

ChatGPT in Nuclear Medicine Education

Geoffrey Currie¹ and Kym Barry²

¹Charles Sturt University, Wagga Wagga, New South Wales, Australia; and ²Charles Sturt University, Port Macquarie, New South Wales, Australia

Academic integrity has been challenged by artificial intelligence algorithms in teaching institutions, including those providing nuclear medicine training. The GPT 3.5–powered ChatGPT chatbot released in late November 2022 has emerged as an immediate threat to academic and scientific writing. **Methods:** Both examinations and written assignments for nuclear medicine courses were tested using ChatGPT. Included was a mix of core theory subjects offered in the second and third years of the nuclear medicine science course. Long-answer-style questions (8 subjects) and calculation-style questions (2 subjects) were included for examinations. ChatGPT was also used to produce responses to authentic writing tasks (6 subjects). ChatGPT responses were evaluated by Turnitin plagiarism-detection software for similarity and artificial intelligence scores, scored against standardized rubrics, and compared with the mean performance of student cohorts. **Results:** ChatGPT powered by GPT 3.5 performed poorly in the 2 calculation examinations (overall, 31.7% compared with 67.3% for students), with particularly poor performance in complex-style questions. ChatGPT failed each of 6 written tasks (overall, 38.9% compared with 67.2% for students), with worsening performance corresponding to increasing writing and research expectations in the third year. In the 8 examinations, ChatGPT performed better than students for general or early subjects but poorly for advanced and specific subjects (overall, 51% compared with 57.4% for students). **Conclusion:** Although ChatGPT poses a risk to academic integrity, its usefulness as a cheating tool can be constrained by higher-order taxonomies. Unfortunately, the constraints to higher-order learning and skill development also undermine potential applications of ChatGPT for enhancing learning. There are several potential applications of ChatGPT for teaching nuclear medicine students.

Key Words: artificial intelligence; tertiary education; higher education; academic integrity; generative AI; language model

J Nucl Med Technol 2023; 00:1–8
DOI: 10.2967/jnmt.123.265844

Although contract cheating and ghostwriting in academic or scientific writing are not new concepts, they have become more efficient with advances in information technology (1). Nuclear medicine technologist or scientist students and authors are not immune to this scourge. At the heart of the issue is academic integrity. There is potential

for significant reputational damage to institutions when authorship is claimed for work that has been produced by another or assessment is fraudulent. For our students, public safety is an issue if graduates cheat to produce evidence of skills and capabilities (2). Indeed, contract cheating among university students has reached epidemic proportions with developments in artificial intelligence (AI) algorithms and with the coronavirus disease 2019 pandemic having driven a move to online or flexible learning and assessment.

ChatGPT (OpenAI) overcomes the limitations of early algorithms for generative writing and contract cheating sites. The ghostwriting capability of ChatGPT poses an immediate threat to the academic integrity of student assessments despite being publicly released only recently, on November 30, 2022. Less than 2 mo after the launch, ChatGPT had more than 100 million users (3). Numerous universities and colleges have reacted to the emergence of ChatGPT by banning its use. Banning use to prevent misuse also eliminates ChatGPT as a potential tool for enhancing learning and writing.

The role of ChatGPT and other AI tools in education is not an easy debate. AI can significantly enhance student learning and capability development and should be supported from that front-door context because it is aligned with the underlying goals and strategies of a teaching institution. Nonetheless, AI use can hide lack of understanding or can fabricate evidence of capability that does not exist—a misuse that should not be acceptable because, at the back door, it undermines the evidence that students meet the graduate outcomes. Indeed, this misuse may relate to the definition of AI; when the term *AI* is used to mean “artificial intelligence,” the student has not developed real knowledge or capability, but when the term is used to mean “augmented intelligence,” student understanding and capability have been enhanced. In either case—misuse or enhanced learning—access to ChatGPT can create inequity typical of the social asymmetry for AI in education and health (4).

The suitability of ChatGPT as an educational tool also needs to consider the currency of GPT 3.5, which powers ChatGPT. At the time of writing, the publicly available ChatGPT learning cutoff date was September 2021. ChatGPT does not have real-time access to information, including the Internet, and does not learn new information based on user input. This limitation is particularly important in nuclear medicine because the field enjoys rapidly advancing technology and techniques; ChatGPT responses may not reflect information currency. For the new edition of ChatGPT powered by

Received Apr. 12, 2023; revision accepted May 13, 2023.
For correspondence or reprints, contact Geoffrey Currie (gcurrie@csu.edu.au).
Published online Jul. 11, 2023.
COPYRIGHT © 2023 by the Society of Nuclear Medicine and Molecular Imaging.

