can be used to help develop individualised resources that allow learners to work through material at their own pace, as well as support teachers to evaluate work and give meaningful feedback (Gill & Kaur, 2023).

ChatGPT works via a natural language, and in essence it is waiting for feedback. Once information is generated, there are opportunities to ask for an extension of the work or a complete rewrite. If this is not undertaken the first iteration of answers is typically lacking depth and there can be major errors present both in information validity and in writing style (Gill & Kaur, 2023; Nur Fitria, 2023). Furthermore, if the input question from the user is vague and closed natured then the output will also be the same and will not develop further despite multiple iterations.

It is acknowledged that students in higher education settings can use GenAI to generate information and submit it as their own, and many current studies focus on the negative ethical implications of GenAI. However, there is a use for this type of technology in both study and the workplace if implemented correctly. Instead of banning GenAI in the academic setting, the normalisation and embedding it within courses will help students to understand the limitations of the GenAI, and apply critical thinking when using it for their own work.

As such, a case study into the use of ChatGPT in assessment was conducted as part of the first-year engineering technology and society class at the University of Southern Queensland (UniSQ). The three goals of the assessment item were developed. The first was to normalise the usage of GenAI software in a tertiary environment. The second goal was to determine if analysing and evaluating tasks leads to creativity – essentially investigating if students analyse bad or mediocre work, do they then create work to a higher level? Lastly, to start capacity building students to use the higher levels of Bloom's taxonomy. Given the development of this assessment, part of the discussion will also record if the generated essay passes the university rubric, not just marked by students but academics as well. As such, the aims of this research are:

1. Does a critical analysis of a body of text develop students' ability to critically analyse?
2. Does evaluation of work lead to improved student results?
3. Can a ChatGPT generated essay pass a university rubric?

Methods

Assessment Pedagogy

To develop these skills it was important to allow students a chance to investigate what good looks like. As such the assessment was split into two sections. The first part of the assessment asked students to review and provide feedback on an exemplar essay, both qualitative and quantitative through a written response and marking rubric respectively. The second part was to write a short essay that answered three questions about the use of ChatGPT in their (future) professional and study careers.

Part A consisted of a short essay that was generated for students to assess using Chat GPT software. ChatGPT was selected for several reasons over other AI platforms. ChatGPT has some glaring issues when it comes to the information that it uses to generate a response (Gill & Kaur, 2023). For example, some of the information that it uses is created and then referenced however, does not exist (Fitria, 2023). Secondly, there is no information past 2021 that it uses to generate a response (Gill & Kaur, 2023). Lastly, students were aware of ChatGPT as there was a great deal of public awareness at the end of 2022, beginning of 2023 resulting in students entering the classrooms, talking and using the software. The input prompt used in ChatGPT was as follows.

> "Write a 600-word essay on the social, economic and environmental implications of ChatGPT using Harvard style referencing".

The specific input prompt was designed so that it would produce some errors in output generated by ChatGPT (Castillo-González, 2022; Nur Fitria, 2023). In Part A of the assessment students were required to analyse the ChatGPT generated essay using a standardised marking rubric and provide an overall score on the essay as well as comments responding to the rubric. This serves as the analyse and evaluate steps of Bloom's taxonomy. The rubric had five sections which were: key ideas, organization and structure, use of sources, spelling and grammar, and referencing. Each of the five criteria had five grades associated with them which were: insufficient (0-49%), limited (50%-64%), developing (65%-74%,) effective (75% - 84%) and comprehensive (85% - 100%). As a part of the study, several (five) academic staff, including the three authors marked the ChatGPT generated essay.

Students were marked on the graded exemplar essay and the feedback they provided based on five criteria which were marking to the rubric, discussion of the content, identification and justification of the information within the essay, assessment of the writing style, and analysis of the provided sources. These criteria overall assessed the students' ability to critique and provide feedback.

In Part B of the assessment students were asked to write a 500-word personal reflection on the likely impact of ChatGPT on their own study and professional career. The reflection was structured and marked using the same rubric that they analysed in Part A.

During the reflection students were asked to respond to the following prompts:

1. Where is ChatGPT likely to be used in your chosen discipline (engineering, surveying, built environment)? Do you consider these potential uses for the program to be positive or negative?
2. What impacts can you see ChatGPT having in your study life? Are these impacts likely to be positive or negative?
3. Where are you most likely to use the program in the future?

