BMJ Neurology Open

Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK Neurology Specialty Certificate Examination

Panagiotis Giannos [1,2,3]

To cite: Giannos P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK Neurology Specialty Certificate Examination. *BMJ Neurology Open* 2023;**5**:e000451. doi:10.1136/bmjno-2023-000451

Accepted 06 June 2023

[1]Department of Life Sciences, Imperial College London, London, UK
[2]Society of Meta-Research and Biomedical Innovation, London, UK
[3]Promotion of Emerging and Evaluative Research Society, London, UK

Correspondence to
Panagiotis Giannos; panagiotis.giannos19@imperial.ac.uk

ABSTRACT

Background Large language models such as ChatGPT have demonstrated potential as innovative tools for medical education and practice, with studies showing their ability to perform at or near the passing threshold in general medical examinations and standardised admission tests. However, no studies have assessed their performance in the UK medical education context, particularly at a specialty level, and specifically in the field of neurology and neuroscience.

Methods We evaluated the performance of ChatGPT in higher specialty training for neurology and neuroscience using 69 questions from the Pool—Specialty Certificate Examination (SCE) Neurology Web Questions bank. The dataset primarily focused on neurology (80%). The questions spanned subtopics such as symptoms and signs, diagnosis, interpretation and management with some questions addressing specific patient populations. The performance of ChatGPT 3.5 Legacy, ChatGPT 3.5 Default and ChatGPT-4 models was evaluated and compared.

Results ChatGPT 3.5 Legacy and ChatGPT 3.5 Default displayed overall accuracies of 42% and 57%, respectively, falling short of the passing threshold of 58% for the 2022 SCE neurology examination. ChatGPT-4, on the other hand, achieved the highest accuracy of 64%, surpassing the passing threshold and outperforming its predecessors across disciplines and subtopics.

Conclusions The advancements in ChatGPT-4's performance compared with its predecessors demonstrate the potential for artificial intelligence (AI) models in specialised medical education and practice. However, our findings also highlight the need for ongoing development and collaboration between AI developers and medical experts to ensure the models' relevance and reliability in the rapidly evolving field of medicine.

INTRODUCTION

The rapid advancements in artificial intelligence (AI) have led to the development of sophisticated language models, such as ChatGPT by OpenAI, which have attracted significant attention in various industries, including education, healthcare and entertainment.[1–3] These models employ large amounts of data and advanced computing techniques to generate meaningful responses based on human prompts. In the medical domain, AI-driven language models have shown potential in assisting with medical education, clinical decision-making and even medical writing.[4–6]

Recent literature has reported on the varied performance of ChatGPT across different subject domains, with some studies showing outstanding performance in economics and programming, while others reported unsatisfactory results in mathematics.[7] A rapid review highlighted the potential benefits of ChatGPT as an assistant for instructors and a virtual tutor for students, but also raised concerns about its generation of incorrect or fake information and the threat it poses to academic integrity.[7] Despite this growing body of research, the performance of AI models such as ChatGPT in specific fields, such as neurology and neuroscience, particularly at a specialty level, remains unexplored.

Large language models, such as ChatGPT, represent a new generation of models that combine clinical knowledge and dialogic interaction more effectively. They have been explored for personalised patient interaction and consumer health education, but their success in testing clinical knowledge through generative question-answering tasks has been limited. Most of the current literature has focused on a recent study by Kung *et al*, in which ChatGPT performed at or near the passing threshold for all three US Medical Licensing Exams (USMLE), suggesting its potential as an innovative tool for medical education.[8] In parallel, ChatGPT has also been evaluated on medical standardised admission tests in the UK, such as the BioMedical Admissions Test, showing promise in areas that assess aptitude, problem-solving, critical thinking and reading comprehension, while facing challenges in specialised domains such as scientific and mathematical knowledge and applications.[9] Yet, no studies have been conducted on ChatGPT's performance in the UK medical education context, specifically the UK Specialty Certificate Examination (SCE) in neurology.