GPT 4, accuracy is improved by 60%, including enhanced interpreting of context and reasoning. When available to the public, ChatGPT powered by GPT 4 will allow voice and image inputs for interpretation, will correct or write code, will allow 25,000 words to be inputted for editing and refinement, and will produce outputs of up to 52 pages (3 is the current limit), all of which will broaden the applications and flexibility for students in nuclear medicine (5,6).

Numerous universities, including Charles Sturt University, have banned use of ChatGPT. The challenge remains in policing such bans, especially in an era of online or flexible learning and open-book, noninvigilated online examinations. An opposing view relates to authentic assessment and learning. With ChatGPT emerging as a tool for use among nuclear medicine professionals in the clinical and research environment, should assessment not afford that same environment? ChatGPT could enhance student critical thinking, problem solving, and writing skills and could be especially helpful when English is a second language. ChatGPT could craft realistic scenarios for case-based learning, help personalize learning, and distil complex learning topics (e.g., from textbooks or lectures) for improved understanding (3). Deeper insight into potential misuse is required before these potential benefits are discarded.

MATERIALS AND METHODS

To evaluate the capabilities of ChatGPT use among undergraduate nuclear medicine students, a sample of the theory-based assessment requirements for second- and third-year undergraduate subjects was used. There are no nuclear medicine-specific subjects in the first year. The subjects included a general second-year subject ("Imaging Pathology") and 3 nuclear medicine-specific subjects ("Nuclear Medicine 1," "Radiopharmacy," and "Instrumentation"). Additionally, 2 general third-year subjects ("CT" and "Pharmacology") and 2 nuclear medicine-specific subjects ("Nuclear Medicine 2" and "Nuclear Medicine 3") were sampled. For each of the 8 subjects, final examination questions were individually entered into ChatGPT. Additionally, written assessment tasks for 6 of the subjects were also entered into ChatGPT, along with the task expectations and requirements (e.g., topic, fully referenced, word count, specific inclusions). Both the subject "Radiopharmacy" and the subject "Pharmacology" also had a second examination comprising calculation-based questions that were individually entered into ChatGPT. The copy-and-paste function was used to lift questions into the ChatGPT window. Examination and written task answers provided by ChatGPT were transferred to an examination sheet and sent for scoring against the standard rubric for each task. Scoring was out of sequence with actual student submissions, and as a result, scoring was not masked; the scorers were aware that the submission was from ChatGPT. This lack of masking could produce a bias in results; however, all scorers were required to score objectively against the standardized rubric and against the expectations for each question and to justify those scores for moderation. Consequently, the scores are expected to be representative and realistic compared with the corresponding student cohort.

Turnitin software detects plagiarism (similarity report) and generates an AI score. This function was introduced in April 2023 to combat generative AI in academic submissions. The score represents the percentage of the submission that Turnitin is 98% certain

was generated by AI. Each of the examinations and written tasks was submitted to Turnitin, and both similarity reports and AI reports were generated.

RESULTS

Both the second-year subject "Radiopharmacy" and the third-year subject "Pharmacology" had calculation examinations with passing scores of 60%. ChatGPT was particularly poor at calculation-style questions. For the subject "Radiopharmacy," ChatGPT was particularly poor, with a score of 24.0% compared with a student mean of 67.3% (Fig. 1). This comprised 31.7% in short calculations and 8.7% in more complex problems. ChatGPT was particularly confounded by decay calculations and on several occasions performed the calculations starting with the premise "assuming no decay" for ^{99m}Tc with several hours of nonnegligible decay. Indeed, even when prompted to recalculate incorporating decay, ChatGPT produced incorrect answers. For the subject "Pharmacology," among the shorter calculation questions, ChatGPT provided the correct answers with full working for 92.7% of available scores but was unable to provide correct answers for any of the more complex questions (zero scores). Overall, in the calculation examination, ChatGPT received a score of 38.8% whereas the mean among 81 students was 67.6%. For several of the more complex questions, ChatGPT had the correct formula and the correct numbers in the formula but the wrong answer, which then impacted subsequent calculations; it got the