Data Collection

To respond to research question 1 there were three parts of the students' responses from Part A of the assessment that were examined. The first part was to examine and analyse the students' marks from this section. This would provide context of if the students were able to meet the academic requirements of the task. If students were able to pass this would show that students possess a passing knowledge of critical analysis and feedback. The second was to investigate the students' qualitative feedback. If the comments made had justification based on sound knowledge, then this would again show that students had begun/have developed an introductory critical analysis skill. Lastly to determine if the students, as a cohort were able to understand and interpret the ChatGPT generated essay, a statistical analysis of mean, variance and standard deviation of the grade given to the essay by students was undertaken. This was then compared to the academics' marks to determine if they were evaluating at the same level of ability.

To respond to research question 2 the students grades from their personal reflection were compared against the mark that they gave the ChatGPT essay. This would allow the authors to see if students were able to examine a piece of writing and carry the same mistakes into their own. As such, each of the criteria was written so that it could be matched between the assessment in Part A against what they were marked in their reflections. As the personal reflection did not require the use of sources, this criterion was omitted and only key ideas, organisation and structure, spelling and grammar, and referencing were used.

To respond to research question 3, five university academics from the school of engineering at UniSQ marked the essay against the same rubric that students were given. The surveyed academics had various positions within the school, from lecturer in civil engineering, associate professor in electrical engineering to professor in electrical engineering. Each of the marks were reviewed for differences and comments on why the mark was given were recorded. An average mark was then allocated and compared to the average mark given by students. This revealed any potential mismatch of staff and academics on the perception of assessment rubrics.

Results and Discussion

There were 53 out the 64 students in the cohort that completed the evaluation of the ChatGPT generated essay and five university academics from the school of engineering. The results generated from this analysis can be seen in the table below.

Table 1: Student and academic evaluation of ChatGPT essay.

	Rubric criteria					Overall
	Key ideas	Organisation and structure	Use of sources	Spelling and grammar	Referencing	
Academics	73%	71%	42%	83%	40%	64%
Students	66%	69%	55%	85%	45%	66%

As seen in Table 1 there was a difference between the students and academics for three criteria: key ideas, use of sources and referencing. The differences in referencing were not major and this came down to a lack of understanding of the wording used in the passing category. The failing statement for referencing was “no referencing provided”. Referencing was provided, and as such no less than 50% was to be given for this criterion, as a mark of 0-49% was designed for not providing any referencing. However, it could be argued that the referencing was severally incorrect and justified the lower mark. It should be noted that it was not just the students who interpreted the rubric this way, two academic staff also rated the referencing criteria lower than 50%. Another point of difference was that the use of sources criteria was rated higher by students than the academics. All the academics had rated the paper a failing grade, except one, and the students averaged a passing mark. When investigating the wording of this criterion, it was again, the difference between not using sources and missing sections/incorrect usage that contributed to the difference in marks. In this criterion, the minimum mark should have been 50% as the essay contained in-text referencing. Again, they were not correctly formatted and there was no corresponding reference in the reference list. This particular criterion could have been argued as the lower grade as the sources themselves were not referenced and in some cases nonfactual. The last criteria that was different was the key ideas. Students rated this section lower than the academics by close to 10%. Both average marks were in the same grading column for this criterion – developing. When investigating this criterion it was noted that academics tended to choose the upper boundary as this essay was designed to meet a first year criteria. Students were in fact more critical of the work provided. During the feedback stage of this assessment, it was noted that several students expected ChatGPT to be “a lot better” than they had assumed.

Students and academics agreed on the organisation, structure, and spelling and grammar of the ChatGPT generated essay. Out of the 53 students that evaluated the ChatGPT essay only five gave a mark that was less than a passing grade of 50%. For these students, they had given very low to no marks for the use of sources and referencing criteria. All five academics passed that essay with an average grade of 64% compared to the students’ average grade of 66%. This result was not surprising, but it did also validate some of the fears for the community over the use of generative AI in academia. This essay was passed by close to 60 people, including 5 academics that have 120 years experience combined in the tertiary education sector. This shows that ChatGPT can generate an essay that is able to pass as a first year engineering essay.