To be useful, ChatGPT must perform comparably to humans on assessments of medical knowledge and reasoning. The UK SCE in neurology consittues a critical milestone for trainees in higher specialty training to assess their knowledge and skills in the field. In this brief report, we examine the performance of ChatGPT on sample questions from the SCE in neurology, addressing the gap in the literature and providing insights into the potential applications and challenges of using AI-driven language models in the field of neurology and neuroscience.

METHODS

We evaluated the performance of ChatGPT in the field of neurology and neuroscience using a sample of 69 questions from the Pool—SCE Neurology Web Questions bank, which encompasses a wide range of common and important disorders, as outlined in the curriculum syllabus.

The questions are structured in a 'best of five' format, testing not only knowledge but also intuitive clinical thinking. Each question presents a brief clinical scenario followed by the lead-in question and five possible answers, with one being the most correct among them.

For this study, we compared the performance of the legacy and default GPT-3.5 models of ChatGPT with the latest ChatGPT-4 model. To maintain consistency in our evaluation, we exclusively used multiple-choice questions, formatted for proper structure and readability (5 questions, 7%). We recorded the total number of questions attempted by each ChatGPT model and the number of correct answers provided by the models during the evaluation process. Additionally, we estimated each model's grade and candidate ranking based on their performance, using data from students who previously took the SCE in neurology.

RESULTS

Our sample dataset comprised 69 questions from the Pool—SCE Neurology Web Questions bank, with a primary focus on the field of Neurology, representing 55 questions (80%). Besides neurology, the dataset featured questions from various disciplines, including clinical science (2 questions, 3%), endocrinology and metabolic medicine (1 question, 1%), gastroenterology (1 question, 1%), ophthalmology (2 questions, 3%), psychiatry (4 questions, 6%) and therapeutics and toxicology (4 questions, 6%). The most frequent subcategories across multiple topics were symptoms and signs (68%), diagnosis (46%), interpretation (17%) and management (16%). Questions pertaining to specific patient populations were also present but appeared less frequently.

The performance of ChatGPT 3.5 Legacy, ChatGPT 3.5 Default and ChatGPT-4 on the dataset exhibited varying levels of accuracy (figure 1). ChatGPT 3.5 Legacy displayed an overall accuracy of 42%, performing better in endocrinology and neurology while showing relative weakness in clinical science and gastroenterology. ChatGPT 3.5 Default demonstrated improvement with a 57% overall accuracy, making progress in clinical science, endocrinology and ophthalmology. Nevertheless, its performance in gastroenterology remained relatively weak. ChatGPT-4, the latest version, achieved the highest accuracy (64%), showcasing consistent performance across all disciplines.

In terms of subtopics, trends indicated that ChatGPT 3.5 Default and ChatGPT-4 outperformed their predecessor in several areas. For symptoms and signs, both ChatGPT-3