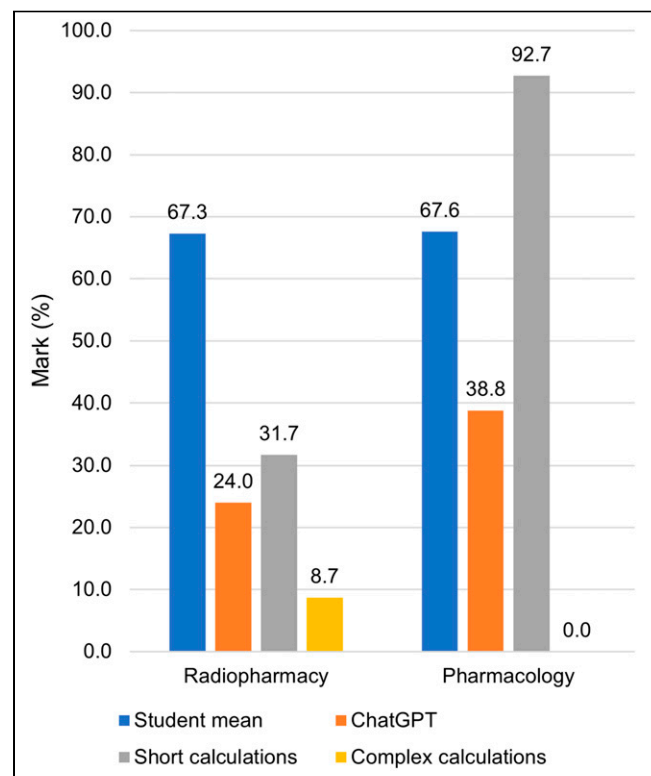


FIGURE 1. Bar chart for student mean and ChatGPT score for each calculation examination, including performance for short problems and complex problems.

simplest part incorrect. Interestingly, ChatGPT handled first-order concentration calculations in the course “Pharmacology” at a higher standard than decay questions in the subject “Radiopharmacy” although the mathematics were identical:

$$C = C_0 e^{-k \cdot t} \text{ vs. } A = A_0 e^{-\lambda \cdot t}$$

The 6 written assignment tasks were scored against the task rubric (Fig. 2). In all 6 subjects, ChatGPT scored significantly more poorly than the mean student score. A general trend suggested that the gap between student mean and ChatGPT scores widened with course progression, indicating that students were developing research and writing skills against higher-order expectations. Each subject was scored independently; however, the feedback in scoring rubrics was consistent for ChatGPT. For example, answers provided shallow insight not connected strongly to practice; research was shallow and narrow, which left answers well short of expectations; language for some aspects of answers was deemed colloquial rather than professional; significant portions of text for which a supporting citation would be expected had no referencing; and there was some repetition without connection, resulting in writing that was not integrated in nature although it did read seamlessly. In addition, the responses were well short of the word count (950), reflecting the lack of depth in discussion and insight that would connect to student or professional capabilities; there was lack of currency of insights, information, and references, creating a significant barrier to both quality and academic integrity; and there was reliance on obscure or fabricated references in preference to mainstream literature, along with omission of key citations from the professional literature.

Among 112 students in the subject “Imaging Pathology,” the mean student score for the writing task was 65.9% whereas ChatGPT scored 49.3%. Among 12 students in the

subject “Nuclear Medicine 1,” the mean student score for the writing task was 69.0% whereas ChatGPT scored 41.0%. Among 13 students in the subject “Instrumentation,” the mean student score for the writing task was 71.0% whereas ChatGPT scored 46.0%. For the third-year subjects, among 81 students in the subject “Pharmacology,” the mean student score for the writing task was 67.7% whereas ChatGPT scored 26.0%. Among 12 students in the subject “Nuclear Medicine 2,” the mean student score for the writing task was 66.0% whereas ChatGPT scored 41.2%. Among 11 students in the subject “Nuclear Medicine 3,” the mean student score for the writing task was 63.5% whereas ChatGPT scored 30.0%. There was a statistically significant difference between the student scores and the ChatGPT scores, with the mean score being 28.3% lower for ChatGPT ($P < 0.001$). Although all 6 ChatGPT written task were well short of expectations, they reflect authentic student submissions, sharing close parallels to failing-grade submissions from students who leave the task to the final hour and hastily cobble together a shallow and poorly researched response. This finding questions the capacity of ChatGPT to be used to enhance student writing and research skills at the university level (Fig. 3), with its benefits perhaps limited to the high school level. A key issue common across the written tasks was the lack of in-text citations, which for a student would constitute plagiarism. Furthermore, ChatGPT had a tendency to fabricate references that cannot be verified or found. Such fabrication, if done by a student on a submission, would constitute serious fraud and misconduct.

