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

A Study of Postgraduate Generative AI Usage in Engineering

Michael Crocco
Monash University[a]
Corresponding Author Email: michael.crocco@monash.edu

ABSTRACT

CONTEXT

At a large, research focused, Australian university, Generative Artificial Intelligence (GenAI) usage is not only unrestricted, but encouraged. This study focuses on a cohort of more than 200, predominantly international, postgraduate engineering students encountering this environment of encouragement for the first time in their studies. The usage of generative pretrained transformers (GPTs) is detailed, and their utility is summarized by students in the studied cohort. These details are aggregated to build a picture of how students engage with these resources and the recording instrument's suitability for purpose is analysed.

PURPOSE

The study intends to articulate the extent to which students use GenAI in the described context, and, in so doing, estimate its utility to this cohort. While *policy* at this university has been set in a way that enables GenAI use in all but the most specific cases, specific *procedures* are still evolving. Thus, educators are compelled to establish appropriate and effective means of documenting AI usage; a proposed instrument is evaluated to inform practice.

APPROACH

A detailed record of all GenAI interactions, critical summaries of their utility at the individual prompt-response level and across various categories of usage. This includes an assessment of the extent to which the generated content was incorporated into final submissions. The data are then statistically analysed to understand the impact of AI usage on students' assessment scores.

OUTCOMES

Students were found to benefit most when they put considerable effort into research and ideation in conjunction with GenAI (co-creation). Meanwhile, those who relied too heavily on report composition conducted by AI models saw a relative drop in scores. Generally, students will benefit from more sophisticated prompt engineering skills, an emerging area of expertise which requires further educative efforts.

SUMMARY

The outcomes of this study will be used to guide future implementations of similar assessment tasks and supporting instruction. The utility of GenAI as an enabler of critical analysis in entry-to-practice engineering students is noteworthy, and appropriate incorporation in teaching and assessment is recommended.

KEYWORDS

Generative AI, policy

Introduction

This paper focuses on the unique setting of postgraduate engineering students at Monash University, exploring how artificial intelligence (AI) tools influence their report writing skills. Given the emphasis in engineering education on blending theoretical knowledge with practical skills, this context is particularly relevant. By analysing students' experiences and outcomes, this study aims to provide insights that can inform and support educators in the evolving landscape of generative artificial intelligence (genAI) assisted learning. The implications of this research could offer valuable perspectives for those looking to enhance their teaching practices through the thoughtful application of AI technologies.

Ethical approval for this study was granted by the Monash University Human Research Ethics Committee (MUHREC) with approval number 42162.

Background

The emergence of widely accessible and highly sophisticated genAI tools in late 2022 and early 2023 has caused considerable disruption in higher education (Kelly et al., 2023) and universities have taken varied approaches to permitting these tools in the classroom (Sullivan et al., 2023). Monash University has taken what it views as a proactive approach, encouraging usage of AI by requiring justification to opt out of its use on any given assessment (reference withheld).

Among the claims about the impacts of genAI on higher education that have been made since early 2023 are that it will require teachers to assume a different role, that it should be part of the curriculum, and that it will require new assessment practices (Jensen et al., 2024). This view is driven by findings of enhanced student independence as their thinking and writing is supported by these tools, as well as a recognition that the use of genAI may decouple thinking from the writing process.

Attempts to incorporate AI in the classroom first emerged in the late 1950s (Brown et al., 1978), but it seems a tipping point has been reached in the last 18 months since the emergence of, especially, free and fast large language models. Several surveys of students have been conducted since then, showing students are generally willing to use these tools and that it can improve content comprehension when compared to those using traditional search engines (Elkhodr et al., 2023).

The cohort studied here consists of students who have never studied in the presence of genAI, either because it was forbidden at their enrolled institution or due to them not studying in the preceding year. As such, they are likely comparable to a typical cohort studied in January 2023 where some researchers found that the vast majority of students thought use of ChatGPT, the most popular resource at the time of writing, constituted cheating (Intelligent, 2023). The same study found that 30% of students had used genAI to complete a written assignment, which holds with the prediction of Cummings et al. (2024) that the development of AI will outpace their adoption in higher education writing as students become familiar with their integration in process and exhibit hesitance to use them, lest they be seen as cheating.

