JOURNAL OF CHEMICAL EDUCATION

pubs.acs.org/jchemeduc

Communication

Evaluating Academic Answers Generated Using ChatGPT

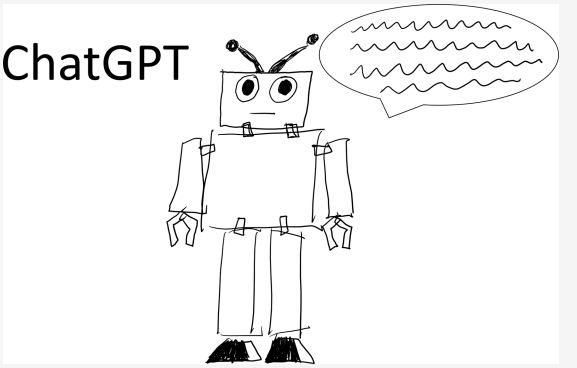
Suzanne Fergus,* Michelle Botha, and Mehrnoosh Ostovar

Cite This: *J. Chem. Educ.* 2023, 100, 1672–1675

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The integration of technology in education has become ever more prioritized since the COVID-19 pandemic. Chat Generative Pre-Trained Transformer (ChatGPT) is an artificial intelligence technology that generates conversational interactions to user prompts. The trained model can answer follow-up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. The functionality of ChatGPT in answering chemistry assessment questions requires investigation to ascertain its potential impact on learning and assessment. Two chemistry-focused modules in year 1 and year 2 of a pharmaceutical science program are used to study and evaluate ChatGPT-generated responses in relation to the end-of-year exam assessments. For questions that focused on knowledge and understanding with "describe" and "discuss" verbs, the ChatGPT generated responses. For questions that focused on application of knowledge and interpretation with nontext information, the ChatGPT technology reached a limitation. A further analysis of the quality of responses is reported in this study. ChatGPT is not considered a high-risk technology tool in relation to cheating. Similar to the COVID-19 disruption, ChatGPT is expected to provide a catalyst for educational discussions on academic integrity and assessment design.



KEYWORDS: *First-Year Undergraduate, General, Public Understanding, Outreach, Internet, Web-Based Learning, Applications of Chemistry*

The importance of digital tools in higher education was acutely experienced during the global COVID-19 pandemic where an immediate pivot to online delivery was essential for the continuation of education studies. Blended approaches to support learning and teaching have placed a focus on digital literacies and capabilities.[1,2] The digital capability for each individual student will depend on their subject specialism, career choice, personal factors, and other contextual factors. Embracing technology is non-negotiable as a graduate attribute in the 21st century.[3]

With the pivot to online assessments during the COVID-19 pandemic, there was a necessary shift to adapt assessment alternatives that were inclusive, accessible, reliable, and valid.[4−7] For example, time-constrained unseen examinations in invigilated rooms or in-class tests were adapted to online "take-away" exams in which questions or tasks were administered virtually and students submitted their responses electronically within a set period.[8] Assessment design required reframing from recall-based tasks to questions that required students to demonstrate how they use information rather than reiterate what they have learned. Knowledge-based questions were adjusted to problem solving, data interpretation, or case-study-based questions that were suitable to an open-book online assessment format.[5] A key aspect with good assessment design is promoting academic integrity and preventing opportunities for academic misconduct.[9] A well-written question item aims to create intellectual challenge and to require interpretation and inquiry. Questions that cannot be easily "Googled" or easily answered through a single click in an internet search engine is a focus. Good assessment design is considered a key factor to reduce cheating, although not exclusively.[10,11] File-sharing sites are known and have been highlighted as routes to contract cheating, particularly during the COVID-19 pandemic.[12]

A recent technology development is Chat Generative Pre-Trained Transformer (ChatGPT),[13] which has been trained using deep-learning algorithms to generate conversational interactions to user prompts. The trained model can answer follow-up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. The technology was released for public use on November 30, 2022, and as of the publication of this Communication, ChatGPT is free for public use.

