



# Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be?

Ian L. Alberts<sup>1</sup> · Lorenzo Mercolli<sup>1</sup> · Thomas Pyka<sup>1</sup> · George Prenosil<sup>1</sup> · Kuangyu Shi<sup>1</sup> · Axel Rominger<sup>1</sup> · Ali Afshar-Oromieh<sup>1</sup>

Received: 9 February 2023 / Accepted: 19 February 2023 / Published online: 9 March 2023  
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

There has been substantial press recently regarding the impressive performance of large language models (LLM), particularly the OpenAI tool “Chat Generative Pre-Trained Transformer,” commonly known as “ChatGPT” [1]. LLM represent artificial intelligence (AI) tools based on multi-layer recurrent neural networks that are trained on vast amounts of data to generate human-like text (<https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html>). Whereas traditional language models are programmed to use statistical techniques to predict the next word in a sentence, ChatGPT uses transformer-based models that allow for the processing of vast amounts of data in parallel. The result is a revolution in the ability of these models to understand and generate text. Their performance is remarkable, e.g., ChatGPT is capable of writing lines of code, producing plays, stories, poetry as well as simulated scientific content such as abstracts. While there has been much fanfare in the media regarding this undoubtedly impressive performance, there is much less information available about how this might affect the nuclear medicine community, or its reliability in producing nuclear medicine and molecular imaging-related content. It is currently unclear to what extent ChatGPT might help as a collaborative tool, for example correcting or helping to improve upon a researcher’s writing or as a tool for summarising nuclear medicine literature.

Within seconds, ChatGPT is capable of producing convincing and grammatically fluent prose that is indistinguishable from content produced by human researchers.

The threat that this poses to academic publishing models is already apparent [2]. Controversially, ChatGPT has (at the time of writing) already been listed as a co-author on four academic publications [3]. Anecdotally, students are already using the tool as a writing assistant, raising issues of academic integrity and plagiarism [4]. There are already 25 PubMed entries for “ChatGPT”, this will likely grow rapidly in the coming weeks and months.

In response, a number of journals are already implementing editorial policies about the acceptability of AI-assisted writing or clarifying issues around authorship [3, 5]. Some internet fora have already banned ChatGPT-generated answers owing to their unreliability (<https://meta.stackoverflow.com/questions/421831/temporary-policy-chat-gpt-is-banned>).

Recent experience has shown that AI tools can be harnessed to mass-produce questionable content on social media networks or social media bots that can deliberately amplify misinformation [6]. This experience might portend the future of ChatGPT-generated academic content. A report from the Copenhagen-based Institute for Future Studies estimates that 99% of the internet could be produced by generative AI by 2025 (<https://cifs.dk/news/what-if-99-of-the-metaverse-is-made-by-ai>). At present, ChatGPT is not capable of producing an entire research paper *sua sponte*, although it is predicted and indeed conceivable that this might soon be the case [7]. Nevertheless, it can already, even in the currently available beta version, produce a very convincing abstract [8]. We wonder whether conferences might soon be flooded with AI-generated abstracts or whether predatory publishers [9] might be catalysed by the ability of ChatGPT to churn out convincing but ultimately unreliable content. Even the review process could be influenced: there are already proposals to harness the ability of LLM to summarise text as a tool for the sifting out of low-quality studies submitted to a journal. Once can imagine a not-too distant future where AI might generate and review research [10],

---

This article is part of the Topical Collection on Advanced Image Analyses (Radiomics and Artificial Intelligence)

✉ Ian L. Alberts  
ian.alberts@insel.ch

<sup>1</sup> Department of Nuclear Medicine, Inselspital, Bern University Hospital, University of Bern, Freiburgstr. 18, 3010 Bern, Switzerland

which could then be cited by other AI-generated research or commented upon via an AI-generated letter to the editor. Until recently, such a future might have sounded far-fetched. In light of the astonishing pace with which LLM have been implemented, we feel that the academic nuclear medicine community urgently needs to confront this issue.

