# *Ensuring Authentic Assessments in Higher Education – Comparing AI-Generated Responses to Case Study-Based Assessment Questions*

Nicholas Netto, Singapore University of Social Sciences, Singapore

The Asian Conference on the Social Sciences 2024
Official Conference Proceedings

## Abstract

The launch of OpenAI's ChatGPT in late 2022 sent ripples across higher education, particularly in attempts to detect its use by students using it to complete assignments. This paper involved a simple experiment that compared responses generated by artificial intelligence (AI) tools, namely ChatGPT and Google Gemini, in response to a case study-based Social Work assessment question. The outputs generated by these AI tools were analysed vis-à-vis a set of marking rubrics, and found that the responses were generally not of a high quality and were unable to integrate information from the case study in their responses. Although the AI tools were able to provide succinct summaries of the case study's key points, they generally performed poorly in accurately applying the specific tool to the case study, in identifying relevant legislation to the local context of the case study, as well as in producing sufficient words that addressed the question. Strategies to improve the authenticity of assessments revolve around enhancing their complexity viz. real life scenarios, and could include incorporating recent events (e.g. in the news) in the case study, having a mix of essential and peripheral information requiring students to discern and synthesize information in case study, and for marking rubrics to rewards points that explicitly apply information from the case study. In sum, the quality of AI-generated responses were generally inadequate and well-crafted case studies appear to be AI tools' Achilles heel which could ensure 'authentic assessments.'

Keywords: Authentic Assessment, Artificial Intelligence, ChatGPT, Case Studies, Higher Education

iafor
The International Academic Forum
www.iafor.org

## Introduction

The launch of OpenAI's ChatGPT in late 2022 sent ripples across higher education, with many excited about its immense potential to facilitate adaptive learning, provide personalised feedback, support research and data analysis, offer automated administrative services, aid in developing innovative assessments (Rasul et al., 2023), as well as in facilitating the crowdsourcing and curation of articles and learning materials, and nurturing partnerships with students to leverage "student-centered and student-guided approaches (Mills et al., 2023) in higher education.

Despite these purported benefits to enhance learning, there were also concerns about disruption within higher education, mainly in academic integrity concerns and to some extent, the ability to detect the use of AI (Artificial Intelligence) tools such as ChatGPT to complete assignments (Sullivan et al., 2023). Rasul and colleagues (2023) had articulated concerns about fairness and equity, particularly in how educators could assess students' learning if they 'cut and pasted' AI-generated responses without having engaged in the learning material.

In reality, there is a tension between and need to balance between these risks and benefits (Yahaya et al., 2023). The response to these concerns have ranged from calling for its complete ban in academic writing, to advocating for its integration to enhance students' experiences as well as learning outcomes (Benuyenah, 2023). Currently, there does not appear to be any compelling reason to endorse its use in assessments, although anecdotally there is recognition that it is currently being used by students to complete assignments. This begs the question of whether our assessments are still 'authentic' in this age of AI tools such as ChatGPT.

Proponents that espouse benefits of AI tools tend to articulate its use among academics, such as in its use to create innovative lesson plans; from the student's perspective, it could also assist in the organization and structuring of an essay, not in answering assessment questions. For the use of AI tools to become mainstream, there is need for leadership as a root support mechanism, character development as an antidote, and authentic assessment as an enabler (Crawford et al., 2023). However, at the point of writing, there is no globally accepted guidance on the use of ChatGPT in assessments.

In this regard, a group of researchers led by Dr Cacciamani from the University of Southern California have committed to undertaking a Delphi Global Cross-discipline Consensus Survey as part of the CANGARU (ChatGPT, Generative Artificial Intelligence, and Natural Large Language Models for Accountable Reporting and Use) Guidelines project. The objectives of the CANGARU project are to craft guidelines that uphold the integrity of academic and scientific endeavors when utilizing AI tools such as ChatGPT, and to develop guidelines that ensure academic integrity as well as provide guidance on the disclosure of their use in academic writing and research (Cacciamani et al., 2023).

In sum, AI tools have several benefits to enhance higher education but it is uncertain how its use might also undermine learning. Therefore, it might be prudent to educate students to avoid using ChatGPT (Sullivan et al., 2023) and examine it further before actively encouraging its use for assessments.

