BMJ Neurology Open

Assessment of ChatGPT's performance on neurology written board examination questions

Tse Chiang Chen,[1] Evan Multala,[2] Patrick Kearns,[2] Johnny Delashaw,[3] Aaron Dumont,[3] Demetrius Maraganore,[1] Arthur Wang [3]

To cite: Chen TC, Multala E, Kearns P, *et al*. Assessment of ChatGPT's performance on neurology written board examination questions. *BMJ Neurology Open* 2023;**5**:e000530. doi:10.1136/bmjno-2023-000530

Received 07 September 2023
Accepted 19 October 2023

[1]Neurology, Tulane University School of Medicine, New Orleans, Louisiana, USA
[2]Tulane University School of Medicine, New Orleans, Louisiana, USA
[3]Neurosurgery, Tulane University School of Medicine, New Orleans, Louisiana, USA

Correspondence to
Dr Arthur Wang;
awang15@tulane.edu

ABSTRACT

Background and objectives ChatGPT has shown promise in healthcare. To assess the utility of this novel tool in healthcare education, we evaluated ChatGPT's performance in answering neurology board exam questions.

Methods Neurology board-style examination questions were accessed from BoardVitals, a commercial neurology question bank. ChatGPT was provided a full question prompt and multiple answer choices. First attempts and additional attempts up to three tries were given to ChatGPT to select the correct answer. A total of 560 questions (14 blocks of 40 questions) were used, although any image-based questions were disregarded due to ChatGPT's inability to process visual input. The artificial intelligence (AI) answers were then compared with human user data provided by the question bank to gauge its performance.

Results Out of 509 eligible questions over 14 question blocks, ChatGPT correctly answered 335 questions (65.8%) on the first attempt/iteration and 383 (75.3%) over three attempts/iterations, scoring at approximately the 26th and 50th percentiles, respectively. The highest performing subjects were pain (100%), epilepsy & seizures (85%) and genetic (82%) while the lowest performing subjects were imaging/diagnostic studies (27%), critical care (41%) and cranial nerves (48%).

Discussion This study found that ChatGPT performed similarly to its human counterparts. The accuracy of the AI increased with multiple attempts and performance fell within the expected range of neurology resident learners. This study demonstrates ChatGPT's potential in processing specialised medical information. Future studies would better define the scope to which AI would be able to integrate into medical decision making.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Artificial intelligence shows potential in healthcare. The performance of ChatGPT has been examined for the United States Medical Licensing Examination and neurosurgery written board exams.

WHAT THIS STUDY ADDS

⇒ From this study, we learnt that ChatGPT can correctly answer neurology-simulated written board exam multiple-choice questions in an accurate and reproducible way. This study demonstrates the ability of ChatGPT's ability to recall factual knowledge and think critically.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study paves the way for future studies investigating ChatGPT's ability to think critically and ways for ChatGPT to be integrated into medical education.

INTRODUCTION

The development of artificial intelligence (AI) models has risen significantly. Machine learning and augmentation of the human–technology interface are applied across a wide range of professional settings such as robotics, transportation and software design. AI systems have also made their way into the healthcare field.[1]

AI is trained by processing data sets provided by the user which are then analysed through working algorithms to output a solution. The intelligence of AI is synonymous with this acquisition of data which can then be patterned for recall to navigate future tasks. Generative Pre-trained Transformer (GPT4, Open AI, San Francisco, California, USA), often referred to as ChatGPT, is a natural language processing (NLP) model programmed to generate responses to human prompts. It has billions of parameters that allow it to learn through massive amounts of data and can interpret complex text to answer in a human-like fashion. It does this by learning and fine-tuning responses through analysis of massive amounts of data, thus enabling it to generate improved responses over time. In healthcare, ChatGPT has been used in helping to diagnose and manage rare complex cases.[2] Additionally, ChatGPT's effectiveness in completing specific radiologic tasks to augment the workflow of physicians has been demonstrated.[3]

The role and efficacy of AI in medicine is currently being assessed. Some view it as a necessary advancement to facilitate the work of physicians while others view AI as a direct