Trustworthiness is a key concept in the academic AI literature, and reports of misrepresentation of simple facts or hidden bias means that there is much work to be done in this regard [11]. As a group of clinicians and scientists working in a clinical nuclear medicine department, we wonder can the nuclear medicine content presently produced by ChatGPT be trusted? A particularly important milestone in the training of any physician is the summative board or licensing examination. Such examinations have the purpose of protecting the public by holding physicians accountable and testing them according to a defined body of knowledge and maintaining public confidence through the maintenance of professional standards [12]. In tort law, when defining what is negligent practice, the knowledge and practice of a body of professionals is taken to be the reference standard in common law jurisdictions [13]. If AI tools are to be potential tools to assist, or even replace physicians, then their performance might be held to a similar standard. Indeed, this has been done recently yielding mixed results. For example, Shelmerdine et al. tested whether a commercially available AI tool (which interestingly also had a CE-conformity label) was capable of passing the radiograph reporting section of the United Kingdom Fellowship of the Royal College of Radiologists (FRCR). The tool performed poorly. Only with special dispensation was it capable of passing only two of ten mock examinations and it ranked last alongside its 26 human peers [14]. In contrast, ChatGPT was capable of passing or being close to passing all three parts of the United States Medical Licensing Exam (USMLE) without additional training or prompts [1] and a Chinese AI tool Xiaoyi (meaning “little doctor”) was, with training, capable of passing the Chinese Medical Licensing Exam [15]. Readers of this letter will doubtless be familiar with the hyperbole surrounding the role of AI in diagnostic radiology and nuclear medicine, exemplified by Hinton’s famous and now controversial statement in 2013 that “we stop training radiologists now” [16]. The effect of this hype, along with many other similar statements, has given cause for medical students to re-evaluate their career options [17], potentially at the cost of recruitment in a field where there is an increasing demand for services and shortage of trainees. We were therefore intrigued to see how ChatGPT might fare in a nuclear medicine examination, to assess whether it might pose a risk to the integrity of online nuclear medicine examinations and whether it has a potential role in assisting candidates preparing for nuclear medicine examinations. There is also the potential that patients might turn to these tools to help

answer their questions or allay their fears regarding nuclear medicine treatments and investigations.

The Fellowship of the European Board Examination is a two-stage examination administered by the Nuclear Medicine Section of the European Union of Medical Specialists in close cooperation with the European Association of Nuclear Medicine [18, 19]. The first step is a written examination, which since the SARS-CoV-2 pandemic has been delivered remotely and online [20]. The second is an oral examination that is delivered alongside the EANM annual congress. The online nature of the written exam is particularly prescient to ChatGPT, since conceivably, a candidate could use OpenAI tools dishonestly to help pass an online examination (<https://arxiv.org/abs/2212.09292>). Alternatively, LLM hold promise as tools for teaching and learning. They could, for example, assist a medical trainee by providing tailored guides to the literature [21].

To simulate the written part examination, the first author (and fellow of the EBNM) provided ChatGPT with 50 example multiple-choice questions which are openly available online as part of the training material for candidates (<https://link.springer.com/article/10.1007/s00259-011-1949-z>). They can be taken to be indicative of the breadth and depth of knowledge expected of candidates. The questions did not require image interpretation. ChatGPT states that its training cut-off was 2021—the questions presented to it date from 2009, meaning that ChatGPT was not expected to answer questions based on state-of-the-art knowledge on which it had not been trained.

The multiple-choice questions require the candidate to choose the single correct answer from four or five possibilities. In all 50 cases, ChatGPT provided a definitive answer. Marking these against the model answer provided in the training material revealed that ChatGPT was correct only 34% of the time (17/50). With 11 answers requiring the candidate to choose from five possible responses and the remainder having four possible responses, the mean probability of choosing the correct by random chance was 0.24, suggesting that ChatGPT was likely able to draw on some knowledge rather than simply guessing. With this performance, it is fair to say that ChatGPT would be unlikely to pass the exam if it were to take it in real life, although this could change in the future with better training for the model.

