An Investigation Into ChatGPT Generated Assessments: Can We Tell the Difference?

Mauryn C. Nweke, University College London, United Kingdom
Matthew Banner, University College London, United Kingdom
Manal Chaib, University College London, United Kingdom

The Barcelona Conference on Education 2023
Official Conference Proceedings

Abstract
The impact of ChatGPT has been revolutionary in many capacities however institutions are beginning to see the gradual increase in students passing off AI generated work as their own. This has negative impacts for student learning and academic integrity. One way to help combat this is to understand if we can tell the difference between AI generated assignments and original pieces of work. This will help those involved in assessing to distinguish between AI generated work compared to original work. In the initial phase of this study, we use ChatGPT to generate assessments for 3 modules in the department of Biochemical Engineering, UCL. These assignments capture the interdisciplinary nature of Biochemical Engineering as well as the diversity in assignment complexity and include mathematics, business and bioprocess validation and quality control. We then convene academic leads and marking staff to assess scripts, compare them to previous cohorts through use of peer-observation to find out what indicators there are of generated work. Results so far have shown that out of the 3modules, 2 modules receive a pass mark with minimal prompts and only 1 module lead was able provide indicators to identify generated work. Results also show that ChatGPT was unable to provide solutions for complex mathematical problems, bioprocess piping and instrumentation technical drawings and critical analysis required for M-level bioprocess quality control. Subsequent phases of the study will expand the number of modules tested on ChatGPT, embed its use into the engineering curriculum and upskill academics on the use of AI tools.

Keywords: ChatGPT, Engineering Education, Assessments

Introduction

ChatGPT is a large language model developed by Open AI. Chat GPT-4 launched on 14th March 2023 and is currently the newest version of the software. It is able to generate human-resembling text responses to the text prompts. It works on a conversational approach generating responses from its wide dataset.

The diverse range of functionalities and applications of ChatGPT includes:

- Answering questions
- Language translation
- Grammar correction
- Writing assistance

ChatGPT serves as an easily accessible knowledge repository, providing quick and comprehensive explanations on a wide array of topics. Its ability to synthesize and present information aids in understanding complex concepts, making it an invaluable tool for students seeking clarification or exploring new subjects (Nikolic et al., 2023). It encourages active learning. By formulating questions and engaging in discussions with the model, students can refine their critical thinking and problem-solving skills. The model can present different perspectives and suggest relevant resources, fostering a deeper exploration of subjects beyond traditional methods. Additionally, ChatGPT can act as a writing assistant, aiding students in composing and refining their academic work. It offers suggestions for structuring essays, improving clarity, and enhancing overall writing quality, thereby boosting students' communication skills (Sánchez-Ruiz, Moll-López, Nuñez-Pérez, Moraño-Fernández, & Vega-Fleitas, 2023).

While ChatGPT offers several benefits for student learning, several studies have highlighted notable disadvantages and limitations that need to be considered (Ali, Shamsan, Hezam, & Mohammed, 2023; Muñoz et al., 2023; Sallam, Salim, Barakat, & Al-Tammemi, 2023; Tyson, 2023):

1. **Lack of Contextual Understanding:** ChatGPT may not fully comprehend the context or nuances of a student's question, leading to inaccurate or irrelevant responses that could potentially confuse learners.
2. **Inaccuracies/Hallucinations and Errors:** The model's responses are generated based on patterns in its training data, and it might provide incorrect information or misconceptions, especially in rapidly evolving fields. Further, if told information is incorrect, the model can hallucinate and provide further inaccurate information to support the users prompt/query.
3. **Dependence on Technology:** Overreliance on ChatGPT could hinder students' development of independent research and critical thinking skills. Students might rely on the model instead of exploring diverse learning resources.
4. **Limited Interaction Depth:** ChatGPT's responses can be shallow, lacking the depth that a knowledgeable teacher or peer might provide in a real classroom setting.
5. **Reduced Effort in Learning:** If students find it too convenient to rely on ChatGPT for quick answers, they might skip the effort of critical thinking and problem-solving that is essential for genuine learning.
6. **Stifled Creativity:** Depending on predefined algorithms, ChatGPT might not encourage the same level of creativity and innovative thinking as human interactions and explorations would.