### *Use of Case Studies in Social Work Education*

Case studies are a common assessment method in Social Work education as they facilitate experiential understanding (Stake, 1978), and have been known to test procedural knowledge and develop strategic skills to plan, monitor and revise approach depending on the details of the case (Carpenter, 2011).

As educators in a practice-based profession, there is an innate responsibility to ensure students are trained and able to apply knowledge to formulate an assessment and intervention. Thus far, case studies have been able to fulfill this requirement, and there is a curiosity as to how case studies in formal assessments would fare if students used AI tools to answer them.

There is minimal research in this area, with only two studies from other disciplines documenting that case studies have not been well answered by AI tools.

First, a study which tested the functionality of ChatGPT in answering exam assessment questions from two chemistry modules reported that ChatGPT was not a high-risk technology tool in relation to cheating for questions that focused on application of knowledge and interpretation in two Chemistry modules (Fergus et al., 2023).

Second, another study tested the use of ChatGPT for academic writing in the biomedical sciences and found that despite the responses generated being "systematic, precise and original", it also lacked academic merit and depth, and was short on word count (Kumar, 2023).

Following the above, this study aims to:

1. Compare the responses generated by two AI tools (i.e. ChatGPT and Google Gemini) in response to a case study-based assessment question in the Social Work programme
2. Critically analyse if case studies can still ensure authentic assessments in the Social Work programme

## Methodology

### *Data Collection Procedure*

For the purposes of this research, two commonly used AI tools were selected to be part of this study, namely ChatGPT version 3.5, and Google Gemini (which was known as Google Bard when this research was conceptualised).

The following prompt was formulated to be input into the two AI tools and for a response to be generated in response to a typical case study:

1. "Write a 1400-word essay in response to the following case study and two questions, allocating about 80% of the response to the second question."

The pre-determined prompt above was entered into each of the respective AI tools on 11 April 2024. Arising from the inadequate responses obtained from the first prompt, two additional prompts were devised to increase the word count and obtain relevant information to the part of the question on legislations:

2. “The word count was only about [number] words. Could you revise it so that it is closer to 1400 words?”
3. “What are some relevant legislations in Singapore that you must consider for this case study?”

All outputs were then compiled into a Word document for analysis.

### *Data Analysis*

The two AI-generated outputs were analysed with respect to the corresponding marking rubrics that had been written for the case study, with an emphasis placed on analysing the response vis-à-vis the marking rubrics to ascertain the quality of the response, and comparing the responses across the two AI tools to identify differences between generated responses as well as their quality. Notes were written alongside the outputs during analysis.

## Findings

### *Comparison of AI-Generated Outputs*

The output generated by the first prompt for ChatGPT and Gemini were 722 and 548 words respectively. The word counts improved to 958 and 1,065 words respectively, after the second prompt. However, both AI tools still struggled to reach the word count, although both noticeably improved with the second prompt, and this was more pronounced for ChatGPT. This finding is aligned with what Kumar (2023) found in biomedical sciences field viz. AI-generated responses being ‘short of word count.’

| **ChatGPT** | |
|---|---|
| Word count for 1st prompt: | 722 words |
| Word count for 2nd prompt: | 958 words |
| Word count for 3rd prompt: | 450 words |
| Total word count after all prompts: | 1,408 words |

Table 1: Output generated by ChatGPT

| **Google Gemini** | |
|---|---|
| Word count for 1st prompt: | 548 words |
| Word count for 2nd prompt: | 1,065 words |
| Word count for 3rd prompt: | 218 words |
| Total word count after all prompts: | 1,283 words |

Table 2: Output generated by Gemini

### *AI Tools Capable of Summarizing Case Study*

One strength of the AI tools was in their ability to summarise key points of the case study, which is evident in the following extract:

> *"This case study presents a multifaceted ethical dilemma for the social worker. G, a 17-year-old student with a history of anxiety and depression, finds herself unexpectedly pregnant. Facing an emotionally volatile home environment with her recently incarcerated father, she juggles part-time work to support the family while battling her mental health struggles. Her new relationship with K, a colleague at the fast-food chain, becomes complicated by the unplanned pregnancy. K pressures her towards an abortion, leaving her feeling unsupported and overwhelmed." – Output from Gemini based on 2nd prompt (expanded version)*

Gemini appeared to perform better at this task; however, both AI tools were unable to assess and identify the crux of the case study despite being able to summarise the key details relevant in the case.

