GPT-4 in Nuclear Medicine Education: Does It Outperform GPT-3.5?

Geoffrey M. Currie

Charles Sturt University, Wagga Wagga, New South Wales, Australia

The emergence of ChatGPT has challenged academic integrity in teaching institutions, including those providing nuclear medicine training. Although previous evaluations of ChatGPT have suggested a limited scope for academic writing, the March 2023 release of generative pretrained transformer (GPT)-4 promises enhanced capabilities that require evaluation. **Methods:** Examinations (final and calculation) and written assignments for nuclear medicine subjects were tested using GPT-3.5 and GPT-4. GPT-3.5 and GPT-4 responses were evaluated by Turnitin software for artificial intelligence scores, marked against standardized rubrics, and compared with the mean performance of student cohorts. **Results:** ChatGPT powered by GPT-3.5 performed poorly in calculation examinations (31.4%), compared with GPT-4 (59.1%). GPT-3.5 failed each of 3 written tasks (39.9%), whereas GPT-4 passed each task (56.3%). **Conclusion:** Although GPT-3.5 poses a minimal risk to academic integrity, its usefulness as a cheating tool can be significantly enhanced by GPT-4 but remains prone to hallucination and fabrication.

Key Words: artificial intelligence; scientific writing; patient education; higher education; academic integrity; generative algorithms

J Nucl Med Technol 2023; 51:314–317
DOI: 10.2967/jnmt.123.266485

ChatGPT (OpenAI) is a versatile and powerful large language model driven by a generative pretrained transformer (GPT) (*1*). The opportunities, limitations, and challenges of ChatGPT have been previously outlined (*2*,*3*). Since its public release on November 30, 2022, there has been an unprecedented adoption rate among new users (*4*). In education, the key issue continues to be the trade-off between benefits (e.g., research, writing, and problem solving) and misuse (scientific fraud, cheating, and academic integrity) (*2*). Consistent with the ChatGPT and artificial intelligence (AI) hype, the purported benefits of ChatGPT were considered a threat to academic integrity, and many educational institutions banned its use (*2*).

In previous reports adopting the GPT-3.5 architecture, examinations and written tasks on undergraduate nuclear medicine and radiography subjects were performed using ChatGPT, marked against standard rubrics, and compared with student cohort means (*3*,*5*). Although ChatGPT performed more poorly than the average student, ChatGPT performance worsened as the task expectations increased. ChatGPT was reported to have limited capacity to assist student assessment because answers lacked the appropriate depth of insight, breadth of research, and currency of information and because the risk of misconduct (plagiarism and information fabrication) was virtually certain (*5*). Among nuclear medicine–specific undergraduate subjects, ChatGPT was reported to perform poorly in calculation examinations (mean of 31.7% compared with 67.3% for students) and written tasks (mean of 38.9% compared with 67.2% for students) (*3*). The findings of these evaluations suggested that for nuclear medicine subject content, there appears to be no academic advantage to students who misuse ChatGPT, whereas paradoxically, ChatGPT misuse is likely to prove disadvantageous to the student (*3*,*5*,*6*).

One should consider that the bulk of research on ChatGPT, including predecessor publications to this article, relates to the GPT-3.5 architecture. Such research conclusions generally provide a warning that GPT-4 will bring superior capability (*3*,*5*). GPT-4 was released in March 2023 to paid users, although some of the features were not yet enabled (e.g., image inputs). It should also be noted that the current subscription version of GPT-4, although adopting the advanced architecture, is also constrained by the September 2021 learning cutoff of GPT-3.5. GPT-4 problem solving and accuracy are superior (by 60%) to those of GPT-3.5 (*6*,*7*). The recently published evaluations of capabilities and limitations of ChatGPT need to be repeated for GPT-4.