Incorporating ChatGPT into education should be a carefully considered decision, balancing its advantages with the potential drawbacks. To maximize its benefits, educators should encourage students to use ChatGPT as a supplementary tool while fostering critical thinking, independent research skills, and a holistic learning experience.

Aims and Objectives of This Study and Methodological Approach

The use of ChatGPT in assessments carries significant implications particularly in ensuring fairness and preventing cheating in online assessments, as the reliance on ChatGPT can aid in academic dishonesty. Many institutions fear that AI and ChatGPT can potentially obtain a degree, which would lead many to question the efficacy of university institutions in this domain. In this study, we aim to understand how ChatGPT performs in our assessments at UCL Biochemical Engineering with the aim of making the relevant adjustments to assessments whilst still maintaining rigour and learning outcomes.

Our main objectives for this project are as follows:

- Assess whether staff can identify the difference between artificially generated assessments made by ChatGPT and previous student assessments.
- Assess what threat ChatGPT poses to academic fairness
- Provide solutions to change the assessments to ensure academic fairness is preserved under ChatGPT.

The methodological approach for this phase of the study involved selecting 3 modules in the first instance to conduct preliminary studies. In order to ensure breadth in scope, the selection of these modules involved considerations such as level (pertaining to cohort year) and subject area (considering that biochemical engineering is an interdisciplinary field), among other factors. As a result, the 3 modules selected were ENGF0003 Mathematical Modelling and Analysis (a first year module – level 4), BENG0035 Business Planning in Bioprocessing and Life Science (a second year module – level 5) and BENG0041 Bioprocess Validation and Quality Control (a fourth year module – level 7). The next step involved using ChatGPT to generate assignments on these modules. The generated assignments were mixed with real student assignments and were given to module staff to assess. Staff were given 3 pieces of work each. This part of the study was to understand whether staff could identify which pieces of work were generated and which were authentic (by providing minuted feedback) and also for staff to assign grades to the generated assignments in order to understand how these assignments performed.