### *Inability to Correctly Apply Tool to Case Study*

A key finding from this exercise found that both AI tools were unable to correctly apply Dolgoff and Loewenberg's Ethical Principle Screen (Dolgoff et al., 2012). Specifically, both ChatGPT and Gemini referred to broader ethical principles rather than the principles of the EPS tool taught in the module.

For example, the output from Gemini mentioned "Autonomy" and elaborated that "G has the right to make decisions about her own life, including whether to disclose sensitive information to her mother." Similarly, Gemini stated "Beneficence" and that G should be encouraged to "explore all options with a non-judgmental approach."

Of significance in the AI-generated responses was that the focus of the output was on resolutions rather than the actual application of the EPS tool. ChatGPT generated only 186 words that were assessed to be relevant to the application of the EPS, whereas Gemini generated 90 words for the same component. Interestingly, the words generated for resolutions were comparatively higher at 295 and 230 words respectively. This means that the AI-generated responses both failed to adequately address the application of the EPS.

### *AI-Generated Responses Struggled to Apply Relevant Local Legislations*

Both ChatGPT and Gemini could not identify relevant legislations based on the first prompt. This improved with the third prompt but the responses were assessed not to demonstrate any real application. For example, although ChatGPT identified the Children and Young Persons Act, it narrated broadly what this entailed: "This act provides for the protection and welfare of children and young persons in Singapore. It addresses issues such as parental responsibilities, child protection, and juvenile justice. The social worker must consider Geraldine's status as a minor and ensure that her rights and best interests are protected under this legislation."

This contrasts with a good human-generated response which would mention her age as qualifying to be protected under the Act, and also make linkages with pertinent information in the case study such as her mental health status and whether she possessed the mental capacity to make an informed decision under her current circumstances, etc.

In this regard, ChatGPT fares slightly better than Gemini, although both fail to contextualise the legislations viz. the case study. Some legislations were also not stated correctly and were likely from another jurisdiction (country) and not Singapore, such as the "Mental Health Care (Amendment) Act (2014)" which could be from South Africa; in Singapore the relevant legislation would be the Mental Health (Care and Treatment) Act 2008.

Generally, the outputs from both AI tools were unable to capture the nuances, and the socio-cultural and legal uniqueness of the Singapore context. Similar to what Fergus et al. (2023) found, questions that focused on "application of knowledge and interpretation" were not easily answered.

## Implications for Practice

In general, based on the experiment carried out as part of this study, it would appear that case studies can ensure authentic assessments in Social Work education, and facilitate the training of students to apply their knowledge to scenarios they may encounter in future practice settings.

What is evident is that sufficient task complexity that is encapsulated in realistic case studies that require higher-order thinking skills, and that require students to synthesise information rather than just discuss issues may not be readily answered by AI tools such as ChatGPT and Gemini. This is congruent with what Carpenter (2011) mentioned about the use of "carefully curated case studies" to test procedural knowledge.

At this juncture, it appears that AI tools are susceptible to what exists in their data banks which may not include information that is specific to more diverse socio-cultural or legal contexts. Therefore, to further enhance case studies is to also exploit the seeming Achilles heel of ChatGPT and Gemini, which is to "bring in the local context" (Liu & Bridgeman, 2023) and to involve "real-life examples and contextually specific situations" (Lee, 2023) which could increase the authenticity of assessments.

To further improve the authenticity of assessments, the following strategies could be deployed. First, case studies could be written based on recent events which may not have been integrated into the AI tool's language model. Second, case studies should be written with both essential and peripheral information; this requires students to synthesise the information, and in so doing, to discern relevance and prioritise courses of action to take. Third, the marking rubrics should be written such that it rewards points made by the student that are explained within the context of the case study, thus demonstrating genuine application.