For the 8 written examinations across the 2y of theoretic study (the fourth year of the course is a residency focused on capability development rather than theory mastery), scoring was analyzed individually and collectively (Fig. 3). For the second-year subject “Imaging Pathology,” the mean among 112 students was 44.3%, compared with 55.7% for ChatGPT. For the subject “Nuclear Medicine 1,” the mean among 12 students was 66.8%, compared with 72.5% for ChatGPT. For the subject “Radiopharmacy,” the mean among 12 students was 60.0%, compared with 55.2% for ChatGPT. For the subject “Instrumentation,” the mean among 13 students was 60.4%, compared with 47.1% for ChatGPT.

For third-year students, for whom theoretic learning represents minimum standards for a practitioner, 4 subjects were evaluated. For the subject “CT,” the mean among 89 students was 54.1%, compared with 37.9% for ChatGPT. For the subject “Pharmacology,” the mean among 81 students was 57.5%, compared with 59.1% for ChatGPT. For the nuclear medicine-specific third-year subjects, the mean

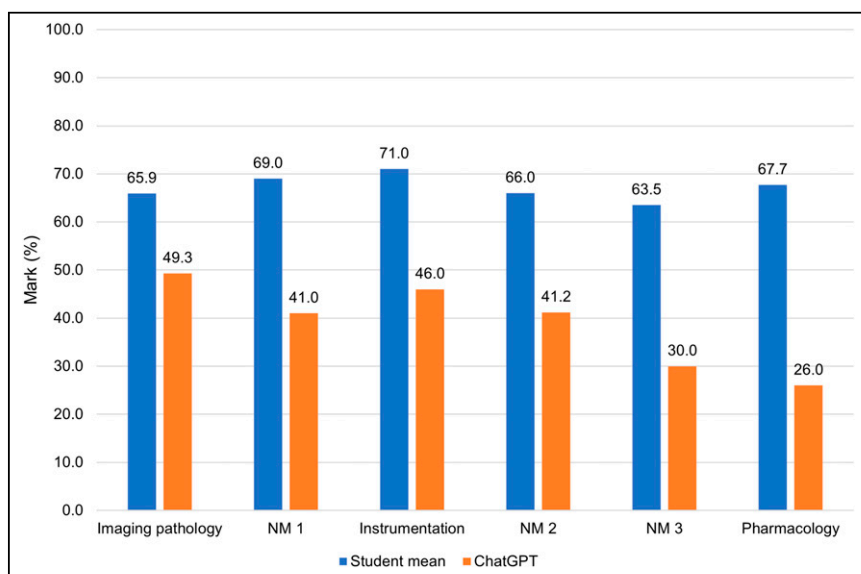


FIGURE 2. Bar chart for student mean and ChatGPT score for each of 6 subjects that had written tasks evaluated. NM = nuclear medicine.

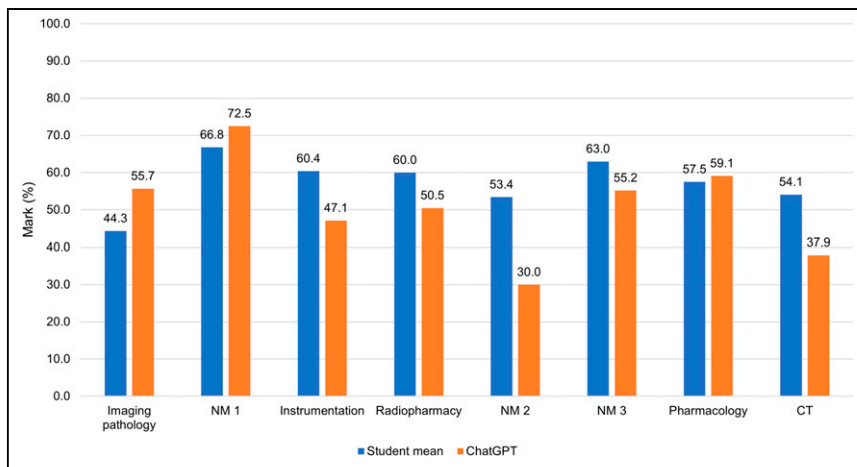


FIGURE 3. Bar chart for student mean and ChatGPT score for each of 8 subjects that had examinations evaluated. NM = nuclear medicine.

among 12 students in the subject “Nuclear Medicine 1” was 53.4%, compared with 30% for ChatGPT, and the mean among 11 students in the subject “Nuclear Medicine 3” was 63.0%, compared with 55.2% for ChatGPT.