Results & Discussion

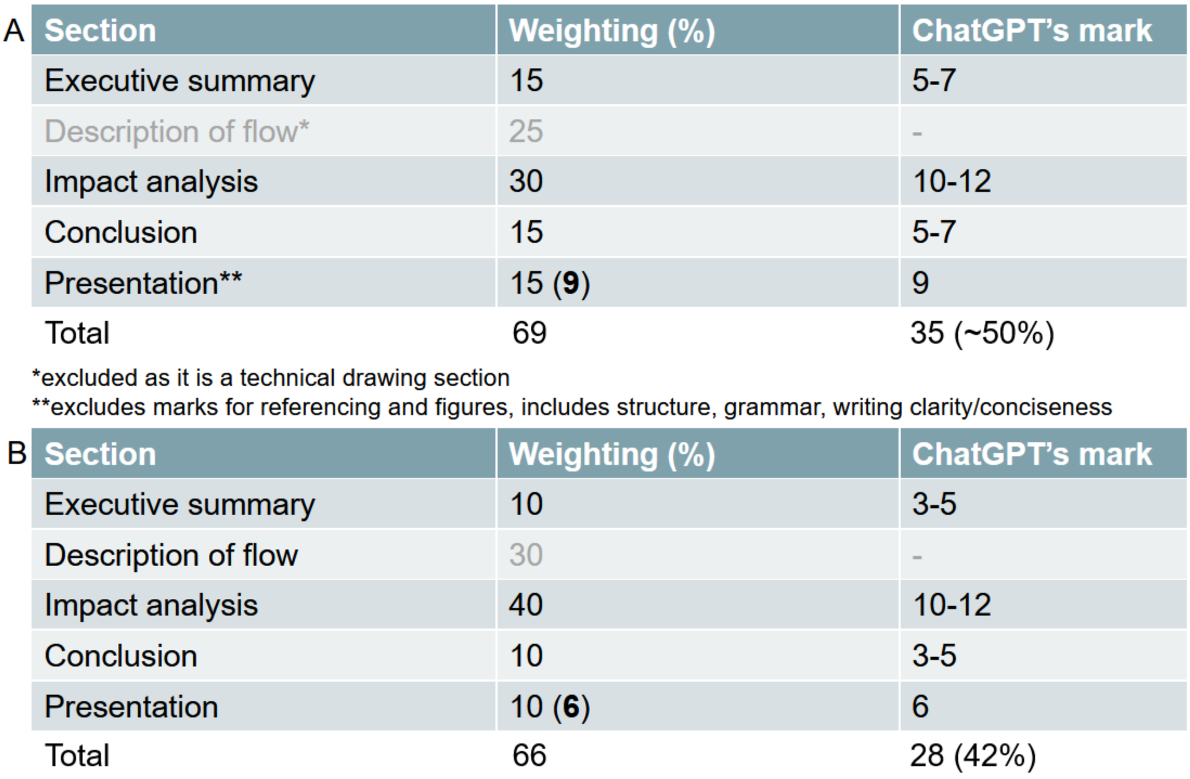
The results in Table 1 summarise the performance of ChatGPT generated assignments. What can be observed is that as the level of difficulty increases, the number of prompts needed for the generated assignment to pass also increases. It should be noted that the pass mark for levels 4-6 is 40% whilst the pass mark for level 7 assignments is 50%. The prompts used relate to the questions posed in the assessment itself. The level 4 and 5 modules required little to no modification of the assignment question in order for a passable answer to be generated. However the level 7 module required significant modifications to be made to the question/s and required an average of 3-4 prompts per question to generate a passable answer. It should be noted that in order to have modified the questions to that extent, the candidate would have had good knowledge of the subject matter and therefore would likely not be using ChatGPT to fabricate their learning, however this cannot be said with certainty for the level 4 and 5 modules.

Module	**Level**	**Average no. of prompts**	**Result (%)**
ENGF0003 (maths)	4 (1st year)	1	60-65
BENG0035 (business)	5 (2nd year)	1-2	55-60
BENG0041 (validation)	7 (4th year)	3-4	~50

Table 1 – results obtained by ChatGPT generated assignments

Staff then provided feedback on which pieces of work were generated and which ones weren't. Staff on ENGF0003 were least able to identify the generated work, followed by BENG0035 and then BENG0041. There are a number of reasons why this may have been the case. Mathematics as a subject generally requires binary/fixed answers whereas the level 5 and 7 assignments are writing assignments (reports) with a lot more ambiguity in what is considered a right or wrong answer. Staff on the level 5 and 7 modules reported on being able to identify differences in grammar, the overuse of transitional and superlative words as well as the general vagueness and lack of specificity in answers provided. This is corroborated by Waltzer, Cox, & Heyman, 2023 who conducted a similar study but also included the perceptions of students in being able to detect AI generated work. Their study revealed that staff tended to think that better written work (with no grammatical errors) was AI generated and idiosyncratic language was an indicator that the work was produced by a student. These findings may be useful within the faculty of engineering sciences, UCL, as a large proportion of the student body has English as a second language (>70%). However it should also be noted that the use of AI is accepted for the purposes of assistance with correcting grammar, so perfect grammar in itself is not an indicator of a student's misuse of AI.