MATERIALS AND METHODS

The capabilities of ChatGPT powered by GPT-4 were evaluated and compared with both student means and GPT-3.5 using a sample of assessment tasks from 4 second- and third-year subjects in an undergraduate nuclear medicine science course. The subjects included second-year radiopharmacy and instrumentation and third-year nuclear medicine and pharmacology. For each of the 4 subjects, final examination questions were individually entered into ChatGPT powered by GPT-4 and separately for GPT-3.5. Additionally, written assessment tasks for 3 of the subjects were also entered into ChatGPT, along with the task expectations and requirements. Both radiopharmacy and pharmacology also had separate calculation examinations that were entered into ChatGPT individually. Responses provided by ChatGPT were marked against the

Received Aug. 2, 2023; revision accepted Sep. 12, 2023.
For correspondence or reprints, contact Geoffrey M. Currie (gcurrie@csu.edu.au).
Published online Oct. 18, 2023.

standard rubric for each task. Additionally, written assessment tasks were entered into the Turnitin Similarity Report (Turnitin LLC) plagiarism portal (similarity report) to generate an AI score (*3*).

RESULTS

The second-year nuclear medicine instrumentation subject (unit of study) included 13 student responses for both a written assignment and a final examination. The third-year clinical nuclear medicine subject included 11 student responses for both a written assignment and a final examination. The second-year radiopharmacy subject included 12 student responses for both a calculation examination and a final examination. The third-year pharmacology subject included 81 medical radiation science student responses for a written assignment, a calculation examination, and a final examination. The results are summarized in Table 1.

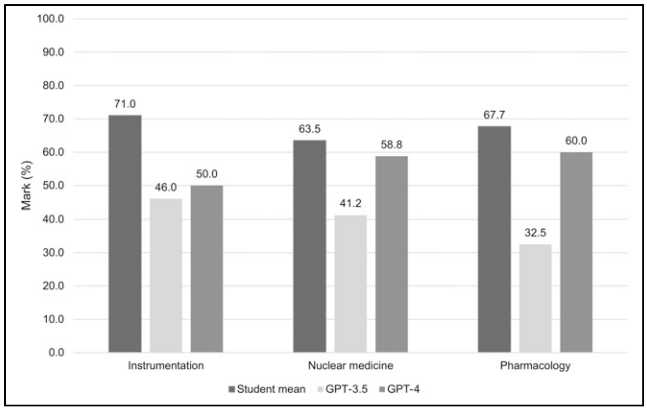

FIGURE 1. Bar chart for student mean, GPT-3.5 score, and GPT-4 score for each of 3 subjects who had written tasks evaluated.

Written Assignment

The 3 written assignment tasks were marked against the task rubric. In all 3 subjects, ChatGPT scored significantly more poorly than the mean student score, although GPT-4 was superior to GPT-3.5 (Table 1). Although a general trend suggested the gap between student mean and GPT-3.5 scores, the opposite trend was noted for GPT-4, which supports the purported improved reasoning of the tool (Fig. 1). There was a statistically significant difference (Wilcoxon signed rank test) between the student mean scores and both the GPT-3.5 scores (mean difference, −27.5%; $P = 0.0062$) and the GPT-4 scores (mean difference, −11.1%; $P = 0.0362$). Although mean scores were higher for GPT-4 than GPT-3.5 (mean difference, 16.4), this difference was not considered statistically significant at the 0.95 level ($P = 0.0534$). Turnitin Similarity Report-generated AI scores ranged from 67% to 100% for the written assignments (Table 1).

Calculation Examination

GPT-3.5 performed poorly in calculation-style questions, whereas GPT-4 showed improved capability (Table 1). For radiopharmacy, GPT-3.5 had a score of 24.0% compared with a student mean of 67.3% and a GPT-4 score of 66.1% (Fig. 2). The GPT-4 scores comprised 31.7% in short calculations and 8.7% in more complex problems for GPT-3.5, compared with 50.8% and 35.3%, respectively, for GPT-4. Both GPT-3.5 and GPT-4 had difficulties with decay calculations. For pharmacology, among the shorter calculation questions, GPT-3.5 scored 92.7%, compared with 100% for GPT-4, whereas more complex questions received scores of 0% and 17.5% for GPT-3.5 and GPT-4, respectively. Overall, in pharmacology GPT-3.5 had a score of 38.8%, compared with a student mean of 63.3% and a GPT-4 score of 52.0% (Fig. 2). There was a statistically significant difference (Wilcoxon signed rank test) between the student mean scores and GPT-3.5 scores (mean difference, −33.9%; $P = 0.0375$) but not GPT-4 scores (mean difference, −6.25%; $P = 0.1987$). Although mean scores were higher for GPT-4 than for GPT-3.5 (mean difference, 26.65), this was not considered statistically significant at the 0.95 level ($P = 0.1662$).