We then sought to test ChatGPT’s ability to learn or be corrected. The correct responses were provided via ChatGPT’s feedback tool. The incorrect question was asked again in an open and Socratic fashion, in the way a kindly examiner might give a wayward candidate a second chance to demonstrate his or her knowledge in an oral examination. For example, when first asked, “which benign lesion does not show increased uptake on a bone scan?” ChatGPT incorrectly chose osteoid osteoma, which is entirely wrong; this entity is famous for its exquisite avidity for

bone-seeking radiopharmaceuticals. When corrected and asked the question again but in a Socratic fashion “Name me a benign lesion which does not show increased uptake on a bone scintigraphy” ChatGPT responded as follows: “A benign lesion that typically does not show increased uptake on a bone scintigraphy is osteoarthritis. Osteoarthritis is a degenerative joint disease that does not involve abnormal bone metabolism and therefore does not result in increased uptake of the radiotracer on a bone scan”. This is of course nonsense: osteoarthritis does involve increased bone turnover and is a common incidental finding on bone scanning. Moreover, ChatGPT showed evidence of confabulation — osteoarthritis was not part of the question stem asked previously (the options were the correct choice of the bone cyst, or the incorrect options of Engelmann’s disease, Paget’s disease, or fibrous dysplasia). It is also non-reproducible, context sensitive and non-linear; asking the same question can reveal different answers. Medical regulators demand that licensed physicians recognise and practice within the limits of their own knowledge and competence (<https://www.gmc-uk.org/ethical-guidance/ethical-guidance-for-doctors/good-medical-practice/duties-of-a-doctor>). Rather than stating that it did not know, or could not answer, ChatGPT ventured to provide a superficial and possibly convincing answer which was wrong, and if relied upon to help interpret a scan finding, might be harmful. Although a very preliminary analysis of ChatGPT’s performance, it bears some similarity to adversarial examples. AI models can be fooled to produce incorrect output with a high degree of confidence [22, 23] and act as a test of the model’s robustness that goes beyond standard metric such as out-of-sample error [24]. In the case of LLM, however, constructing adversarial examples is notoriously difficult and this small analysis was able to probe the limits of ChatGPT’s performance in generating nuclear medicine-related content [25–27]. As the technology matures, future studies might be performed to systematically test the robustness of LLM and their reliability in producing nuclear medicine related content.

We must highlight that the producers of ChatGPT would not claim that it is, or should, be capable of passing a board examination. Nor can it be expected that it be capable of insight or self-awareness. We asked ChatGPT: “Are you capable of passing a nuclear medicine board examination?” It gave us the following response, which we quote verbatim: “No, as a language model AI, I do not have the capability to take exams or have the physical ability to perform medical procedures. I am trained to assist with answering questions and generating human-like text based on the input provided to me”. We then asked ChatGPT its thoughts on the reliability of its answers, eliciting the following response: “My answers are generated based on patterns in large amounts of text data and while they can be helpful as a reference, they are not a substitute for professional medical advice and should

not be used as the sole basis for making medical decisions. It’s always best to consult with a licensed medical professional for accurate and up-to-date information. Additionally, the information I have is based on the data I was trained on, which has a knowledge cut-off date of 2021, so newer developments or updates in the medical field may not be reflected in my answers.” At face value, this modest answer acknowledges the limits of an LLM. However, in our view, it does not adequately address the fact that the answers it provided were not just unhelpful, but factually incorrect, misleading and delivered without hesitation. We would therefore urge caution when weighing claims that LLMs might be used for summarising medical records, drafting authorisation letters to insurers justifying treatment costs or in decision support tools for diagnosis [28].

