Designing Exam Questions in the Era of ChatGPT

Sheng Lu, University of Delaware
Xingqiu Lou, Kent State University

Background

While academic plagiarism by students is not new, ChatGPT, a rapidly advancing artificial intelligence (AI) tool, has introduced significant new challenges. For example, ChatGPT can produce academic papers that closely resemble those written by humans and answer technical questions with high accuracy (Alkaissi & McFarlane, 2023). Neither are most conventional anti-plagiarism tools effective in detecting text generated by ChatGPT, as the content is technically "original" and not copied directly from existing sources (Sullivan et al., 2023).

Based on the results of two open-book exams conducted in different institutions, this study empirically examined the relationship between question types and students' utilization of ChatGPT. The findings provide new insights into students' academic integrity behaviors in the age of ChatGPT and shed light on effective strategies to prevent AI-based academic plagiarism, particularly in the textile and apparel discipline (Benuyenah, 2023; Ventayen, 2023).

Literature review

Existing studies suggest that exam question types could significantly impact students' academic integrity behaviors. **First,** questions that evaluate students' comprehension of a theory or a specific knowledge point may be more susceptible to cheating since students can efficiently use ChatGPT to generate responses based on the tool's extensive data and pre-existing knowledge (Geerling et al., 2023). **Second**, questions that ask students to apply critical thinking and utilize course-specific learning reflections tend to be less vulnerable to AI plagiarism, as ChatGPT may not possess the necessary background knowledge to generate an accurate response (Novick et al., 2022). **Third,** since the current public version of ChatGPT cannot read or interpret visual inputs, in theory, questions that ask students to analyze and interpret graphs and figures could be less susceptible to plagiarism (Shen, et al., 2023).

Method

The data for the study was collected from two junior/senior-level fashion merchandising courses offered by two U.S.-based four-year universities. In March 2023, both classes held an online, open-book exam that consisted of three short-answer questions, each falling into one of the following three types:

- *Theory question*: students need to apply a specific theory learned in the class to explain a real-world phenomenon;
- *Reflection question*: students need to reflect on their learning experiences in the course to make an argument and reason their viewpoint;

- *Graphic question*: students need to explain the meaning of a graph or figure using a particular knowledge point learned in the course;

While students were permitted to reference any course material during the exams, **the instruction explicitly prohibited using AI tools, including ChatGPT, to generate answers**.

Altogether, 37 students from institution A and 41 students from institution B participated in the exam (i.e., n=78). Students' response to each exam question was checked individually using GPT-2 Output Detector, one of the most widely used tools for detecting content generated by ChatGPT (GPT-2, 2023).

Given the categorical nature of the data, logistic regression was adopted to assess the relationship between question type and students' use of ChatGPT (Lawal &Lawal, 2003). The model used *Cheat* as the dependent variable, measuring if a student's answer contained AI-generated content detected by GPT-2 at a threshold of 30% (i.e., 1=yes; 0=otherwise). The independent variables measuring the exam question types:

- *Theory* (1=theory question; 0=otherwise);
- *Reflection* (1=refletion question; 0=otherwise);
- *Graph* (1=graphic question; 0=otherwise);

Results and discussions

The results showed that students used ChatGPT in the two exams with varying frequency by question type. Students were found mostly using ChatGPT for Theory questions (n=13) and Reflection questions (n=10), but only one student used the tool for Graphic questions.

Further, the logistic regression was statistically significant at the 99% confidence level (likelihood ratio (L.R.) statistics $p<.001$). Specifically, **First,** when holding other factors constant, students would be 4.38 times more likely (Wald X^2=42.9, $p<.001$) to plagiarize using ChatGPT when the question asked about a theory (i.e., *Theory*=1) than otherwise. **Second**, the results showed that using reflection questions (i.e., *Reflection* =1) would also increase the odds of ChatGPT plagiarism by 3.9 times (Wald X^2=32.4, $p<.001$). **Additionally**, no clear statistical evidence shows that using graphic questions (i.e., *Graph*=1) would result in more use of ChatGPT in exams (Wald X^2=2.5, $p<.001$).

Implications and future research agendas

The study's findings provide valuable first-hand empirical evidence regarding students' use of ChatGPT in fashion merchandising courses and have several important implications. On the one hand, the results underscore the importance of rethinking the exam question types and methods of evaluating students' learning in the ChatGPT area. As the tool's popularity continues to grow alongside other AI technologies, students will inevitably use them in the learning process. In light of the changing learning environment, it is imperative for educators to adapt their teaching practices to ensure academic integrity and optimize students' educational experiences.

On the other hand, the results call for more university policy guidance on dealing with students' plagiarism using AI technologies such as ChatGPT. While tools like GPT-2 are capable

of detecting AI-generated content, whether the "*AI-based evidence*" can be used to challenge students' "AI-based plagiarism" remains unclear and controversial (Benuyenah, 2023).

Despite the interesting findings, future studies can continue to use surveys or in-depth interviews to understand students' perspectives on using ChatGPT to assist with assignments and exams. Exploring the potential of incorporating ChatGPT into the learning process and examining its impacts on the learning outcome could be meaningful also.

References

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*, *15*(2).

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.

Geerling, W., Mateer, G. D., Wooten, J., & Damodaran, N. (2023). Is ChatGPT Smarter than a Student in Principles of Economics?. *Available at SSRN 4356034*.

GPT-2 Output Detector, GPT-2 (2023). GPT-2 Output Detector. Retrieved from https://openai-openai-detector.hf.space/

Lawal, B., & Lawal, H. B. (2003). *Categorical data analysis with SAS and SPSS applications*. Psychology Press.

Novick, P. A., Lee, J., Wei, S., Mundorff, E. C., Santangelo, J. R., & Sonbuchner, T. M. (2022). Maximizing academic integrity while minimizing stress in the virtual classroom. *Journal of Microbiology & Biology Education*, *23*(1), e00292-21.

Shen, Y., Heacock, L., Elias, J., Hentel, K. D., Reig, B., Shih, G., & Moy, L. (2023). ChatGPT and other large language models are double-edged swords. *Radiology*, 230163.

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).

Ventayen, R. J. M. (2023). ChatGPT by OpenAI: Students' Viewpoint on Cheating using Artificial Intelligence-Based Application. *Available at SSRN 4361548*.