There was a statistically significant difference between the student scores and the ChatGPT scores, with the mean score being 6.4% lower for ChatGPT ($P = 0.009$). Despite this lower mean score, ChatGPT performed better than the student mean in 3 of the subjects: “Imaging Pathology,” “Nuclear Medicine 1,” and “Pharmacology.” Each of these subjects has content that is well established, and understanding of the content is reflected by describing processes, for which ChatGPT is well equipped. “Nuclear Medicine 1” is the first clinical nuclear medicine subject that students undertake, and learning outcomes represent lower-order taxonomies that are well handled by ChatGPT. The subject “Pharmacology” covers important content, with the expectation being more of acquiring a working understanding

than of achieving mastery; as such, the examination questions tend to be more superficial and to cover content that is well established. The remaining subjects saw ChatGPT perform worse than the student mean. These subjects are specific in nature and require mastery of content that requires not only deep insights not typical of ChatGPT but also command of current innovations and developments occurring outside the training data for ChatGPT.

Turnitin generated similarity scores that ranged from 3% to 18% for examinations and from 13% to 34% for written assignments. The difference between lower and higher similarity scores related to the question itself for examinations and to the reference list for assignments. No instances of plagiarism were identified. Conversely, the AI scores ranged from 9% to 75% for examinations (although the 9% was an outlier, with the next lowest being 43%) and from 47% to 100% for written assignments (Fig. 4). For normalization, the entire introduction above was assessed through Turnitin, with a 0% similarity index and 0% AI score.

DISCUSSION

The performance of ChatGPT in nuclear medicine assessment was enlightening. ChatGPT performed well when brief and shallow insights were required, typical perhaps of first-year subject topics and early or general aspects of second-year subject topics. Third-year topics are far from the shallows, and as a result, the expectations of depth and insight were beyond the capabilities of ChatGPT, even when prompted by a specific word count. Importantly from an education perspective, ChatGPT provided answers with no evidence, with outdated evidence, or with fabricated evidence. Such

answers not only are devoid of current insight into the fluid nuclear medicine environment but also would represent academic misconduct if used by students in their submissions. ChatGPT was particularly poor for calculations, despite providing convincing working and justification for the same. Our findings are consistent with those reported for medical examinations, for which ChatGPT scored 43%–68% in open-ended questions and 40%–65% in multiple-choice questions (7,8). The authors similarly reported lower scores correlating with more complex questions.

Alarming, one of the chief benefits (and risks) of ChatGPT is in written tasks, for which ChatGPT was shown

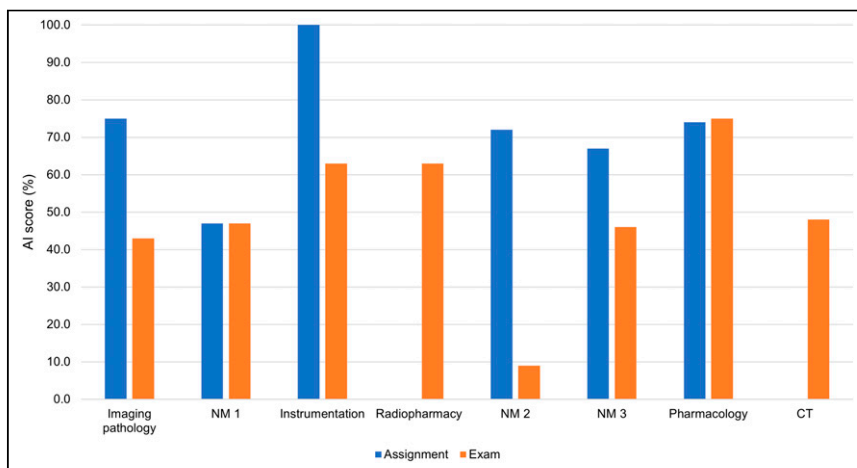


FIGURE 4. Bar chart for Turnitin AI score for ChatGPT assignments for 6 subjects that had ChatGPT assignments generated and ChatGPT examination answers for each of 8 courses that had examinations generated. NM = nuclear medicine.

to perform well short of expectations across all levels and courses. The depth of research and writing, the insight and understanding demonstrated in the writing, and the writing style itself (e.g., professional language and tone, integration across the piece, and integration with practice) not only penalized scoring but could not be used to positively impact skill development in students. This shortfall raises serious questions about using ChatGPT to generate questions (revision or assessment) when it lacks the insight to answer them. The performance of ChatGPT on very general topics or at a lower level of education (e.g., high school) might be better, and in the current evaluation, ChatGPT performed well on topics requiring shallow information and on topic that were widely evidenced before September 2021 (e.g., the role of bone scans in prostate cancer). Nonetheless, nuclear medicine education requires depth and specific detail in a rapidly evolving domain that confounds ChatGPT. It is possible for students to plan sections or topics within a written response and ask ChatGPT more targeted questions to produce a higher-scoring paper; however, it will remain constrained by lack of depth and insight, language that is less than professional, incorrect information, and inadequate or fraudulent referencing. Regardless of well-constructed responses or poorly constructed responses, Turnitin software confidently predicted when responses were AI-generated.