Objectives

This study aims to describe the genAI usage of a cohort of Masters of Engineering students enrolled in the first semester of the first year of this degree and relate it to assessment performance, specifically expert-assessed grades. This is intended to illuminate the level of self-sufficiency students have when using these new tools, identify what usage modes, such as ideation or general report composition, students will gravitate towards, and inform educators of how they may direct teaching efforts to optimise learning experiences. Specific attention is paid to large language AI tools (as opposed to AI writing assistants or AI-powered educational tools) and the ethics associated with AI use in the classroom were generally ignored in favour of permitting students the latitude required by such an exploratory study.

Methods

Cohort

This study was conducted within a unit of study entitled Engineer in Society, a compulsory unit for all students enrolled in a Masters of Engineering at Monash University. The students are overwhelmingly of international origin (with only two of 228 identified as domestic) and, according to in-class surveys, have never been permitted to use genAI in the completion of academic assessment tasks, either at this or any other institution.

These informal surveys, conducted in face-to-face class sessions, indicate that previous experience with genAI was varied, despite an absence in the classroom, ranging from none to extensive, with some (though limited) students using genAI tools in their professional careers.

Resources and Instruction

Students are provided a template for recording their genAI interactions and to be submitted as an appendix to their reports. This template captures: a) The genAI model used and b) it's version (e.g., ChatGPT / 3.5), c) the share link to the chat, where available, and d) the date of the interaction. It includes a table where students are to provide a summary of the overall interaction pointing out positive and negative points and briefly summarise the outcomes, such as whether the output from AI was directly copied, or modified in the final report.

The template also provides students space to record their prompts and to critically summarise responses from the genAI model. This summary is requested to avoid overly lengthy documents (as genAI text output can be very large) and to encourage students to think critically about whether the output is useful or needs further refinement.

The template document also provided an abridged example of how to summarise and a link to a chat conducted by the researcher to create content for a similar assessment task (assigned a year prior).

This document and process was explained in detail in face-to-face class sessions, 4 weeks before the assignment was due. Also in this session, was a live trial of the logical limitations of current (at the time) models. The need for critical analysis was explained and examples of hallucinations and inefficiencies were explored.

Questions were taken and the class was revisited by the researcher for question and answer sessions twice before the submission deadline.

Assessment

The assessment in question is an individual assignment worth 10% of the overall mark and submitted at the midpoint of the semester. Despite this seemingly low weighting, it is considered the major individual assessment and it forms the basis for later group-based activities; without a pass, students cannot participate in the group project and, thus, cannot earn above a P (pass) in the unit. Significant time is devoted to marking and providing individual feedback as this is considered pivotal by the examiner to demonstrate key learning outcomes in the unit of study.

Students are asked to analyse a scenario, the merger of hypothetical engineering firms, and apply professional practice knowledge gained in the first half of the semester in an engineering context to define a successful merger and how to achieve it in the format of a business report to the (hypothetical) CEO of the newly formed company.

For many, the concepts explored and leveraged in this assessment will be new and the integration of such topics as cultural inclusiveness planning into engineering studies may be encountered for the first time.

Approach

GenAI usage acknowledgements were collected from each student and this usage was characterised by the researcher. This involved the classifications as described in Table 1. As well as basic (logical) indicators describing a) whether any acknowledgement of AI was provided, b) whether AI was used (note that students were requested to declare if they did not use AI resources at any stage of their assessment creation), c) which AI models were used, and d) additional modes of AI usage (grammar, paraphrasing, translation).

