

ChatGPT and Patient Information in Nuclear Medicine: GPT-3.5 Versus GPT-4

Geoff Currie¹, Stephanie Robbie², and Peter Tually^{1,3}

¹School of Dentistry and Medical Sciences, Charles Sturt University, Wagga Wagga, New South Wales, Australia; ²Queensland X-Ray, St. Andrews Hospital, Toowoomba, Queensland, Australia; and ³Telem Health, Kalgoorlie, Western Australia, Australia

The GPT-3.5–powered ChatGPT was released in late November 2022 powered by the generative pretrained transformer (GPT) version 3.5. It has emerged as a readily accessible source of patient information ahead of medical procedures. Although ChatGPT has purported benefits for supporting patient education and information, actual capability has not been evaluated. Moreover, the March 2023 emergence of paid subscription access to GPT-4 promises further enhanced capabilities requiring evaluation. **Methods:** ChatGPT was used to generate patient information sheets suitable for gaining informed consent for 7 common procedures in nuclear medicine. Responses were generated independently for both GPT-3.5 and GPT-4 architectures. Specific procedures were selected that had a long-standing history of use to avoid any bias associated with the September 2021 learning cutoff that constrains both GPT-3.5 and GPT-4 architectures. Each information sheet was independently evaluated by 3 expert assessors and ranked on the basis of accuracy, appropriateness, currency, and fitness for purpose. **Results:** ChatGPT powered by GPT-3.5 provided patient information that was appropriate in terms of being patient-facing but lacked accuracy and currency and omitted important information. GPT-3.5 produced patient information deemed not fit for the purpose. GPT-4 provided patient information enhanced across appropriateness, accuracy, and currency, despite some omission of information. GPT-4 produced patient information that was largely fit for the purpose. **Conclusion:** Although ChatGPT powered by GPT-3.5 is accessible and provides plausible patient information, inaccuracies and omissions present a risk to patients and informed consent. Conversely, GPT-4 is more accurate and fit for the purpose but, at the time of writing, was available only through a paid subscription.

Key Words: GPT-4; patient education; ChatGPT; generative AI; language model

J Nucl Med Technol 2023; 51:307–313

DOI: 10.2967/jnmt.123.266151

ChatGPT (OpenAI) is an artificial intelligence (AI), generative pretrained transformer (GPT) language model trained to generate humanlike text (1). It is a chatbot driven by GPT-3.5, first available publicly on November 30, 2022.

The ghostwriting capability of ChatGPT has significant potential in all aspects of medical writing (1), although there are few formal evaluations of accuracy, appropriateness, and currency for medical information. Indeed, in a dynamic industry such as nuclear medicine, in which changes are frequent, currency may be a limitation for ChatGPT, whose learning ended in September 2021. In earlier research, ChatGPT was also revealed to lack the depth of insight and professional language to be useful to nuclear medicine and medical imaging students (2–4). ChatGPT does not access information live or from memory. It was trained to learn textual relationships in a large language dataset up to and including September 2021. Sequences of words whose use may have emerged after this date, because it is new technology, will not be word sequences that ChatGPT can reliably predict. The key concept is that ChatGPT does not look for answers to prompt questions in memory or the Internet; it simply predicts the next word repeatedly until an answer forms.

At the time of writing, there was a paucity of scientific evaluations of ChatGPT medical writing for either health professionals or patients. Moreover, the research that has been undertaken for ChatGPT relates predominantly to the GPT-3.5 architecture. Evaluations of the accuracy, appropriateness, and currency of ChatGPT to date invariably issue a disclaimer that advances associated with the inevitable GPT-4 architecture will require new considerations (2–5). A paid subscription version of GPT-4 was released in March 2023, although some features were yet to be enabled (e.g., image inputs). The GPT-4 architecture is associated with improved language capabilities but remains constrained by the September 2021 learning cutoff. The problem solving and accuracy of GPT-4 are improved by 60% over GPT-3.5 (6,7).