## Conclusion

The pervasive use of AI tools such as ChatGPT and Gemini has had an impact on various spheres in society including higher education. A genuine concern exists about the ability to ensure authentic assessments in an age where AI tools can supplant the student's ability to think for themselves and answer questions testing their learning. This study, through a simple experiment, sought to compare the responses generated by ChatGPT and Google Gemini to a case study-based Social Work assessment question, as well as to ascertain if case studies could still ensure authentic assessments.

This study found that although AI tools were capable of summarizing the key points of a case study, their responses were lacking in several areas. AI-generated responses were short of word count, were unable to correctly and adequately apply the specific tool to the case study, and struggled to identify legislation relevant to the local context of the case study. Overall, the quality of AI-generated responses were poor vis-à-vis marking rubrics. At the time of writing, case studies appear to be able to ensure ‘authentic assessments’ in a Social Work module that adopts case studies for assessments.

## References

Benuyenah, V. (2023). Commentary: ChatGPT use in higher education assessment: Prospects and epistemic threats. *Journal of Research in Innovative Teaching & Learning*, *16*(1), 134–135. https://doi.org/10.1108/JRIT-03-2023-097

Cacciamani, G. E., Collins, G. S., & Gill, I. S. (2023). ChatGPT: Standard reporting guidelines for responsible use. *Nature*, *618*(7964), 238. https://doi.org/10.1038/d41586-023-01853-w

Carpenter, J. (2011). Evaluating Social Work Education: A Review of Outcomes, Measures, Research Designs and Practicalities. *Social Work Education*, *30*(2), 122–140. https://doi.org/10.1080/02615479.2011.540375

Crawford, J., Cowling, M., & Allen, K.-A. (2023). Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). *Journal of University Teaching & Learning Practice*, *20*(3). https://doi.org/10.53761/1.20.3.02

Dolgoff, R., Loewenberg, F., & Harrington, D. (2012). *Ethical Decisions for Social Work Practice*.

Fergus, S., Botha, M., & Ostovar, M. (2023). Evaluating Academic Answers Generated Using ChatGPT. *Journal of Chemical Education*, *100*(4), 1672–1675. https://doi.org/10.1021/acs.jchemed.3c00087

Kumar, A. H. (2023). Analysis of ChatGPT Tool to Assess the Potential of its Utility for Academic Writing in Biomedical Domain. *Biology, Engineering, Medicine and Science Reports*, *9*(1), Article 1. https://doi.org/10.5530/bems.9.1.5

Lee, J. (2023, May 8). *Effective assessment practices for a ChatGPT-enabled world*. Times Higher Education Campus - Resources for Academics and University Staff. https://www.timeshighereducation.com/campus/effective-assessment-practices-chatgptenabled-world

Liu, D., & Bridgeman, A. (2023, January 23). *How can I update assessments to deal with ChatGPT and other generative AI? – Teaching@Sydney*. https://educational-innovation.sydney.edu.au/teaching@sydney/how-can-i-update-assessments-to-deal-with-chatgpt-and-other-generative-ai/

Mills, A., Bali, M., & Eaton, L. (2023). How do we respond to generative AI in education? Open educational practices give us a framework for an ongoing process. *Journal of Applied Learning and Teaching*, *6*(1), Article 1. https://doi.org/10.37074/jalt.2023.6.1.34

Rasul, T., Nair, S., Kalendra, D., Robin, M., Santini, F. de O., Ladeira, W. J., Sun, M., Day, I., Rather, R. A., & Heathcote, L. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *Journal of Applied Learning and Teaching*, *6*(1), Article 1. https://doi.org/10.37074/jalt.2023.6.1.29

Stake, R. E. (1978). The Case Study Method in Social Inquiry. *Educational Researcher*, *7*(2), 5–8. https://doi.org/10.2307/1174340

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), Article 1. https://doi.org/10.37074/jalt.2023.6.1.17

Yahaya, M., Umagba, A., Obeta, S., & Maruyama, T. (2023). Critical Evaluation of the Future Role of Artificial Intelligence in Business and Society. *Journal of Artificial Intelligence, Machine Learning and Data Science*, *1*(1), 21–29.