Table 1: Generative AI usage classifications with definitions

Variable	Classification	Definition
Sophistication	Low	Simple prompting with minimal to no background information or expectation declaration
	Medium	Background and expectation provided in prompts, but no critical response to model output
	High	Well formed initial prompt, followed by well-considered successive prompting for fine tuning of output
Research and Ideation	Limited	Minimal search with minimal background information
	Moderate	More than one topic researched but with simple prompting
	Extensive	Significant engagement with AI models for ideation and/or research of the topic of focus
	None	No research or ideation conducted using AI
Composition	Limited	Minimal writing composition, e.g., overall outline only
	Moderate	The AI model is used to write one or two subsections of the report
	Extensive	Several sections are initially written by the AI.
	None	No report writing is done by the AI
Prompt Count	Numerical	A simple tally of the number of total AI Model prompts (where available)

The variables in Table 1 are used in multiple linear regression modelling to determine the importance of individual variables in prediction of assessment mark (and the sense of this effect, be it subtractive or additive). The typical model of a least squares multiple linear regression is shown in Figure 1. In this case, the predicted value $\hat{y}$, is the assessment mark. a is a constant (the y intercept), x_n is the nth independent variable, and b_n is the coefficient of the nth independent variable. A machine learning model is used to compute the regression "surface" in a $n + 1$ dimensional space.

$$\hat{y} = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

Figure 1: Linear equation of a least squares multiple linear regression

When interpreting statistical analyses of these types, p-values are often helpful as a measure of statistical significance that helps determine whether the observed relationship is likely to represent a genuine effect or not. In this study:

- p-values less than 0.05 are considered statistically significant, indicating strong evidence against there being no effect or relationship

- p-values between 0.05 and 0.1 suggest marginal significance or a trend that may be worth further investigation
- as p-values increase above 0.1, this indicates that the observed effect is not statistically significant, meaning the ability to rule out chance as an explanation for the observation decreases with this increase in p-value

For some submissions, it was not possible to determine the number of prompts used (Prompt Count), so a study was carried out to determine the importance of this variable among the majority for whom this could be determined. Considering p-values in this linear regression model showed that the Prompt Count variable was not significant (p-value = 0.8481), providing confidence that this variable can be removed from the analysis, and allowing a larger dataset to be analysed without having to impute numerical values to the unavailable observations. In other words, null numerical entries cannot be considered in the model used and imputing a false value would likely skew any results obtained.

So that the assessment and sophistication of genAI usage could be judged independently of the report submission, these items were analysed separately with the assessor ignoring AI usage and the researcher ignoring the final submission. This approach ensures that genAI usage can be judged on its own merits, while the assessment marking remains unbiased.

When creating a linear regression model with categorical values (as opposed to numerical ones) it is necessary to establish a reference category. The approach is clear with categories such as research and ideation or composition as their absence is a clear reference. When judging the sophistication of prompting, however, the modal value (low sophistication) was selected as a reference.

Findings

Despite a university policy requirement for students to acknowledge AI usage, the provision of a template to do so, written instructions provided with the assignment, and multiple, face-to-face class sessions during which this requirement was discussed, only 33% (76 of 228) of students provided the requisite acknowledgement appendix or a facsimile.

While the absence of reporting could be interpreted as two thirds of the cohort having not used any genAI in the creation of their report, students were provided clear instructions on how to report the non-usage of these tools. Meanwhile, several students who provided the appendix did as instructed and stated that no genAI was used in the creation of their submission.

Students may have misunderstood instructions, intentionally rejected what may be seen as extra work, or neglected this requirement due to other, unseen reasons. Whatever the cause, the number of participants will limit the utility of statistical methods for accurate prediction, but these approaches nonetheless remain useful for indicating trends.

This study focuses on those 76 students, hereafter referred to as “the cohort”, who submitted an acknowledgement of their use (or absence thereof) of genAI. First, though, a brief comparison between groups who followed the provided documentation instruction and those who did not. Those who neglected to provide an acknowledgement had a mean score of 5.6 compared to those who did provide an acknowledgement, with a mean score of 6.1 (out of 10).

General Usage Trends

Of the cohort of 76 students identified for detailed study, 69 students (or 91%) used genAI in the process of composing their report, achieving a mean score of 6.0. In comparison, the 9% who didn’t use any genAI had a mean score of 6.7. Here, it is likely that students who were diligent enough to follow instructions and chose not to engage with AI tools felt confident in addressing the requirements of this assessment task due to previously acquired aptitude in the skills and/or knowledge assessed.