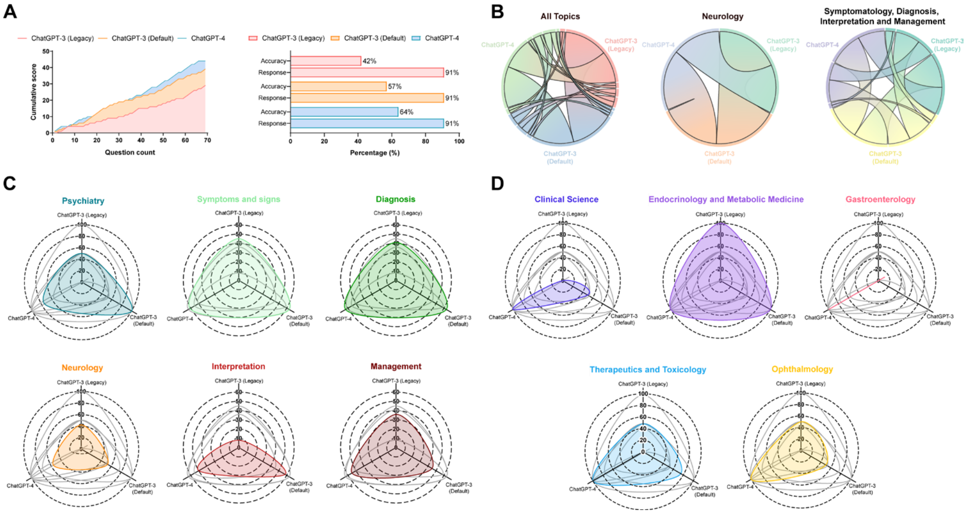

Figure 1 Comparative performance of ChatGPT-3 Legacy, ChatGPT-3 Default and ChatGPT-4 on SCE Neurology Questions. Accuracy and rate of each model presented as percentage of correct responses and score count (A). Co-occurrence of accurate responses across different disciplines and subtopics (B). Performance on relevant topics and subtopics in the field of neurology (C). Performance of each model in the remaining disciplines outside of neurology (D). SCE, specialty certificate examination.

Default and ChatGPT-4 outperformed the legacy version with respective accuracies of 63% and 61%, compared with the 45% accuracy of the legacy model. The trend continued in diagnosis category, where the ChatGPT 3.5 Default and ChatGPT-4 models again showcased superior performance, both achieving an accuracy of 63% against the 41% of the legacy version. The shift was even more dramatic in interpretation, with the legacy model lagging at 8% while the default and ChatGPT-4 models achieved accuracies of 58% and 50%, respectively. Finally, in the management category, all three models showed improvement, but ChatGPT-4 took the lead with an accuracy of 55%, surpassing both the default and legacy models which scored 45% and 36%, respectively. Overall, the default and ChatGPT-4 models showed a significant performance boost, both averaging an accuracy of 57% in contrast to the 33% accuracy of the legacy model.

According to the metrics for the 2022 SCE in neurology, the pass mark was set at 409, corresponding to 58% or 114 out of 197 questions. In this context, ChatGPT-3's performance of 42% and 57% falls short of the passing threshold, while GPT-4's performance of 64% surpasses it. The pass rate for UK trainees in 2022 was 79.6%, while the pass rate for all candidates was 60.2%. The highest pass rate among UK trainees was observed for ST5 trainees, with a sitting number of 32 and a pass rate of 87.5%.

DISCUSSION

We evaluated and compared the performance of ChatGPT 3.5 Legacy, ChatGPT 3.5 Default and ChatGPT-4 models on a diverse dataset of questions from the Pool—SCE Neurology Web Questions bank, assessing their proficiency across various medical disciplines, primarily within the neurology specialisation. Key subtopics included symptoms and signs, diagnosis, interpretation and management along the understanding of specific patient populations and underlying mechanisms. While all three models demonstrated varying levels of performance, ChatGPT-4 showed the most promise in terms of overall accuracy and consistency. According to the metrics for the 2022 SCE in neurology, ChatGPT-3's performance of 42% and 57% falls short of the passing threshold, while GPT-4's performance of 64% exceeds the required 58% to pass.

The findings of this study indicate that ChatGPT3s' performance in the specialised field of neurology and neuroscience is lower than its performance in more general medical examinations, such as the USMLE. This discrepancy can be attributed to several factors that differentiate entry-level examinations from specialty-level examinations such as the SCE in neurology.

Specialty examinations require a deeper understanding of specific medical domains compared with entry-level exams. While ChatGPT has been trained on a vast amount of medical literature, it may not possess the same level of expertise in the nuances of neurology and neuroscience as a specialist would. Furthermore, specialty exams often present more complex clinical scenarios that require higher-level reasoning skills and the ability to synthesise information from multiple sources. ChatGPT, while capable of understanding context, may struggle with the intricacies of these specialised scenarios, leading to a lower success rate.