There were three consistent themes with the students’ qualitative feedback to the essay. The first commented on the referencing style provided by ChatGPT. Most students noted that the referencing style was incorrect for the one reference provided. Furthermore, the information that was used within the essay was also heavily commented on. Students noticed the amount of misinformation with one student stating “The essay is misleading, as it suggests that ChatGPT will dramatically increase the unemployment rate with many people losing”. Another student noted the specific misinformation that could be present when describing itself “ChatGPT has a bias when writing about itself”. Lastly, students commented on the structure and depth of the essay. There was a consensus that the information was mostly shallow, unless referencing itself. The essay itself was just a series of short paragraphs that do not follow a typical essay structure.

Students specifically noted

> *"The essay does not emphasise how intertwined social, economic and environmental implications are."*
>
> *"The introduction does not properly establish or explain what the essay is about."*

All 64 students submitted Part A of the assessment. As noted previously, not all students provided marking of the rubric but all did include qualitative discussion. As one of the criteria that the students were marked against was "marking the rubric" the students who did not submit the rubric marking were not considered in the average results below.

Table 2: Average student results of Part A of the assessment (%).

	Rubric criteria					Overall
	Marking of rubric	Discussion of content	Identification of misinformation	Assessment of writing style	Analysis of sources provided	
Students	71.72	71.21	70.00	70.40	67.07	70.08

The average mark for Part A was 70.08% for the 53 students who undertook the assessment item. This high credit (65% +) mark showed that students were able to critically analyse and evaluate others' work. The results from this part of the study showed that students are well aware of what is solid assessment writing and can provide feedback to the written source. What was slightly concerning, was that there were some students that still did not understand why the sources provided by ChatGPT could contain misinformation. As such it is recommended that future assessment in first year should involve a detailed investigation into the validity and reliability of sources of information.

Table 3: Comparison of rubric criteria between students' assessment of the ChatGPT generated essay and the students' reflection marks including differences between the two criteria.

	Key ideas	Organisation and structure	Spelling and grammar	Referencing	Overall
Students assessment of ChatGPT essay (%)	66.09	68.83	84.63	45.09	65.93
Students reflection (%)	71.37	77.27	75.44	61.03	71.41
Change in grade (%)	5.28	8.44	-9.19	15.94	2.37

The students' average result for Part B, the reflective essay, was 71.41%. Compared to the 65.93% students graded the ChatGPT generated essay, this result showed that on average the premise of the study was correct that students would write to the same if not higher level than what they had seen. It does come into question whether the result is due to the students' skills already being up to that level though. Anecdotally, this result is higher than the average marks of other similar reflective essays given to first year engineering students. As this was a reflective essay the results could not be compared to previous cohorts as this type of task has not been ran within this course previously.

Breaking down the different criteria, it was shown that three of four of the rubric criteria were the same. It should be noted that there were only four criteria that overlapped between the students' grading of the ChatGPT generated essay and the reflective essay. As such, the marks were scaled to correspond to the new weightings. Key ideas, organisation and structure, and referencing were all higher than that of what the students marked the generated essay as. The key ideas section was the most interesting as this is the section of essays that students typically have struggled with in the past. It is worth noting that the organisation and structure of the essay was again higher at ~8.5%. This showed that student was understanding of the requirements of this task as most students gained a distinction level mark. The largest difference was the referencing marks. This,

whilst not surprising, is still a very good trend. Students were able to reference at a level more than 15% higher than what they had seen. Whilst the bar was not set high by the ChatGPT essay, getting students referencing correctly in first year is something that is good to see. Lastly, the only negative difference between the two essays was the spelling and grammar section. Students still scored quite highly in this section but did not reach the high distinction levels of the AI generated essay. This shows, on average, that students' levels of spelling and grammar are not at the higher end of expectations of academics in first year. This does bring into question whether students should be doing more English based courses and assessments to improve their skills in this area. Engineers write a lot of technical information and documents, therefore having a high level of written English is important.

Conclusions

The outcomes of the study were quite successful and overall positive. Students were able to review a body of work, critically analyse it and create something that was not only at the same quality, but arguably better. First year students were given the opportunity to engage with ChatGPT as part of an assessment item and gained aware of the limitations of such a platform. The results also revealed that more work is required for first year students in grammar and spelling and investigating the quality and reliability of sources. Alarmingly, the essay that was generated by the ChatGPT platform passed almost 60 reviews by students and staff - with all five staff passing the essay when grading to a rubric. Future works will involve more interaction with generative AI to push the limits of generative AI in the tertiary environment. A sample of the upcoming work include investigations into the ethical application of GenAI in tertiary settings, utilising GenAI to customise assessment items for diverse cohorts and year levels. The authors recommend incorporating generative AI into engineering curriculum to destigmatise and show the strengths and limitations of the platforms.