Received: February 1, 2023
Revised: March 15, 2023
Published: March 31, 2023

We are interested in investigating the use of this artificial intelligence technology in academic assessments. The research questions that informed this investigation are as follows:

1. Can ChatGPT generate answers to chemistry assessment questions?
2. What is the quality of answers generated by ChatGPT?
3. How similar are answers from requests using different user accounts? Does plagiarism-matching software Turnitin report similarities with the answers generated by ChatGPT?

CONTEXT OF STUDY

Pharmaceutical Science at the University of Hertfordshire, United Kingdom, is a general applied chemistry program. Foundations of Pharmaceutical Chemistry is a compulsory 30-credit year 1 chemistry module that focuses on developing an understanding of organic chemistry nomenclature and reactivity in preparation for more detailed medicinal chemistry modules that follow and introduce basic principles of thermodynamics, atomic orbital theory, and chemical structure. Methods in Drug Design is a compulsory 30-credit year 2 chemistry module that focuses on medicinal chemistry principles, molecular modeling, and related in silico methodologies, in conjunction with structure−activity relationships (SAR) to inform rational drug design. The end-of-year exam assessment for both modules consists of an individual 24-h open-book assessment with short-answer questions (SAQs) that cover the teaching content and learning outcomes for the modules. An overall average grade of 40% or more is required to pass the assessment. The year 1 module assessment provides five questions, and students need to answer four questions (Supporting Information). The year 2 module assessment provides six questions, and students need to answer all questions (Supporting Information). Students are advised to spend no more than 3 h working on these assessments in total and are limited to a 500-word count per question. These assessments were used in this study to evaluate ChatGPT-generated responses and address the research questions.

METHOD

The authors who are all faculty/staff registered for a ChatGPT account. With use of S. Fergus' account, each assessment item was copied directly as a prompt to generate a response using the artificial intelligence tool. A separate chat was opened for each assessment item. Additional prompts were used including the phrase "using appropriate references" to the original assessment item wording. The word count of the ChatGPT-generated response was noted. Each ChatGPT-generated response was marked using the assessment mark scheme and checked by a second marker. The process was repeated by M. Ostovar and M. Botha so that an additional two sets of responses could be compared. The questions from the year 1 and year 2 modules were categorized using Bloom's Taxonomy to evaluate the level of intellectual demand required.[14] Turnitin, which is used to perform similarity checking of students' academic work, was utilized to check the ChatGPT-generated responses.[15] Turnitin is a web-based text-matching service that searches a variety of electronic sources to identify any duplication with work submitted by students. As well as websites and databases, the Turnitin software searches for matches against other student submissions. However, it does not search all digital records in existence, nor does it detect matches with nondigital content such as textbooks. The Turnitin software will generate a report highlighting sections of text that have been found to match those in the student's work. It will identify the percentage of each match and provide the source (although not necessarily the original source). Turnitin will not decide as to whether a student has plagiarized, but the information it provides can help to make an informed decision as to whether academic misconduct has occurred.

RESULTS AND DISCUSSION

This section has been divided to address and respond to the research questions of the study. All ChatGPT-generated responses are available in the Supporting Information.

Research Questions: Can ChatGPT Generate Answers to Chemistry Assessment Questions? What Is the Quality of Answers Generated by ChatGPT?

It was possible for ChatGPT to generate answers to chemistry assessment questions but not for all chemistry questions. The responses were articulated well. In the year 1 assessment, question items that referred to a chemical structure presented as a figure could not be answered. This is a common question format for organic chemistry topics. In the year 2 assessment, this limitation was also observed with question items that required the drawing of chemical structures in the answer or the plotting of a graph from data provided. Question 4 incorporated a newly approved medicine—"Vericiguat is a medication used to reduce the risk of cardiovascular death and heart failure, approved for medical use in the United States in January 2021 and for use in the European Union in July 2021"—and ChatGPT's response could not source this medicine: "I'm sorry, I don't have the information about the SMILES string for Vericiguat as the drug was only recently approved for medical use and my knowledge cut off is 2021". Table 1 and Table 2 report the classification of each