The rapid pace of events cannot be overemphasised: what is written about the performance of ChatGPT today could be invalidated within a matter of weeks, if not days. Indeed, in the days preceding the writing of this editorial, Microsoft announced that it would integrate ChatGPT into its Bing search engine and ChatGPT functionality has already been embedded in the beta version of the recently introduced you.com search engine. LambdaBard, dubbed a competitor to ChatGPT and produced by Google, is eagerly anticipated within weeks (<https://blog.google/technology/ai/lambda/>). ChatGPT itself has only been available in stable release form for a matter of days at the time of writing this editorial, and a subscription ChatGPT plus with additional functionality is planned (<https://openai.com/blog/chatgpt-plus/>).

In summary, although ChatGPT is presently capable of providing seemingly convincing content, including referenced abstracts that are capable of fooling peer reviewers [8], our preliminary analysis suggests that it is currently far from demonstrating the knowledge expected of a certified nuclear medicine physician in Europe in the setting of a standardised exam. Candidates preparing for exams or practicing physicians should test the validity of any statement generated by these models for themselves and be aware that the content can be unreliable. Given the performance demonstrated in this preliminary analysis, we do not see any evidence that ChatGPT would pose a threat to the integrity of any online nuclear medicine examination at the present time, although given the rapid pace of development, this could very well change in the near future. Nevertheless, we believe that the power of ChatGPT (or the lack thereof) shows an urgent need to address the ethical challenges of such systems in a systematic way [29–31]. We believe that the education and training of clinicians will have to adapt according to the degree of agency that tools like ChatGPT will have in the medical field. Finally, with tongue in cheek and contrary to Hinton’s advice, it would be prudent to continue training nuclear medicine physicians and radiologists—at least for the time being.

## Declarations

**Ethics approval** This article does not contain any studies with human participants or animals performed by any of the authors.

**Conflict of interest** AR has received research support and speaker honoraria from Siemens. All other authors report no conflict of interest.