The last step in this preliminary study was to evaluate assessment rubrics as a way to provide a solution. Figure 1 provides a summary of the steps taken to (for all intents and purposes) ChatGPT-proof the BENG0041 module. It was noticed in figure 1A that where ChatGPT obtained a larger proportion of marks were parts of the assignment required more description rather than critical analysis (such as the executive summary and conclusion where it obtained up to 46% of the marks attributed to these sections, compared to a maximum of 40% of marks for the section requiring the most critical analysis – Impact analysis). The learning outcomes of this module place a large emphasis on being able to critically analyse, given that it is a level 7 module. Figure 1B shows the results of ChatGPT once the weighting of each section is adjusted to place greater emphasis on the Impact analysis section. The results of this show a reduction in the proportion of marks that ChatGPT obtained in all sections, bringing its overall total far below the pass mark. It should be noted that there is the exclusion of a proportion of the total marks pertaining to technical drawings, references and figures due to ChatGPT's inability to generate those items. As alluded to earlier, where ChatGPT does excel in is its ability to generate grammatically flawless prose with a great level of clarity, hence why it scores highly in the Presentation section.



A

Section	Weighting (%)	ChatGPT’s mark
Executive summary	15	5-7
Description of flow*	25	-
Impact analysis	30	10-12
Conclusion	15	5-7
Presentation**	15 (**9**)	9
Total	69	35 (~50%)

*excluded as it is a technical drawing section

**excludes marks for referencing and figures, includes structure, grammar, writing clarity/conciseness

B

Section	Weighting (%)	ChatGPT’s mark
Executive summary	10	3-5
Description of flow	30	-
Impact analysis	40	10-12
Conclusion	10	3-5
Presentation	10 (**6**)	6
Total	66	28 (42%)

Figure 1 – A – Summary of BENG0041 marking rubric showing ChatGPT’s marks per section and final mark. B – ChatGPT’s marks per section and final mark after rubric weighting adjustment.

Conclusion

The findings have highlighted that ChatGPT is able to pass assessments at various levels in the Biochemical Engineering degree, indicating that it may be able to obtain a degree with students who demonstrate limited knowledge. A promising remedy involves revision of assessment rubrics to ensure that the weighting of the most critical elements of an assignment is adequate whilst considering the learning outcomes of the module as well as the level of study. The subsequent steps in this study have and will involve the implementation of ChatGPT and AI classes within the engineering curriculum and the use of the outcomes of this preliminary study to inform a wider study involving more marking staff and more modules in order to obtain more statistically robust data. As this is a project conducted under the auspices of the equality, diversity and inclusion (EDI) committee, the final phase of this project will look to explore the biases of ChatGPT, particularly around concerns with inherent information bias and equitable access. We will be looking to understand how the use of ChatGPT and AI affects vulnerable student populations such as those that are neurodivergent, those with English as a second language and those that are socio-economically challenged.

Acknowledgements

The authors would like to acknowledge the support of UCL ChangeMakers as project sponsors and the department of Biochemical Engineering, UCL.

References

Ali, J. K. M., Shamsan, M. A. A., Hezam, T. A., & Mohammed, A. A. Q. (2023). Impact of ChatGPT on Learning Motivation: *Journal of English Studies in Arabia Felix*, *2*(1), 41–49. https://doi.org/10.56540/jesaf.v2i1.51

Muñoz, S. A. S., Gayoso, G. G., Huambo, A. C., Tapia, R. D. C., Incaluque, J. L., Aguila, O. E. P., … Arias-Gonzáles, J. L. (2023). Examining the Impacts of ChatGPT on Student Motivation and Engagement. *Przestrzen Spoleczna*, *23*(1), 1–27.

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

Sallam, M., Salim, N., Barakat, M., & Al-Tammemi, A. (2023). ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. *Narra J*, *3*(1), 1–11.

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 (Switzerland)*, *13*(10). https://doi.org/10.3390/app13106039

Tyson, J. (2023). Shortcomings of ChatGPT. *Journal of Chemical Education*, *100*(8), 3098–3101. https://doi.org/10.1021/acs.jchemed.3c00361

Waltzer, T., Cox, R. L., & Heyman, G. D. (2023). Testing the Ability of Teachers and Students to Differentiate between Essays Generated by ChatGPT and High School Students. *Human Behavior and Emerging Technologies*, *2023*, 1–9. https://doi.org/10.1155/2023/1923981

Contact email: c.nweke@ucl.ac.uk