When asked, ChatGPT confidently claims to be able to assist the nuclear medicine clinician in numerous ways (2,8), including by providing patient information. Given the previously outlined (3,4) limitations of ChatGPT for depth and professional language, ChatGPT might be better suited to generating patient information. Despite the confidence of ChatGPT, this capability requires formal evaluation. ChatGPT was asked whether it had such a capability, and although both GPT-3.5 and GPT-4 were optimistic, GPT-4 was more realistic about limitations and caveats. The actual answers by each version were as follows.

Received Jun. 8, 2023; revision accepted Jul. 13, 2023.

For correspondence or reprints, contact Geoff Currie (gcurrie@csu.edu.au).
Published online Sep. 12, 2023.

COPYRIGHT © 2023 by the Society of Nuclear Medicine and Molecular Imaging.

GPT-3.5

“As an AI language model, ChatGPT can be useful in developing patient information for nuclear medicine procedures as part of informed consent. With its vast knowledge base and natural language processing capabilities, ChatGPT can provide detailed and accurate information about nuclear medicine procedures, including what they involve, how they are performed, and what potential risks or side effects patients may experience.

“In addition to providing information, ChatGPT can also help patients better understand the information they are receiving by answering their questions in a clear and concise manner. This can help alleviate any concerns or fears patients may have about undergoing a nuclear medicine procedure and help them make informed decisions about their treatment.

“Of course, it is important to note that ChatGPT should not be used as a substitute for a qualified healthcare professional. Patients should always consult with their healthcare provider to receive personalized information and advice about their individual healthcare needs. However, ChatGPT can be a valuable tool in helping to educate and inform patients about nuclear medicine procedures as part of the informed consent process.”

GPT-4

“As an AI, GPT-4 can certainly assist in generating patient information sheets for the purpose of obtaining informed consent. These sheets could include details about the procedure or treatment, its benefits and risks, alternatives, and the patient’s rights.

“However, it’s important to note a few caveats:

“Limitation in Medical Knowledge. AI models like GPT-4 are trained on a diverse range of data from the internet up to a specific cutoff point, and they do not have the ability to learn or update their knowledge after this cutoff. This means they may not be aware of recent developments or research in the medical field.

“No Personalized Medical Insights. GPT-4 doesn’t have access to personal data about individuals unless it has been shared in the course of the conversation. It is designed to respect user privacy and confidentiality. Therefore, it can’t provide information based on a specific patient’s medical history or condition unless that context is provided within the conversation.

“Validation by Healthcare Professionals. Any information produced by GPT-4, especially related to health and medical treatments, should be reviewed and validated by healthcare professionals. While the model aims to provide accurate and helpful information, it’s not a substitute for professional medical advice, diagnosis, or treatment.

“In summary, while AI can assist in creating draft versions of patient information sheets, the final version should always be produced in collaboration with healthcare

professionals to ensure its accuracy, relevancy, and compliance with healthcare regulations.”

EVALUATING CHATGPT

Approach

To evaluate the capabilities of ChatGPT for generating patient information sheets suitable for use in gaining informed consent, ChatGPT was asked to generate several patient information sheets for common procedures: bone scanning, myocardial perfusion scanning, thyroid scanning, ventilation–perfusion lung scanning, ¹⁸F-FDG PET scanning, captopril renal scanning, and ⁸⁹Sr palliation.

In separate ChatGPT sessions, patient information sheets were generated using the GPT-3.5 and GPT-4 architectures. No prompt engineering was adopted, and default settings were used for temperature and tone. Three expert nuclear medicine technologists or scientists independently assessed each ChatGPT-generated information sheet for accuracy, appropriateness, currency, and fitness for the purpose, with rankings for each subcategory as poor, below average, average, above average, or excellent. Patient information sheets were also directly compared for content between the ChatGPT offering and information sheets used clinically for patients.