References

Ahern, A., Dominguez, C., McNally, C., O'Sullivan, J. J., & Pedrosa, D. (2019). A literature review of critical thinking in engineering education. *Studies in Higher Education*, *44*(5), 816-828.

Engineers Austrlia (2019). *Stage 1 Competency Standard for Professional Engineers*. Retrieved from https://www.engineersaustralia.org.au/publications/stage-1-competency-standard-professional-engineers

Castillo-González, W. (2022). The importance of human supervision in the use of ChatGPT as a support tool in scientific writing. *Metaverse Basic and Applied Research*, *2*, 29. https://doi.org/10.56294/mr202329

Cropley, D., & Cropley, A. (2005). Engineering creativity: A systems concept of functional creativity. In *Creativity across domains* (pp. 187-204). Psychology Press.

Crosthwaite, C. (2021). Engineering futures 2035 engineering education programs, priorities & pedagogies. *Australian Council of Engineering Deans, Report*.

Crosthwaite, C., Hargreaves, D., Wilson, J., Lee, P., Foley, B., Burnett, I., Goldfinch, T., & Symes, M. (2018). Engineering futures 2035. 29th Australasian Association for Engineering Education Conference 2018 (AAEE 2018),

Fang, T., Yang, S., Lan, K., Wong, D. F., Hu, J., Chao, L. S., & Zhang, Y. (2023). Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *arXiv preprint arXiv:2304.01746*.

Fitria, T. N. (2023). Artificial intelligence (AI) technology in OpenAI ChatGPT application: A review of ChatGPT in writing English essay. ELT Forum: Journal of English Language Teaching,

Gill, S. S., & Kaur, R. (2023). ChatGPT: Vision and challenges. *Internet of Things and Cyber-Physical Systems*, *3*, 262-271.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*(4), 212-218.

Maxwell, A., McAlister, C., & Quince, Z. (2023). Happy Little Trees. Proceedings of the 34th Annual Conference of the Australasian Association for Engineering Education (AAEE 2023),

Meda, L., & Swart, A. J. (2018). Analysing learning outcomes in an Electrical Engineering curriculum using illustrative verbs derived from Bloom's Taxonomy. *European Journal of Engineering Education*, *43*(3), 399-412.

Nur Fitria, T. (2023). Artificial intelligence (AI) technology in OpenAI ChatGPT application: A review of ChatGPT in writing English essay. *ELT Forum Journal of English Language Teaching*, *12*, 44-58. https://doi.org/10.15294/elt.v12i1.64069

Quince, Z., Hills, C., & Maxwell, A. (2023). Are engineering programs meeting the 2035 professional and personal requirements? Proceedings of the 34th Annual Conference of the Australasian Association for Engineering Education (AAEE 2023),

Quince, Z., & Phythian, M. (2023). Educational Diversity in Engineering Entrance Pathways. Proceedings of the 34th Annual Conference of the Australasian Association for Engineering Education (AAEE 2023),

Wardat, Y., Tashtoush, M., Alali, R. M. A., & Jarrah, A. (2023). ChatGPT: A Revolutionary Tool for Teaching and Learning Mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, *19*, 1-18. https://doi.org/10.29333/ejmste/13272

Xia, Q., Chiu, T. K. F., Lee, M., Sanusi, I. T., Dai, Y., & Chai, C. S. (2022). A self-determination theory (SDT) design approach for inclusive and diverse artificial intelligence (AI) education. *Computers & Education*, *189*, 104582. https://doi.org/https://doi.org/10.1016/j.compedu.2022.104582

Zheng, L., Niu, J., Zhong, L., & Gyasi, J. F. (2021). The effectiveness of artificial intelligence on learning achievement and learning perception: A meta-analysis. *Interactive Learning Environments*, 1-15. https://doi.org/10.1080/10494820.2021.2015693

Zhou, J., Muller, H., Holzinger, A., & Chen, F. (2023). Ethical ChatGPT: Concerns, Challenges, and Commandments.

Copyright Statement