OpenAI's ChatGPT 3.5 was, by far, the most commonly used tool used by this cohort, with 66% of identifiable usage being attributed to this model. This was followed by the paid version (at the time), ChatGPT 4.0, at 26%, then Microsoft Copilot with 4% of usage. Only two students made mention of Grammarly, though it is important to note that this service has been available for some time as a plug-in to other word processing applications, and many students may not consider its use out of the ordinary.

The median number of prompts used by students who provided this detail was only 6 (mean of 8), and a maximum of 36. Given that the instructional example of establishing an outline and having a genAI model to begin writing just a single section of the report required 10 prompts, this was unexpectedly low usage. During class sessions, it was stressed that there is likely to be a point of diminishing returns where it takes more prompting to reach a desired outcome than it would require independent writing to achieve that outcome. However, the observed norm was for students to bypass a process of writing refinement except for in select cases; this is what is referred to under the variable "Composition" in this study.

Beyond primary uses of Research / Ideation and Composition, many students (33 of 69) used AI for grammatical checks, and only 3 of those acknowledging its use translated from or into their first language. This seeming underuse may again be attributable to underreporting, as grammar checking is embedded in word processing applications and there are many services that translate without overtly using artificial intelligence.

26 of 69 students sought rewrites with instructions to paraphrase or correct tone for the intended audience. Reasons for the underuse of this capability are unclear, though the expectation of a tonal adjustment appropriate for a business report may not have been apparent to these students.

Regression Model

Of primary interest is understanding the influence of different modes of genAI use on assessment scores. As previously mentioned, the analysed population is small, and, accordingly many variables will be statistically insignificant, but trends are of note for consideration in the classroom. See Table 2 for a summary of the multiple linear regression accounting for prompt sophistication, and extents of research and ideation, as well as composition.

Table 2: Multiple linear regression coefficients and significance

Variable	Estimate	Standard Error	p-value
(Intercept)†	5.5907	0.5644	3.66e-14 ***
Sophistication - Unknown	-0.7568	0.8309	0.3661
Sophistication - Medium	-0.1519	0.5501	0.7835
Sophistication - High	1.4235	0.8198	0.0877 .
Research & Ideation - Limited	-1.1482	0.8340	0.1738
Research & Ideation - Moderate	-0.3540	0.6139	0.5663
Research & Ideation - Extensive	1.0111	0.6716	0.1375
Composition - Limited	0.6250	0.6464	0.3375
Composition - Moderate	1.0489	0.6678	0.1216
Composition - Extensive	0.2937	0.6850	0.6696

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The intercept in table 2 represents the expected score in an instance of low sophistication and no research & ideation or composition. As the variables analysed are not numerical, but categorical, the estimations of effect of scores can be interpreted as the expected impact of being in one of those categories. For example, exhibiting highly sophisticated prompts means that the student is estimated to score 1.42 points (out of 10) higher than a student with low prompt sophistication.

Pairing these estimations with their significance, a few key points emerge:

1. High sophistication of prompts is highly associated with improved outcomes
2. Limited (or superficial) research and ideation is associated with worse outcomes
3. Extensive research and ideation is associated with improved outcomes.
4. Moderate levels of composition assistance are most associated with improved outcomes.

The importance of the variable categories can be visualised by comparing T-values of the categorical variables, per the graph in Figure 2.