In the hands of a student already performing at a passing level or better, ChatGPT is unlikely to boost grades; indeed, it may reduce grades and risk academic misconduct. In theory, when norm-based referenced testing is used, the class

mean improves and those using ChatGPT have an increased representation in the higher grades, potentially relegating non-ChatGPT users to the lower grades. For criterion-based referenced testing, it could allow student performance to improve independently of the class performance. Those using ChatGPT are advantaged but not at the expense of the grades of those not using ChatGPT. In reality, the current version of ChatGPT (GPT 3.5) does not provide that capability, and students relying on ChatGPT for either cheating or enhancing responses are likely to be penalized in their scores, independently of academic misconduct issues.

On the basis of the evidence in this evaluation, ChatGPT does not pose a risk of masking the students' shortcomings against the learning outcomes. Among nuclear medicine students, ChatGPT also does not appear to enhance grades by honing skills in writing and expanded learning. The significant risk to academic integrity is raised by poor, absent, or fraudulent referencing in ChatGPT responses. ChatGPT does not raise concerns about students graduating without the requisite knowledge and skills for safe clinical practice because reliance on ChatGPT will not allow the student to accrue a passing grade. It is appropriate to reevaluate the ChatGPT claimed benefits to student learning and assess whether these are conceivable (Table 1). Among the potential roles of ChatGPT for nuclear medicine student learning, the following would likely be the most appropriate: use of ChatGPT as a language assistant or language practice tool for those for whom English is a second language (for this role, ChatGPT would be developing basic English language

TABLE 1
Summary of Potential Applications of ChatGPT in Nuclear Medicine Student Education, with Evaluation of That Capability Against Findings of This Investigation

ChatGPT capability	Evidence	Comment
Language assistant	✓	Is useful to develop language skills when English is second language
Accessibility	✓	Is useful for supportive technology
Generation of lecture notes	?	Could fill gaps but requires confidence in accuracy and currency of content
Interactive learning	?	Requires confidence in accuracy and currency of content
Personalized learning	?	Is useful perhaps for first-year students performing at lower level but not for advanced students
Group collaboration	?	Is useful perhaps for first-year students but lacks insights and depth to facilitate meaningful discussion
Feedback	?	Requires confidence in accuracy and currency of content
Question answering	?	Requires confidence in accuracy and currency of content
Question creation	?	Requires confidence in accuracy and currency of content
Suggestion of group discussion topics	?	Requires confidence in accuracy and currency of content
Provision of case studies	?	Is useful perhaps for basic studies but not for novel, complex, or new applications or old applications with refined approaches
Independent learning	×	Lacks currency, depth of insight, and accuracy of information to benefit students
Assignment help	×	Provides writing that is inadequate, research that is shallow, and information that is not current or accurate
Information and resources	×	Provides research that is shallow and information that is not current or accurate
Writing assistance	×	Lacks professionalism, does not target audience, and ignores important writing conventions (e.g., referencing)

✓ = applications our evidence supports; × = applications for which evidence suggests no role; ? = applications for which evidence suggests limited role.

skills, not professional language or writing skills); use of ChatGPT to support accessibility for students with disabilities, such as through assistive technology for text-to-voice conversion; use of ChatGPT as a training tool in which ChatGPT responses are applied in clinical, theory, or research learning domains to have students critique, refine, and correct; use of ChatGPT to proof, test, and refine assessments and scoring rubrics as part of moderation; and use of ChatGPT to simulate conversations with patients or caregivers as a form of authentic assessment.