Another potential reason for ChatGPT's lower performance in the specialised field is the limitations in its training data. Its knowledge is based on the information available up to September 2021, which may not be up to date with the latest advancements and guidelines in neurology and neuroscience. Additionally, ChatGPT was unable to answer 7% of the questions in our study due to their reliance on images. As specialty exams often require the interpretation of various visual aids, such as brain scans and neurological charts, ChatGPT's inability to process images might be a significant limiting factor in its performance.

ChatGPT-4 demonstrated promising proficiency in the specialised field of neurology, outperforming its predecessors such as ChatGPT-3. This ability to attain specialty-level medical knowledge can be attributed to several factors. A crucial factor is that ChatGPT-4 was trained on a larger and more diverse dataset containing up to date, specialised medical information, enabling the model to develop a deeper understanding of complex medical knowledge. Furthermore, an advanced architecture potentially allows ChatGPT-4 to perform complex reasoning and effectively synthesise information from multiple sources, which is essential for handling intricate clinical scenarios and gaining a comprehensive grasp of the subject matter within specialty-level examinations. Additionally, refined training techniques contribute to the model's enhanced performance, making it more efficient and effective at capturing nuances in specialised medical knowledge. Finally, the implementation of mechanisms that promote creativity and flexibility in the model's responses enables it to generate accurate and contextually relevant answers.

The findings of this study have important implications for the use of AI models such as ChatGPT in medical education and practice. While earlier versions of ChatGPT demonstrated limitations in the specialised field of neurology and neuroscience, the more advanced ChatGPT-4 has shown promise in attaining specialty-level medical knowledge, setting a new benchmark for AI models in the context of specialised medical education and practice. This development suggests that it may be more effective for advanced medical training. Medical professionals and educators should consider these improvements when evaluating the use of AI models such as ChatGPT in their practice or teaching.

These findings also emphasise the importance of continually updating AI models with the latest medical knowledge to ensure that they remain relevant and reliable in the rapidly evolving field of medicine. Collaboration between AI developers and medical experts is crucial in achieving this goal and ensuring that AI models can effectively support medical professionals in their practice and education.

CONCLUSIONS

In conclusion, this study demonstrates that while earlier versions of ChatGPT had limitations in the specialised field of neurology and neuroscience, the more advanced ChatGPT-4 has shown promise in attaining specialty-level medical knowledge. This sets a new benchmark for AI models in specialised medical education and practice, highlighting the potential for further development in the medical domain. These improvements raise important considerations for the use of AI models such as ChatGPT in medical education and practice, emphasising the need for ongoing updates and collaboration between AI developers and medical experts to ensure their effectiveness in supporting medical professionals.

Contributors PG conceptualised the idea, analysed the data and drafted the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no datasets generated and/or analysed for this study.

ORCID iD

Panagiotis Giannos http://orcid.org/0000-0003-1037-1983

REFERENCES

1 Khurana D, Koli A, Khatter K, *et al*. Natural language processing: state of the art, current trends and challenges. *Multimed Tools Appl* 2023;82:3713–44.
2 Kevin S. Microsoft teams up with OpenAi to exclusively license GPT-3 language model. n.d. Available: https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/
3 Nagarhalli TP, Vaze V, Rana NK. A review of current trends in the development of Chatbot systems. 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS); Coimbatore, India.2020:706–10
4 Hutson M. Could AI help you to write your next paper *Nature* 2022;611:192–3.
5 Stokel-Walker C. AI Bot ChatGPT writes smart essays - should professors worry? *Nature* 2022. 10.1038/d41586-022-04397-7 [Epub ahead of print 9 Dec 2022].
6 Sabry Abdel-Messih M, Kamel Boulos MN. ChatGPT in clinical toxicology. *JMIR Med Educ* 2023;9:e46876.
7 Lo CK. What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences* 2023;13:410.
8 Kung TH, Cheatham M, Medenilla A, *et al*. Potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2:e0000198.
9 Giannos P, Delardas O. Performance of ChatGPT on UK standardized admission tests: insights from the BMAT, TMUA, LNAT, and TSA examinations. *JMIR Med Educ* 2023;9:e47737.