TABLE 1
Summary of Performance of GPT-3.5, GPT-4, and Student Cohorts

Category	Student mean score (%)	GPT-3.5 score (%)	GPT-4 score (%)	Students (*n*)
Written assignment marking				
Instrumentation	71.0	46.0	50.0	13
Nuclear medicine	63.5	41.2	58.8	11
Pharmacology	67.7	32.5	60.0	81
Written assignment AI scores (Turnitin)				
Instrumentation	—	100.0	73.0	13
Nuclear medicine	—	67.0	73.0	11
Pharmacology	—	74.0	68.0	81
Calculation examination marking				
Radiopharmacy	67.3	24.0	66.1	12
Pharmacology	63.3	38.8	52.0	81
Final examination marking				
Instrumentation	60.4	47.1	55.8	13
Radiopharmacy	60.0	50.5	52.9	12
Nuclear medicine	63.0	55.2	56	11
Pharmacology	47.3	66.2	79.8	81

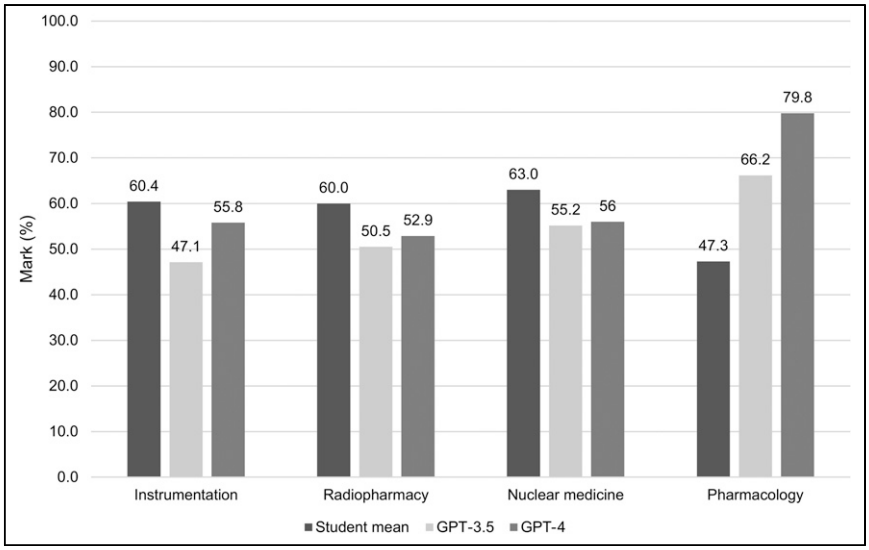

FIGURE 2. Bar chart for student mean, GPT-3.5 score, and GPT-4 score for each calculation examination.

Final Examination

For the radiopharmacy subject, the student mean was 60.0%, compared with 50.5% for GPT-3.5 and 52.9% for GPT-4 (Fig. 3). For the instrumentation subject, the student mean was 60.4%, compared with 47.1% for GPT-3.5 and 55.8% for GPT-4. For the clinical nuclear medicine subject, the student mean was 63.0%, compared with 55.2% for GPT-3.5 and 56.0% for GPT-4. For the pharmacology subject, the student mean was 47.3%, compared with 66.2% for GPT-3.5 and 79.8% for GPT-4. These observations reflect the nuclear medicine–specific nature of the content in the first 3 subjects combined with the level of student understanding of the content through reinforced learning across multiple theory and clinical placement (workplace learning) subjects. Conversely, pharmacology is foreign material for students but is a topic for which a shallow, general response by ChatGPT is suitable. There were no statistically significant differences (Wilcoxon signed rank test) between the student mean scores and GPT-3.5 scores (mean difference, −2.9%; $P = 0.7178$) or GPT-4 scores (mean difference, 3.45%; $P = 0.7456$). There was also no statistically significant difference between GPT-4 and GPT-3.5 (mean difference, 6.37; $P = 0.2500$).