Table 1. Year 1 Assessment Items and Evaluation of ChatGPT-Generated Responses

Question	Categorization of Assessment Item	Grade (%)	Comments
1	Understanding	65	Understanding evident with justification. Some specifics omitted.
2(a)	Understanding/application	0	No answer
2(b)	Understanding/application	0	No answer
2(c)	Understanding/application	0	No answer
2(d)	Application	0	Response does not answer question
3(a)	Understanding/application	0	No answer
3(b)i	Application	0	Incorrect answer
3(b)ii	Understanding	0	Response does not answer question and contains an error
3(c)	Understanding	0	Response does not answer question
4(a)	Understanding	42	General response with some key points
4(b)	Understanding	88	Understanding evident
5(a)	Understanding	25	Some relevant points included
5(b)	Understanding	0	Response does not answer question
5(c)	Understanding/application	15	Some relevant points included

https://doi.org/10.1021/acs.jchemed.3c00087
J. Chem. Educ. 2023, 100, 1672−1675

assessment item (using Bloom's Taxonomy) and the grade percentage awarded by faculty/staff to the associated ChatGPT-generated response.

Table 2. Year 2 Assessment Items and Evaluation of ChatGPT-Generated Responses

Question	Categorization of Assessment Item	Grade (%)	Comments
1(a)	Understanding/application	50	General outline, although some steps omitted
1(b)	Understanding/application	55	Correct steps indicated, although with no chemical structures
1(c)	Understanding/application	55	Correct steps indicated, although with no chemical structures
1(d)	Understanding/application	55	Correct description, although key conclusion omitted
2(a)	Application	0	Incorrect answer
2(b)	Application	0	Incorrect answer
3(a)	Understanding/application	0	No answer and additional information do not answer question
3(b)	Understanding/application	0	No answer and additional information not relevant
3(c)	Understanding/application	0	No answer and additional information not relevant
3(d)	Application/analysis	0	Incorrect answer
4(a)	Application	0	No answer
4(b)	Application	0	No answer
4(c)	Application	0	No answer
5	Understanding	50	General response with some key points
6(a)–(c)	Application/analysis	0	No answer
6(d)	Understanding	25	General response with few key points

It was interesting to note that question 3(b)ii, in the year 1 assessment, contained an error and provided the incorrect answer. This was not expected as the correct answer can be found via an internet search. The overall grade on the year 1 paper calculated from the top four graded answers would be 34.1%, which does not meet the pass criteria. The overall grade on the year 2 paper would be 18.3%, which does not meet the pass criteria.

Research Question: How Similar Are Answers from Requests Using Different User Accounts? Does Plagiarism-Matching Software Turnitin Report Similarities with the Answers Generated by ChatGPT?

The identical question prompts from two different user accounts (M.B. and M.O.) did not produce identical answers. This is a feature of the artificial technology used in ChatGPT where the use of large language models, after reading the input, generates a probability distribution for the first character of the response.[16] The model managing program samples from that distribution, generating the first character, which can be one of several possibilities. ChatGPT is fed the first character, which was sampled from the distribution and generates a distribution for the second character and so on. Therefore, there is randomness in the answers generated, and it always responds differently. Unexpectedly, some questions that generated no response for user S.F. generated responses for the other two different user accounts. There was a one-week time gap between user S.F. requesting ChatGPT-generated responses and those of users M.B. and M.O. because of ChatGPT exceeding the user capacity with high demand and being unavailable. In question Q3(b)ii, in the year 1 assessment, which shows the chemical structure of penicillin followed by "How many stereogenic carbon centers are present in the penicillin molecule?", each user obtained a ChatGPT-generated response with a different error. The correct answer is "three", and yet the answers generated by ChatGPT were "one", "four", and "five". In question 1, in the year 2 assessment, the ChatGPT generated responses that outlined organic reaction schemes in text only. Using text only is not the typical format in representing organic chemistry. In question 3, in the year 2 assessment, ChatGPT generated an answer that acknowledged "I'm sorry, I am unable to draw the structures of the reactants and products as I am a text-based AI model" and additional information provided was general. A week later, the ChatGPT-generated responses did indicate a chemical structure, for example, "Pentan-2-one: H | C−C−C−C−C | C═O". This demonstrates how the technology develops in a short time period and with user input. Similarly, for question 4, in the year 2 assessment, the ChatGPT-generated response a week later provided an answer, although the chemical structure of the SMILES string for Vericiguat that was provided was incorrect. Figure 1 shows the chemical structure of the ChatGPT-generated response compared to the Vericiguat structure required.