Academic misconduct concerns for ChatGPT have been related largely to the potential capability to generate examination answers or responses to written tasks; that is, cheating for advantage. For nuclear medicine subjects, the risk of this use appears to be low, with limitations to the ChatGPT capability. This use could be further limited by structuring assessments that target student insight and understanding at a deeper level and by setting the minimum standards for a passing score in the rubric more rigorously against learning outcomes. That is, if a learning outcome for a subject requires students to be able to demonstrate their understanding by explaining a concept, then a student response that does not explain that concept (lists or outlines key information or facts) or does not show understanding (has errors or lacks integration with

practice) should receive a failing score for that question or task. Typically in a rubric, there is some wiggle room that would recognize some knowledge and produce a passing grade with a credit, perhaps representing what should be the minimum standard (Table 2). ChatGPT will be able to produce those passing grades in which the credit expectations in a rubric represent what should be a minimum passing-grade requirement. Adjusting this approach would minimize the risk that ChatGPT will be used to navigate through subject assessment and would better align assessment with learning outcomes. Perhaps the biggest academic integrity issue for ChatGPT use is the potential plagiarism or fraud that students using generated responses could confront. Writing and referencing are typically below the expected standards; a paucity of citations would represent plagiarism, and the tendency for ChatGPT to add citations that cannot be verified is potential fraud. Indeed, ChatGPT can simply fabricate answers.

Although ChatGPT has surprising accuracy for some topics, it is prone to several different types of errors, and these were apparent throughout this investigation when the specific nature of nuclear medicine confounded ChatGPT responses. The term *hallucination* has been used widely in AI to refer to false or misleading information, yet the term is more specific,

TABLE 2
Selection of Learning Outcomes for Specific Examination or Assignment Tasks Across Some Evaluated Courses

Learning outcome descriptor	Assessment type	Current credit standard	Current passing-score standard
Describe, explain, and implement	Written task	Compare data across some relevant primary literature, explain some relationships across data, and summarize key concepts	Compare data with some relevant literature, describe some relationships among data, and state some key concepts
Identify, explain, and apply	Examination	Recall mostly accurate facts, generally apply relevant knowledge correctly within context most of the time, explain important relationships across data most of the time, and provide plausible interpretation	Recall some accurate facts, apply relevant knowledge correctly within context some of the time, explain important relationships across data some of the time, and provide interpretation
Explain and apply principles of	Written task	Review and contextualize knowledge, with capacity for explanation of practice	Review and describe knowledge, with capacity for discussion of practice
Demonstrate understanding of	Examination	Generally apply relevant knowledge correctly within context most of the time	Apply relevant knowledge correctly within context some of the time
Describe, explain, evaluate, and develop	Examination	Indicate sound knowledge and application of concepts through descriptions, explanations, and discussion of content	Indicate knowledge and application of emerging concepts through descriptions of content
Explain and apply	Written task	Demonstrate sound understanding of themes and course matter; use recent external literature to demonstrate sound comprehension of themes	Demonstrate limited understanding of themes and course matter; use external literature to demonstrate limited understanding of themes
Recognize and describe	Examination	Recall facts from course material accurately, with limited but accurate application of facts to presented scenarios	Recall facts from course materials and record in basic manner; attempt to apply facts to scenarios presented

[AQ9]

Data show misalignment with minimum passing-score standards. In each case, it could be argued that to meet learning outcome, credit standard should be minimum passing-score criteria.

referring to a plausible response that is incorrect (it seems correct to ChatGPT but is not—identifying a stick that is not there) (8). Other types of AI errors have also been named according to this original psychiatry analogy:

- *Illusion* is like a hallucination except it is an error of similarity (mixing up similar items)—mistaking a piece of rope for a stick.
- *Delusion* is false belief or error (wrong information)—after correction and examination, insisting the rope is a stick.
- *Delirium* is either a sophisticated or a nonsensical answer because the algorithm is overwhelmed or confused—describing a ball when asked to describe a stick.
- *Confabulation* or *lying* is fabrication of information—photoshopping images of sticks in a scene to provide evidence that sticks were there without actually checking the scene for real sticks.
- *Extrapolation* or *interpolation* is a logical, although incorrect, extension of known information—based on 3 dogs carrying sticks, declaring a fourth dog also is carrying a stick without seeing the dog.
- *Miscalculation* or *blunder* is a computational error despite the correct equation and data—there are 5 sticks but 6 are counted.

All of these error types were evidenced through the scoring of examinations and written assignments. Students leaning on ChatGPT are ill-equipped to identify such errors. These students risk undermining their understanding and ongoing learning. Information is not knowledge.

The GPT 4-powered chatbot could be a bigger problem, with easier, faster, and more accurate responses. GPT 4 allows voice-to-text conversion, which would enable students to simply read the question to ChatGPT. GPT 4 will allow importing of images for analysis and interpretation, a task that would previously have confounded ChatGPT. ChatGPT is prone to hallucinations (false or misleading information), which a student using ChatGPT will not be aware enough to correct. GPT 4 is 60% more accurate with answers and, in particular, with interpreting context and reasoning. GPT 4 is trained on 3 times more data with a 500-fold increase in capacity, which will allow greater originality and accuracy of responses, confounding even the best plagiarism and AI detection software.