Results

There was general consensus among the 3 assessors for GPT-3.5, although there was a statistically significant difference in overall responses between each pair of assessors ($P < 0.001$). Assessor 1 had a higher proportion of poor results; assessor 2, a higher proportion of below-average results; and assessor 3, a higher proportion of average results (Fig. 1). Conversely, for GPT-4, there was no statistically significant difference among assessors 1–3 ($P = 0.752$), with each providing predominantly average evaluations. There was, however, a statistically lower proportion of above-average evaluations for assessor 2 than for either assessor 1 ($P < 0.001$) or assessor 3 ($P < 0.001$) (Fig. 1).

Across the 7 information sheets, accuracy of information was identified as an issue for GPT-3.5, whereas appropriateness of information was thought adequate (Table 1). Currency of information was generally below average but was particularly poor when considering omitted information. In their totality, the data were presented in an appropriate way but lacked accuracy, and there was important information omitted. As a result, the cumulative impression of GPT-3.5-generated patient information sheets was that they were inadequate (Table 1). The accuracy, appropriateness, and currency were all enhanced by GPT-4 (Table 1; Fig. 2); however, accuracy and fitness for the purpose remained below minimum standards (Fig. 3).

Among the 7 information sheets generated by GPT-3.5, the mode response was below average for all except the bone scan (which was average). The simplicity and commonality of the bone scan explains the ranking (Table 2).

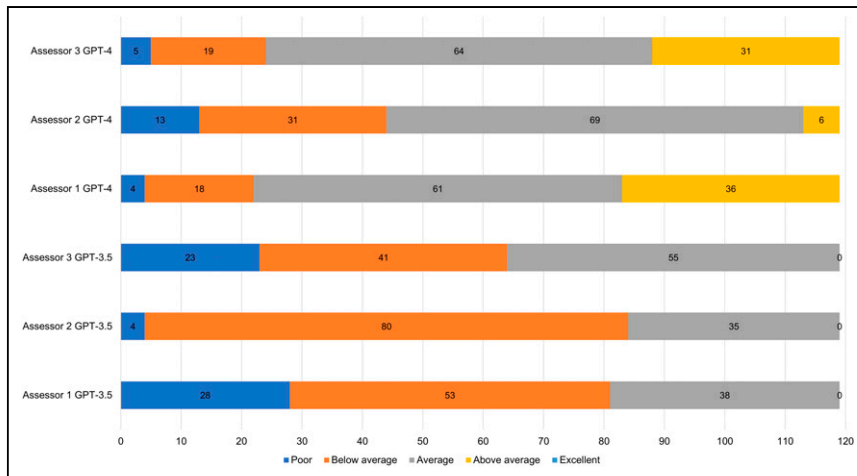


FIGURE 1. Mosaic plot of responses from 3 assessors for each GPT-3.5 and GPT-4 evaluation. Colored column widths are proportion of each evaluation category, and corresponding number is absolute number of questions classified for that category.

GPT-4 produced higher-quality patient information sheets (Table 2), with 3 classified as fit for the purpose: bone scan, ventilation–perfusion lung scan, and ⁸⁹Sr palliation.

Several clinically available patient information sheets were used to critique the information sheets produced by ChatGPT. The major flaw in all GPT-3.5–generated patient information sheets was the regurgitation of the same general, and often incorrect, information for each procedure. For example, “You may be asked to avoid eating or drinking for a few hours before the scan” appeared in every GPT-3.5–generated information sheet, including those that had no preparation and without specific detail when fasting was required. This falls short of adequate information and insight to constitute informed consent. GPT-3.5 tended to provide superficial and general information, at times adopting language too basic even for patients.