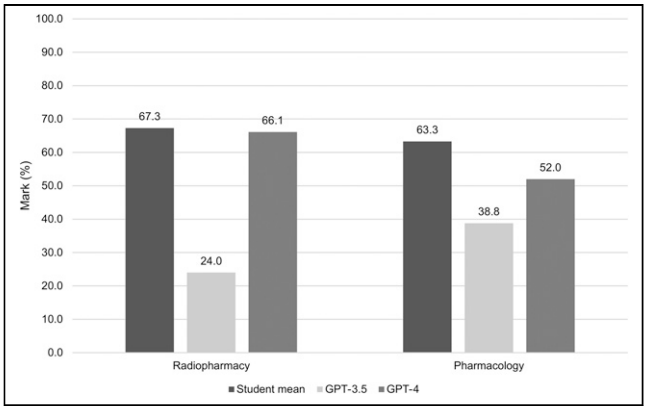

FIGURE 3. Bar chart for student mean, GPT-3.5 score, and GPT-4 score for each of 4 subjects who had final examinations evaluated.

DISCUSSION

Although GPT-4 is orders of magnitude superior to GPT-3.5 in terms of accuracy of responses and the professionalism of the language used, there remain limitations. GPT models generate text by predicting the next word in a sequence, which allows natural language and coherent sentences to be generated. This text generation method also leaves GPT-4 prone to the fabrications and hallucinations that plague GPT-3.5 (*2*,*3*,*5*). Furthermore, GPT-4 remains confounded by complex and discipline-specific questions that require deeper understanding—insight that can help shape formulation of examination questions. Despite this limitation, GPT-4 significantly outperforms GPT-3.5 and raises concern for potential misuse and academic integrity, particularly for topics of a superficial or general nature, for which GPT-4 also outperformed students (e.g., pharmacology).

ChatGPT+ is an innovative tool that, with GPT-4, has an improved capacity for assisting student development and increased potential for misuse. For the student performing at a pass level, ChatGPT powered by GPT-4 remains of limited benefit. There is a significant threat to academic integrity and misconduct associated with the fabrication of information and references and with errors and hallucinations. Written assessment tasks are particularly problematic for ChatGPT in this regard. ChatGPT powered by GPT-4 has improved capability and accuracy for examination responses, especially in multiple-choice questions and questions for which a shallow or general response is suitable.

CONCLUSION

ChatGPT is an innovative educational tool that, despite limited generative capability in nuclear medicine, offers promise for student support and development. There remains the risk of misuse. GPT-4 affords enhancements over GPT-3.5 that provide exciting opportunities for AI-augmented learning while demanding increased awareness and scrutiny for potential misuse.

DISCLOSURE

No potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

I acknowledge the contribution of ChatGPT (GPT-3.5 and GPT-4) in generating some of the text in this article. The model was accessed between May 22 and May 25, 2023.

KEY POINTS

QUESTION: Does ChatGPT powered by GPT-4 outperform and address previous problems associated with GPT-3.5?

PERTINENT FINDINGS: ChatGT powered by GPT-4 addresses previously documented limitations of GPT-3.5, particularly accuracy and professional tone, but remains susceptible to fabrication and hallucination.

IMPLICATIONS FOR PATIENT CARE: ChatGPT has increasing scope for cheating among university students with the enhancements associated with GPT-4.

REFERENCES

1. Bhayana R, Bleakney RR, Krishna S. GPT-4 in radiology: improvements in advanced reasoning. *Radiology.* 2023;307:e230987.
2. Currie G. Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy? *Semin Nucl Med.* 2023;53:719–730.
3. Currie G, Barry K. ChatGPT in nuclear medicine education. *J Nucl Med Technol.* 2023;51:247–254.
4. 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.
5. Currie G, Singh C, Nelson T, Nabasenja C, Al-Hayek Y, Spuur K. ChatGPT in medical imaging higher education. *Radiography.* 2023;29:792–799.
6. Graham F. Daily briefing: what scientists think of GPT-4, the new AI chatbot. *Nature.* March 17, 2023 [Epub ahead of print].
7. Sanderson K. GPT-4 is here: what scientists think. *Nature.* 2023;615:773.