For nuclear medicine courses, there appears to be no advantage to students who misuse ChatGPT; indeed, they will be disadvantaged. It is important, therefore, to educate students about use, misuse, risk versus absence of benefit, professional responsibilities to their future patients, and the consequences, now or in the future, of cheating (e.g., if new technology is developed in 10 y that allows retrospective detection of ChatGPT, degree disqualification and deregistration could and should be a consequence). Education of students about ChatGPT would allow it to be integrated into the learning environment to support students when appropriate and to be used as a learning tool. This use is best supported by reengineering assessments and by recrafting learning outcomes to be both authentic and capability-focused, independently of whether ChatGPT is used.

Although not available at the time of writing, Google plans to release its AI chatbot, Bard, which would be in competition with ChatGPT. Although the chief comparisons relate to their chatbot functions for Google and Bing search engines, respectively, both have competing capabilities as a text generator. It is unrealistic to compare Bard with GPT 3, with Bard having capabilities mirroring the recently released GPT 4; these include context interpretation, image analysis, and mathematic problem solving. There are similar issues with accuracy and bias that need to be interrogated in the public user arena. Unlike GPT 4, Bard lacks plagiarism detection or prevention, accesses the Internet in real time, and updates the corpus of knowledge, providing currency at the expense of increased misinformation and bias. It is a reasonable prediction that ChatGPT will be the preferred tool for generative text for academic and scientific writing whereas Bard may emerge with broader applications in text generation (e.g., list generation, agenda production, and scheduling tasks) and image or video creativity for personal and general professional purposes.

CONCLUSION

ChatGPT is an exciting educational tool that has limited generative capability to assist student writing in the nuclear medicine setting because of limitations on depth of insight, breadth of research, and currency of information. It is particularly inadequate at producing written assessment tasks (e.g., literature reviews) and introduces student risk of misconduct associated with inadequate referencing practices. ChatGPT could be used to build examination answers in real time, but performance is limited to superficial learning evidence produced by shallow or general answers. These limitations that reduce the risk that students will benefit from cheating also limit the educational benefit of ChatGPT for enhancing learning and writing skills. There are, however, several applications of ChatGPT that can enrich student learning in nuclear medicine. GPT 4 will require reimaging of the AI-augmented learning space.

ACKNOWLEDGMENT

As a language model, ChatGPT should not be included as an author on journal articles. Authorship implies a contribution to the content of the article, and although ChatGPT may have been used to generate some of the text, it is not an individual who has contributed to the intellectual content of the article. ChatGPT does not meet the authorship requisites recommended by the International Committee of Medical Journal Editors. ChatGPT does warrant an acknowledgment. We would like to acknowledge the contribution of ChatGPT (version 3.5), a language model developed by OpenAI (<https://openai.com/>), in generating some of the text in this article. The model was accessed between April 1 and April 7, 2023. We would also like to acknowledge Turnitin plagiarism-detection software (www.turnitin.com), used for similarity reports and AI scores (April 4, 2023, release).

KEY POINTS

QUESTION: Does ChatGPT pose a risk to academic integrity?

PERTINENT FINDINGS: ChatGPT powered by GPT 3.5 lacks the capability to provide responses that reflect the depth, breadth, and currency of information; research expectations; and the appropriate professional tone.

IMPLICATIONS FOR PRACTICE: ChatGPT has limited scope for cheating among nuclear medicine students, which also limits the potential beneficial applications of ChatGPT in enhancing learning.

REFERENCES

1. Awdry R, Ives B. International predictors of contract cheating in higher education. *J Acad Ethics*. 2023;21:193–212.
2. Stone A. Student perceptions of academic integrity: a qualitative study of understanding, consequences, and impact. *J Acad Ethics*. 2022;28:1–19.
3. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Med Educ*. 2023;6;9:e46885.
4. Currie G, Rohren E. Social asymmetry and artificial intelligence: the nuclear medicine landscape. *Semin Nucl Med*. 2022;52:498–503.
5. Graham F. Daily briefing: what scientists think of GPT-4, the new AI chatbot. *Nature*. March 17, 2023 [Epub ahead of print].
6. Sanderson K. GPT-4 is here: what scientists think. *Nature*. 2023;615:773.
7. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2:e0000198.
8. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*. 2023;15:e35179.