This included generalizations that could be misleading and errors, both of which threaten professionalism and the validity of informed consent. Conversely, GPT-4 used

TABLE 1
Results Across 7 ChatGPT Generated Information Sheets Among 3 Expert Assessors

Parameter	GPT-3.5				GPT-4			
	Poor	Below average	Average	Above average	Poor	Below average	Average	Above average
Overall accuracy	16.3%	72.8%	10.9%	0%	4.8%	20.4%	60.5%	14.3%
Procedure is accurately explained	11	9	1	0	3	1	15	2
Preparation is accurate	9	10	2	0	3	4	11	3
Postprocedure requirements are correct	1	16	4	0	0	6	12	3
Potential side effects or risks are correctly outlined	0	18	3	0	0	5	12	4
Content is relevant to procedure	1	17	3	0	0	5	11	5
Information is evidence-based	1	17	3	0	0	5	14	2
All information is accurate	1	20	0	0	1	4	14	2
Overall appropriateness	0%	11.4%	88.6%	0%	5.7%	4.8%	52.4%	37.1%
Medical terminology is appropriate and explained in layperson’s terms	0	0	21	0	0	1	12	8
Language and tone are appropriate for target patients and their families	0	0	21	0	0	1	12	8
Information is presented in clear, organized manner	0	6	15	0	0	3	7	11
Any cultural or linguistic considerations have been considered	0	6	15	0	6	0	14	1
Professional tone is used in patient-appropriate way	0	0	21	0	0	0	10	11
Overall currency	21.4%	57.1%	21.4%	0%	8.3%	26.2%	53.6%	13.1%
Content is up to date	1	19	1	0	0	4	13	4
Information reflects current best practice	4	17	0	0	2	6	10	3
Information is free from bias	0	4	17	0	0	1	17	3
There is no key information omitted	13	8	0	0	5	11	5	1
Is this adequate for purposes of informed consent?	12	7	1	0	3	11	4	3

Mode is highlighted in bold. There were no excellent responses.

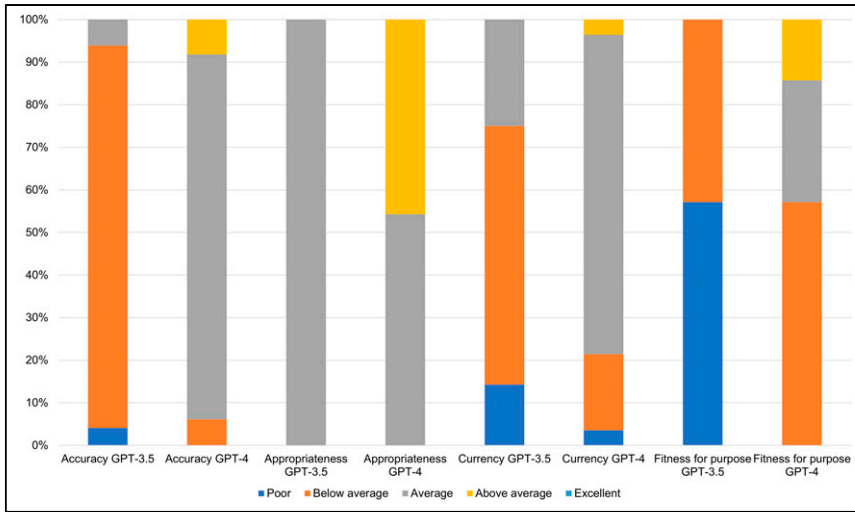


FIGURE 2. Percentage of responses in each category (accuracy, appropriateness, currency, and fitness for purpose) demonstrating improved outcomes for GPT-4 compared with GPT-3.5.

better language structure and provided more accurate information for each specific procedure. Nevertheless, GPT-4 tended to provide less information, thus creating an easy-to-read information sheet but one that omitted key information in parts.