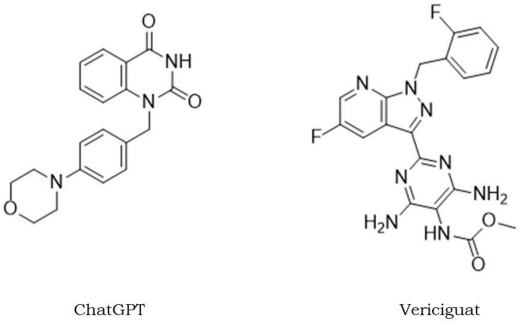


Figure 1. ChatGPT-generated response for question 4, year 2 assessment, provided a SMILES string for Vericiguat, and the chemical structure corresponding to the answer generated is shown in comparison to the Vericiguat structure required.

The appropriate references cited in the ChatGPT-generated responses could not be found through online search checks in some instances. Turnitin did not produce a high-percentage matching score on both the year 1 (21% similarity) and year 2 (25% similarity) assessments.

A look at the Turnitin reports (Supporting Information) showed that there was nothing to alert any further investigation required in relation to academic integrity.

IMPLICATIONS

The findings in this Communication help to highlight how educators can utilize assessment design to appropriately challenge and stretch students at their level of study. This is particularly important given the burgeoning developments with artificial intelligence. ChatGPT will not be the only such tool available for generating chat responses to user prompts and at the time of publication; other options such as Microsoft's Bing and Google's Bard are emerging. It is expected that academic integrity platforms such as Turnitin will develop capacity to detect artificial intelligence generated text. This space will

https://doi.org/10.1021/acs.jchemed.3c00087
J. Chem. Educ. 2023, 100, 1672−1675

continue to develop at pace. There are many potential opportunities with ChatGPT as a learning tool; however, a discussion on this is outside the scope of this Communication.

CONCLUSION

ChatGPT-generated responses to chemistry assessment questions where provided were well-written. The quality of answers in this investigation varied, and ChatGPT demonstrated limitations in relation to application and interpretation questions and nontext information. Application and interpretation of knowledge with more complex analysis is not well processed by ChatGPT. Using problem solving, data interpretation, or case-study-based questions are ways to redesign assessment beyond knowledge-based questions. This disruptive technology can help educators to review the assessment approaches they implement. The ongoing developments with ChatGPT will provide rich discussions and considerations for learning and assessment. This initial investigation demonstrates the limitations of ChatGPT and a better understanding of its functionality in relation to chemistry assessment questions.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available at https://pubs.acs.org/doi/10.1021/acs.jchemed.3c00087.

Turnitin reports (PDF)
ChatGPT-generated responses (PDF)
ChatGPT-generated responses (DOCX)
Year 1 end-of-year exam assessments (PDF)
Year 2 end-of-year exam assessments (PDF)

AUTHOR INFORMATION

Corresponding Author

Suzanne Fergus − *School of Life and Medical Sciences, University of Hertfordshire, Hatfield AL10 9AB, United Kingdom;* orcid.org/0000-0002-7134-0665; Email: s.fergus@herts.ac.uk

Authors

Michelle Botha − *School of Life and Medical Sciences, University of Hertfordshire, Hatfield AL10 9AB, United Kingdom*

Mehrnoosh Ostovar − *School of Life and Medical Sciences, University of Hertfordshire, Hatfield AL10 9AB, United Kingdom*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jchemed.3c00087

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Thank you to colleagues with expertise in artificial intelligence who provided helpful information regarding initial queries.