## References

- Looi MK. Sixty seconds on ... ChatGPT. *BMJ*. 2023;380:205. <https://doi.org/10.1136/bmj.p205>.
- Mann Douglas L. Artificial intelligence discusses the role of artificial intelligence in translational medicine. *JACC: Basic Transl Sci*. n.d.;0. <https://doi.org/10.1016/j.jacbs.2023.01.001>.
- Thorp HH. ChatGPT is fun, but not an author. *Science*. 2023;379:313. <https://doi.org/10.1126/science.adg7879>.
- Stokel-Walker C. AI bot ChatGPT writes smart essays - should professors worry? *Nature*. 2022. <https://doi.org/10.1038/d41586-022-04397-7>.
- Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. *Nature*. 2023;613:620–1. <https://doi.org/10.1038/d41586-023-00107-z>.
- Gisondi MA, Barber R, Faust JS, Raja A, Strehlow MC, Westafer LM, et al. A deadly infodemic: social media and the power of COVID-19 misinformation. *J Med Internet Res*. 2022;24:e35552. <https://doi.org/10.10196/35552>.
- van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature*. 2023;614:224–6. <https://doi.org/10.1038/d41586-023-00288-7>.
- Else H. Abstracts written by ChatGPT fool scientists. *Nature*. 2023;613:423. <https://doi.org/10.1038/d41586-023-00056-7>.
- Beall J. Predatory publishers are corrupting open access. *Nature*. 2012;489:179. <https://doi.org/10.1038/489179a>.
- Checco A, Bracciale L, Loreti P, Pinfield S, Bianchi G. AI-assisted peer review. *Humanit Soc Sci Commun*. 2021;8:25. <https://doi.org/10.1057/s41599-020-00703-8>.
- Kitamura FC, Marques O. Trustworthiness of artificial intelligence models in radiology and the role of explainability. *J Am Coll Radiol*. 2021;18:1160–2. <https://doi.org/10.1016/j.jacr.2021.02.008>.
- Mirzaei S, Hustinx R, Prior JO, Ozcan Z, Boubaker A, Farsad M, et al. Improving nuclear medicine practice with UEMS/EBNM committees. *J Nucl Med: Off Publ Soc Nucl Med*. 2020;61:18N–20N.
- Sokol DK. How good a doctor do you need to be? *BMJ Br Med J*. 2012;345:e7858. <https://doi.org/10.1136/bmj.e7858>.
- Shelmerdine SC, Martin H, Shirodkar K, Shamshuddin S, Weir-McCall JR, Collaborators F-AS. Can artificial intelligence pass the Fellowship of the Royal College of Radiologists examination? Multi-reader diagnostic accuracy study. *BMJ*. 2022;379:e072826. <https://doi.org/10.1136/bmj-2022-072826>.
- Rampton V, Ko A. Robots, radiologists, and results. *BMJ*. 2022;379:o2853. <https://doi.org/10.1136/bmj.o2853>.
- Langlotz CP. Will Artificial intelligence replace radiologists? *Radiol Artif Intell*. 2019;1:e190058. <https://doi.org/10.1148/ryai.2019190058>.
- Hustinx R. Physician centred imaging interpretation is dying out - why should I be a nuclear medicine physician? *Eur J Nucl Med Mol Imaging*. 2019;46:2708–14. <https://doi.org/10.1007/s00259-019-04371-y>.
- Pons F, Bischof DA. The European board of nuclear medicine fellowship examination. *Eur J Nucl Med Mol Imaging*. 2006;33:109–10. <https://doi.org/10.1007/s00259-005-1963-0>.
- Neilly B, Dizdarevic S, Prvulovich L, Buscombe J, Lewington V. Nuclear medicine training and practice in the UK. *Eur J Nucl Med Mol Imaging*. 2016;43:800–3. <https://doi.org/10.1007/s00259-015-3255-7>.
- Ozcan Z, Kulakiene I, Vaz SC, Garzon JRG, Boubaker A. Challenges and possibilities for board exams in the Covid-19 era: experience from the Fellowship Committee of European Board of Nuclear Medicine. *Eur J Nucl Med Mol Imaging*. 2022;49:1442–6. <https://doi.org/10.1007/s00259-022-05716-w>.
- Baidoo-Anu D, Owusu Ansah L. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. 2023. Available at SSRN: <https://ssrn.com/abstract=4337484>; <https://doi.org/10.2139/ssrn.4337484>.
- Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. *arXiv preprint arXiv:13126199*. 2013. <https://doi.org/10.48550/arXiv.1312.6199>.
- Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:14126572*. 2014. <https://doi.org/10.48550/arXiv.1412.6572>.
- Carlini N, Athalye A, Papernot N, Brendel W, Rauber J, Tsipras D, et al. On evaluating adversarial robustness. *arXiv preprint arXiv:190206705*. 2019. <https://doi.org/10.48550/arXiv.1902.06705>.
- Zhuo TY, Huang Y, Chen C, Xing Z. Exploring AI Ethics of ChatGPT: a diagnostic analysis. *arXiv preprint arXiv:230112867*. 2023. <https://doi.org/10.48550/arXiv.2301.12867>.
- Alzantot M, Sharma Y, Elgohary A, Ho B-J, Srivastava M, Chang K-W. Generating natural language adversarial examples. *arXiv preprint arXiv:180407998*. 2018. <https://doi.org/10.48550/arXiv.1804.07998>.
- Zhang H, Zhou H, Miao N, Li L. Generating fluent adversarial examples for natural languages. *arXiv preprint arXiv:200706174*. 2020. <https://doi.org/10.48550/arXiv.2007.06174>.
- Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. *Radiology*. 2023;230163. <https://doi.org/10.1148/radiol.230163>.
- Milano S, Taddeo M, Floridi L. Recommender systems and their ethical challenges. *AI Soc*. 2020;35:957–67.
- Nyholm S. Attributing agency to automated systems: reflections on human–robot collaborations and responsibility-loci. *Sci Eng Ethics*. 2018;24:1201–19.
- Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med*. 2020;3:17.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.