A detailed analysis of the bone scan patient information sheet produced by GPT-3.5 and GPT-4 (Supplemental Appendix 1; supplemental materials are available at <http://jnmt.snmjournals.org>) provides a prototype for comparison of capabilities. Both GPT-3.5 and GPT-4 used a truncated list of applications for bone scanning; this truncation could

leave patients confused about whether bone scanning is the right test for their circumstances. Although GPT-3.5 omitted the important description of the procedure as “very sensitive,” GPT-4 made this point clearly to offset any perceived risk. Both GPT-3.5 and GPT-4 failed to include a section explaining what patients need to bring (e.g., previous scans, medication history, and health card), but only GPT-3.5 erroneously suggested that patients could be asked to fast or restrict fluids. It is important to indicate that a patient history will be taken before the injection, so the patient is aware of the opportunity to receive further information or ask questions. The GPT-3.5 information sheet would benefit from division into 2 parts to better explain the injection with or without initial imaging and then the delayed imaging, whereas GPT-4 includes this division but lacks adequate detail. The injection-to-scan time was incorrect for GPT-3.5, and neither architecture provided information about hydration or resumption of normal activities. Discussion of SPECT and SPECT/CT was omitted from both versions. GPT-3.5 described sounds typical of MRI, a description that is incorrect and misleading, whereas GPT-4 did not suffer this type of hallucination. Both GPT-3.5 and GPT-4, importantly, invited patients to ask further questions and to provide feedback.

Much of the GPT-3.5 bone scan patient information sheet was regurgitated verbatim for the other patient information sheets, which replicated the issues described above—most notably, the patient preparation and MRI-like sounds. Additionally, the following specific errors were noted.

The GPT-3.5 myocardial perfusion scan information sheet lacked specific advice about fasting, the need for caffeine cessation, and the need for comfortable clothes and shoes. The comfortable-clothes issue was addressed by GPT-4, but food and medication advice were only partially addressed. There was no information specific to reporting a history of diabetes or asthma in relation to stress testing. Only GPT-4 provided advice on checking with a medical practitioner before stopping medications. The information needed to be divided into rest and stress components for both



FIGURE 3. Radar plot of accuracy, appropriateness, currency, and fitness-for-purpose criteria for GPT-3.5, GPT-4, and minimum standard expected by assessors.

TABLE 2
Total Responses Across 7 ChatGPT-Generated Information Sheets Among 3 Expert Assessors

Information sheet	GPT-3.5				GPT-4			
	Poor	Below average	Average	Above average	Poor	Below average	Average	Above average
Bone scan	0	8	9	0	0	0	7	10
Myocardial perfusion	3	8	6	0	0	4	11	2
Thyroid scan	0	11	6	0	1	3	13	0
Ventilation–perfusion lung scan	0	11	6	0	0	0	16	1
¹⁸ F-FDG PET scan	2	9	6	0	0	1	12	4
Captopril renal scan	2	9	6	0	0	2	12	3
⁸⁹ Sr palliation	3	8	6	0	0	2	13	2
Total percentage of responses	8.4%	53.8%	37.8%	0%	0.8%	10.1%	70.6%	18.5%

Mode is highlighted in bold. There were no excellent responses.

versions and have the appropriate timings and detail outlined.

The GPT-3.5 thyroid scan information sheet overlooked mentioning iodine-based foods or supplements and implied that the patient was injected in an imaging position that translated to timing and protocol errors. The protocol and timing errors were rectified by GPT-4.

The ventilation–perfusion lung scan information sheet produced by GPT-3.5 repeated the previously described patient preparation and requirements errors and included incorrect procedure information. These were largely rectified by the GPT-4 version of the information sheet. These same observations were made for the ¹⁸F-FDG PET patient information sheets.

The captopril renal scan information sheet produced by both GPT-3.5 and GPT-4 mentioned “plenty of fluids” and “well-hydrated” without quantifying how much fluid needs to be consumed over what period. Neither information sheet provided an adequate breakdown of the 2-part protocol, and both overlooked postcaptopril monitoring. GPT-4 was generally more accurate for timings and the reason for performing the scan.