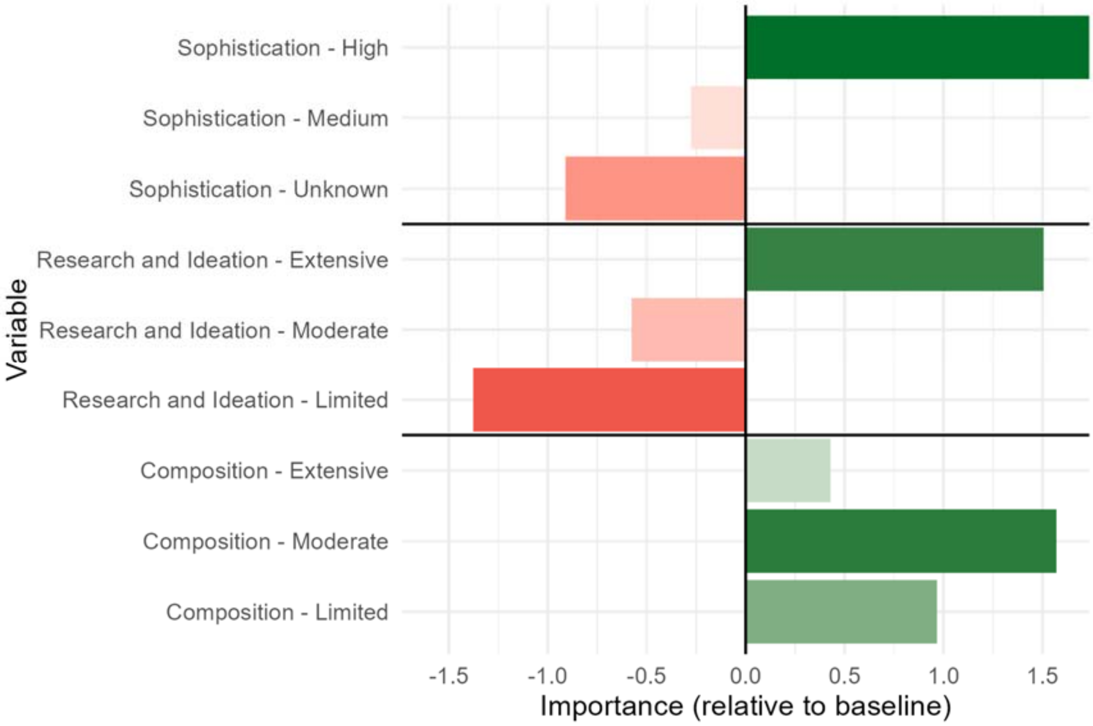

Figure 2: Importance comparison of regression variables

Conclusion

While this study is limited in both cohort size and context, important inferences can be drawn from it for consideration in higher education. While familiarity with genAI models and their interfaces will increase and become more intuitive in the months and years ahead, at present, there exists a significant disparity between students, and the potential benefits from effective use should not be overlooked. Given an understanding that students can use these tools in their studies, whether encouraged or not, and whether permitted or not, there is evidence of an advantage (in terms of both scores and learning outcomes) provided to those who use them well.

Prompting skill is an area of focus for educators aiming to equalise advantages between students. This is often referred to as Prompt Engineering, a term of varied definitions but one that can be viewed as a combination of genAI model and subject matter understanding; that is, one who understands how the models work and can prompt them in a way most effective for the subject at hand can elicit the best response. While this is reassuring on one hand, in that, at least at present, a reasonable amount of subject matter knowledge is key to getting meaningful output, on the other hand there is an emerging skill requirement that is not often being directly addressed, creating a type of paradox in education, and it’s as though we’re teaching students complex computational tasks while they’re still learning how to effectively use their calculators. In essence, higher education is grappling with when and how to teach students how to harness genAI alongside traditional subject matter.

The potential of Generative AI Large Language Models to assist in the research phase of a task should not be overlooked. Recognition of the search potential of these tools, as they can connect by topic and meaning, rather than just selected words, and can provide summaries which can lead to further search, all in a conversational format, is akin to recognising that students will learn a lot on a topic by engaging in a discussion with an expert.

By their nature of being trained on huge amounts of human writing, AI models can draw from this dataset to arrive at conventional outlines for a given assessment task very quickly. This very likely expedites the early writing stages, determined by Kellogg (1988) to demand the highest cognitive load. So, tasking a genAI to write one's report can be beneficial (provided sufficiently sophisticated prompting) initially, but an interesting trajectory emerges in this author's experience. When a model is prompted to write a section of a report following the provision of background information, it generally produces something of high legibility, flow, grammatical precision, and structure. Due to inaccurate conclusions drawn by the models, some refinement prompting is often needed to capture the thoughts and logic of the prompting author; this is the level described as "moderate".