REFERENCES

(1) Scholten, D. J.; Wijtmans, M.; Dekker, S. J.; Vuuregge, A. H.; Boon, E. J.; Vos, J. C.; Siderius, M.; Westbroek, H.; van Muijlwijk-Koezen, J. E. Practical Guidelines to Redesign Introductory Chemistry Courses Using a Flexible and Adaptive Blended Format. *J. Chem. Educ.* **2021**, *98* (12), 3852−3863.

(2) Limniou, M.; Varga-Atkins, T.; Hands, C.; Elshamaa, M. Learning, student digital capabilities and academic performance over the COVID-19 pandemic. *Education Sciences* **2021**, *11* (7), 361−376.

(3) McGunagle, D.; Zizka, L. Employability skills for 21st-century STEM students: the employers' perspective. *Higher education, skills and work-based learning* **2020**, *10* (10), 591−606.

(4) Fergus, S.; Botha, M.; Scott, M. Insights Gained During COVID-19: Refocusing Laboratory Assessments Online. *J. Chem. Educ.* **2020**, *97* (9), 3106−3109.

(5) Dicks, A. P.; Morra, B.; Quinlan, K. B. Lessons learned from the COVID-19 crisis: adjusting assessment approaches within introductory organic courses. *J. Chem. Educ.* **2020**, *97* (9), 3406−3412.

(6) Koh, J. H. L.; Daniel, B. K. Shifting online during COVID-19: A systematic review of teaching and learning strategies and their outcomes. *International Journal of Educational Technology in Higher Education* **2022**, *19* (1), 56.

(7) Holme, T. A. Introduction to the Journal of Chemical Education special issue on insights gained while teaching chemistry in the time of COVID-19. *J. Chem. Educ.* **2020**, *97*, 2375−2377.

(8) Sambell, K.; Brown, S. Changing landscape of assessment: some possible replacements for unseen, time constrained, face-to-face invigilated exams. Staff and Educational Development Association. https://www.seda.ac.uk/wp-content/uploads/2021/04/Paper-3-The-changing-landscape-of-assessment-some-possible-replacements-for-unseen-time-constrained-face-to-face-invigilated-exams-4.pdf (accessed 2023-01).

(9) Nguyen, J. G.; Keuseman, K. J.; Humston, J. J. Minimize online cheating for online assessments during COVID-19 pandemic. *J. Chem. Educ.* **2020**, *97* (9), 3429−3435.

(10) Ellis, C.; Van Haeringen, K.; Harper, R.; Bretag, T.; Zucker, I.; McBride, S.; Rozenberg, P.; Newton, P.; Saddiqui, S. Does authentic assessment assure academic integrity? Evidence from contract cheating data. *Higher Education Research & Development* **2020**, *39* (3), 454−469.

(11) Bretag, T.; Harper, R.; Burton, M.; Ellis, C.; Newton, P.; van Haeringen, K.; Saddiqui, S.; Rozenberg, P. Contract cheating and assessment design: exploring the relationship. *Assessment & Evaluation in Higher Education* **2019**, *44* (5), 676−691.

(12) Lancaster, T.; Cotarlan, C. Contract cheating by STEM students through a file sharing website: a Covid-19 pandemic perspective. *International Journal for Educational Integrity* **2021**, *17* (1), 3.

(13) *ChatGPT*. https://chat.openai.com/ (accessed 2023-01).

(14) Anderson, L. W.; Krathwohl, D. R.; Airasian, P. W.; Cruikshank, K. A.; Mayer, R. E.; Pintrich, P. R.; Raths, J.; Wittrock, M. C. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*, abridged ed.; Longman: White Plains, NY, 2001.

(15) *Turnitin*. https://www.turnitin.com/ (accessed 2023-01).

(16) Jiang, Z.; Xu, F. F.; Araki, J.; Neubig, G. How can we know what language models know? *Transactions of the Association for Computational Linguistics* **2020**, *8*, 423−438.