Given that ⁸⁹Sr palliation uses therapeutic doses of radio-nuclides, information sheets need to have sufficient accuracy and detail to allow informed consent. GPT-3.5 lacked the detail needed for a patient to understand the procedure and provide informed consent. GPT-4 produced more accurate information and, despite being concise, was fit for the purpose.

All GPT-3.5 information sheets that were produced contained errors related to procedure timing and protocol, included information errors and omissions, had insufficient detail for attaining informed consent, and lacked organization of concepts. GPT-4 generally addressed these issues with a more professional presentation of more accurate and

appropriate information, although some information lacked detail or was omitted.

DISCUSSION

Although GPT-3.5–generated patient information sheets provide plausible information for patients having nuclear medicine procedures, there were errors, misinformation, and omissions that rendered all GPT-3.5 versions inadequate. Use of such material in clinical practice could create confusion among patients and contradict the information provided by nuclear medicine professionals. Accessibility of the publicly and freely available GPT-3.5 increases potential reliance as a source of information by patients. How the convenience and accuracy of GPT-3.5 compares with Internet browser search strategies requires investigation. The shortcomings of GPT-3.5 are likely to increase the time demands on nuclear medicine staff for providing clarification and ameliorating any anxiety produced by discrepancies. These observations are counter to the purported benefits of AI generally, and ChatGPT specifically, in supporting patients and clinicians.

Of particular concern are the errors in the information provided by GPT-3.5. The patient information sheets generated by GPT-3.5 included a variety of inaccuracies that can be classified using previously defined terms (2,3): hallucination (e.g., “you may be asked to avoid eating or drinking for a few hours before the scan” for a bone scan); illusion (e.g., suggesting that MRI-type sounds might be experienced during standard nuclear medicine imaging); delusion (e.g., inaccurate waiting times for the thyroid scan); delirium (e.g., omission of crucial information about caffeine cessation for stress myocardial perfusion scans); confabulation (e.g., “you should drink plenty of fluids to help flush the radioactive material from your body” for a biliary system–excreted radiopharmaceutical); and extrapolation (e.g.,

advice about injection-site discomfort that is more typical of CT contrast administration).

Patient information sheets prompt questions about procedures to help patients provide informed consent. The information age has seen increased reliance by patients on the Internet, but any given search can reap inaccurate, confusing, and unrelated information mixed among the websites providing valuable information. ChatGPT (public version) provides patients with immediate answers that appear plausible and professional without dredging through multiple websites while deciphering quality from mediocrity (or worse). The value of GPT-3.5 in the patient information arena might be better targeted at translating existing patient information when English is a second language, although ChatGPT has some reported limitations in non-English language text generation (9).

The emergence of GPT-4 through paid subscription is less accessible to patients but generates patient information sheets of higher quality—insufficient, perhaps, to be used directly as an information sheet distributed by nuclear medicine departments but sufficient to provide a foundation for patients sourcing information before their procedure. Indeed, the quality is sufficient to develop a bank of questions patients want to ask ahead of their procedure, which would support enhanced informed consent, and is certainly sufficiently accurate and consistent to displace the more tortuous general Internet search strategies.

GPT-4 was able to provide plausible information for patients having nuclear medicine procedures, with few errors and constrained only by information omission. The concerning errors in the information that plagued GPT-3.5 are largely overcome with more accurate and professional offerings by GPT-4. Indeed, the language structure itself eliminated hallucination, delirium, and confabulation. The advanced architecture of GPT-4, despite having the same learning cutoff dates as GPT-3.5, improved accuracy including illusion, delusion, and extrapolation. Given that ChatGPT is a language model and predicts text on the basis of adjoining text, GPT-4 does not know more information but rather is better able to predict the appropriate language. For example, ChatGPT does not scour the Internet to find how long a bone scan takes; it starts a sentence structure (e.g., “a bone scan takes”) and predicts what text is most likely to appear next, as with predictive text messaging. GPT-4 has no more information to rely on than GPT-3.5 in making that prediction, but the more advanced architecture has a more accurate prediction. In this specific question, triplicate enquiries of GPT-3.5 each responded with 1–2 h as the length of a bone scan, with variable other information unrelated to time. Conversely, GPT-4 first responded with 3–6 h with a scan time of 30–60 min and a delay between injection and scan of 2–4 h, then responded with several hours with a scan time of 30 min and a delay between injection and scan of 2–4 h, and finally responded with 3–5 h