At this stage of composition, there likely remain some errors, depending on the complexity of the arguments, and the prompting author can continue to attempt to refine the text through prompting ("extensive" composition) which may result in further divergence from the goal, often via conflating or confusing of concepts or so-called "hallucinations", or, having read a potential composition and identifying its shortcomings, may be in a position to continue refinement themselves. It is this transition to human-conducted refinement, in this author's opinion, that leads to a more desirable outcome. The author has benefited from a stage of co-creation, but only they, in a way still unique to humans, can hold a central objective and ensure the composition moves the reader toward it. It is similarly argued by Thanh et al. that the advantages of genAI to learning during conduct of authentic assessments "could be compromised if students resort to using…models to **complete** [the] tasks" (2023, emphasis added).

Monash University faces a challenge associated with increasing cohort sizes (that of the subject of this study was twice the size it had been in previous years) and the resultant difficulty to engage at length with small groups of students or individuals. Assessment drafting feedback and ideation sessions are luxuries no longer afforded to the instructor, and yet, an opportunity is presented by the emergence of genAI models where students can explore topics independently and expose themselves rapidly to connected information. Perhaps we are at the point of realising the opportunity imagined by Brown et al. (1978) when considering the application of artificial intelligence in education and noted, a half century ago, "schools [were] flooded with experimental programs to teach students to "think" (à la problem-solving), but where are these students being taught how to understand something new on their own, let alone what it means to 'understand'?"

References

Brown, J. S., Collins, A., & Harris, G. (1978). Artificial Intelligence and Learning Strategies. In *Learning Strategies* (pp. 107–139). Elsevier. https://doi.org/10.1016/B978-0-12-526650-5.50010-1

Cummings, R. E., Monroe, S. M., & Watkins, M. (2024). Generative AI in first-year writing: An early analysis of affordances, limitations, and a framework for the future. *Computers and Composition*, *71*. https://doi.org/10.1016/j.compcom.2024.102827

Elkhodr, M., Gide, E., Wu, R., & Darwish, O. (2023). ICT students' perceptions towards ChatGPT: An experimental reflective lab analysis. *STEM Education*, *3*(2), 70–88. https://doi.org/10.3934/steme.2023006

Intelligent. (2023, January 23). *Nearly 1/3 of college students have used ChatGPT on written assessments*. Intelligent. https://www.intelligent.com/nearly-1-in-3-college-students-have-used-chatgpt-on-written-assignments/

Jensen, L. X., Buhl, A., Sharma, A., & Bearman, M. (2024). Generative AI and higher education: A review of claims from the first months of ChatGPT. *Higher Education*. https://doi.org/10.1007/s10734-024-01265-3

Kellogg, R. T. (1988). Attentional Overload and Writing Performance: Effects of Rough Draft and Outline Strategies. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *14*(2), 355–365. https://doi.org/10.1037/0278-7393.14.2.355

Kelly, A., Sullivan, M., & Strampel, K. (2023). Generative artificial intelligence: University student awareness, experience, and confidence in use across disciplines. *Journal of University Teaching and Learning Practice*, *20*(6). https://doi.org/10.53761/1.20.6.12

Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning and Teaching*, *6*(1), 31–40. https://doi.org/10.37074/jalt.2023.6.1.17

Thanh, B. N., Vo, D. T. H., Nhat, M. N., Pham, T. T. T., Trung, H. T., & Xuan, S. H. (2023). Race with the machines: Assessing the capability of generative AI in solving authentic assessments. *Australasian Journal of Educational Technology*, *39*(5), 59–81. https://doi.org/10.14742/ajet.8902

Acknowledgements

The author would like to acknowledge Michael Kennedy, who made their classroom and assessment available for the purpose of this research. Always encouraging and proactive in recognition of the importance of the emergence of generative AI in the engineering classroom, the author's thanks cannot be understated.

Generative AI was used as a topic reference search engine (Scopus AI), as a coding syntax assistant during data analysis (OpenAI ChatGPT 4 and 4o), and to gain feedback on writing tone and clarity during final edits (Anthropic Claude 3.5 Sonnet). These outputs were considered, and some were adapted by the author; no direct substitutions of AI generated text were made.

Copyright statement