with a 10- to 15-min injection time, 2- to 4-h delay between injection and scan, and 30–60 min of actual scan time.

The steps are simple, although the algorithmic function is complex (10). First, the input text is broken down into units called tokens. Second, each token is mapped against semantic and syntactic properties learned in the training phase, which was truncated at September 2021. Third, the transformer model (GPT-3.5 or the more powerful GPT-4) determines the context of each word and creates a new token or representation that includes context. Fourth, the GPT produces a probability distribution of potential next words based on the preceding token context. Fifth, a word is chosen on the basis of the highest probability or randomly to introduce variability in responses. This process occurs in approximately real time for every word generated by ChatGPT. Although learning cutoff dates are the same, it is easy to understand how errors are produced in steps 3–5 and how the more powerful architecture of GPT-4 reduces those errors.

Despite the enhanced performance of GPT-4, the inaccuracies and omissions demand direct supply of accurate and appropriate information to each patient from the nuclear medicine department at the time the patient makes the appointment. This information should also be distributed directly to patients by referring specialists and should be developed in patient-facing language. As a consequence, patients will be less likely to lean on the variable reliability of ChatGPT.

CONCLUSION

ChatGPT powered by GPT-3.5 has a limited role in helping nuclear medicine departments produce patient information or in supporting patients seeking information ahead of nuclear medicine procedures. GPT-3.5 is inaccurate and omits key information, undermining informed consent. GPT-4 is less accessible at the time of writing (paid subscription) but provides more accurate information that, although generally inadequate for informed consent alone, is valuable and provides fodder for patient queries ahead of procedures. The potential for patients to use ChatGPT to source information ahead of nuclear medicine procedures presents a risk of misinformation, confusion, and an increased demand for nuclear medicine staff time, while concurrently offering an exciting opportunity to enrich patient experiences and empower the informed consent process.

DISCLOSURE

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

ACKNOWLEDGMENT

We acknowledge the contribution of ChatGPT (versions 3.5 and 4) in generating some of the text in this article. The model was accessed between April 13 and June 6, 2023.

KEY POINTS

QUESTION: Can ChatGPT generate accurate patient information sheets?

PERTINENT FINDINGS: ChatGPT powered by GPT-3.5 lacks the capability to provide responses that reflect the depth, accuracy, and currency of information for patients having nuclear medicine procedures. GPT-4 provides enhanced capability with more accurate and appropriate information sheets that may be useful for informed consent.

IMPLICATIONS FOR PATIENT CARE: GPT-3.5 has limited scope to support patient information and education but may emerge as a risk because of its accessibility and purported patient benefits. The improved capabilities of GPT-4 appear set to change that landscape.

REFERENCES

1. Biswas S. ChatGPT and the future of medical writing. *Radiology*. 2023;307:e223312.
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. Currie G, Singh C, Nelson T, Nabasenja C, Al-Hayek Y, Spuur K. ChatGPT in medical imaging higher education. *Radiography*. 2023;29:792–799.
5. 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.
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.
8. Currie G. A conversation with ChatGPT. *J Nucl Med Technol*. 2023;51:255–260.
9. Seghier ML. ChatGPT: not all languages are equal. *Nature*. 2023;615:216.
10. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv website. <https://arxiv.org/abs/1706.03762>. Published June 12, 2017. Revised August 2, 2023. Accessed August 9, 2023.