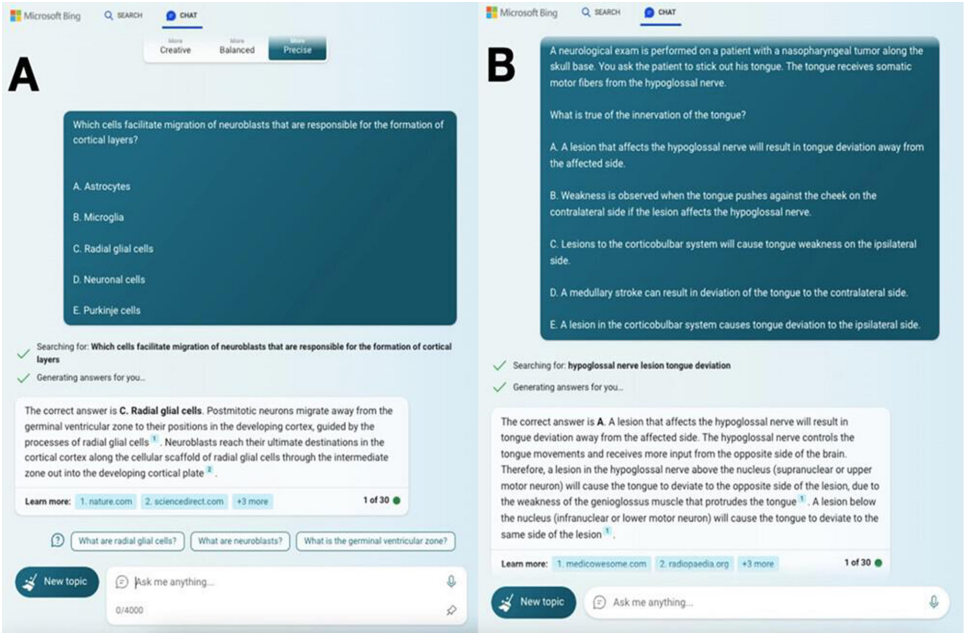

Figure 1 Neurology question stems. (A) Example of a question input followed by a correct answer response from the ChatGPT. (B) Example of a question input followed by an incorrect answer response from the AI (the correct answer is B). AI, artificial intelligence.

encroachment on the integrity and safety of medical practice.[4–6]

In order for AI services to be applied in a clinical setting, they must facilitate the workflow of health practitioners through efficiency and accuracy. It is therefore important to assess these AI systems' ability to analyse and answer specialty-specific board-style exam questions. ChatGPT has already demonstrated the ability to pass the United States Medical Licensing Examination (USMLE) Step One and Two, as well as advanced licensing exams such as the neurosurgical and radiology primary board examinations.[7–10] Therefore, it is reasonable to explore ChatGPT's performance on neurology board-like exam questions, which has not been reported in the literature to date.

Neurology, being a complicated and specialised field that requires meticulous training, provides an optimal environment to test AI on advanced board questions.[11–15] We assess the performance of ChatGPT on a widely used commercial question bank to simulate performance on the American Board of Psychiatry and Neurology board certification exam. We aim to quantify ChatGPT's ability to process highly specialised medical information.

METHODS

BoardVitals (BV) is an online question bank used by medical residents to prepare for a variety of medical specialty board certification exams. It is accredited by the Accreditation Council for Continuing Medical Education to provide continuing medical education to physicians. BV covers more than 50 specialty fields across medicine, dentistry and pharmacy. Content is written by physician authors and references major medical/healthcare textbooks and publications. ChatGPT is an AI chatbot developed by OpenAI in 2022. ChatGPT interacts with users by following a human-generated prompt and provides a detailed response. This translates into a seamless dialogue, simulating a human-like conversation.

Protocol

We used ChatGPT (GPT4) to analyse a total of 560 BV neurology questions (14 blocks of 40 questions each). For each question, the difficulty level of the question (ie, easy, medium and hard) and the topic category were recorded. Specifically, question categories included basic neuroscience; behavioural, cognitive, psychiatry; cerebrovascular; child neurology; congenital; cranial nerves; critical care; demyelinating disorders; epilepsy and seizures; ethics; genetic; headache; imaging/diagnostic studies; movement disorders; neuro-ophthalmology; neuro-otology; neuroinfectious disease; neurologic complications of systemic disease; neuromuscular; neurotoxicology, nutrition, metabolic; oncology; pain; pharmacology; pregnancy; sleep; and trauma.

For each question, the entire question stem along with the answer choices was copied and pasted into the ChatGPT input section (figure 1). Two researchers (TCC, AW) analysed the responses provided by ChatGPT and compared them to the answers provided by BV. The chat input was cleared and refreshed before beginning a new question. ChatGPT-generated answers were recorded for every question; specifically, two scores were recorded: one for the first attempt/iteration, and if the initial answer was incorrect, a second score was recorded for up to three attempts/iterations. The chat was cleared prior to each attempt. The reasoning for allowing multiple attempts is that ChatGPT is an NLP model that creates responses

one word at a time based on probabilities such that each new attempt at the same question can generate a different answer. It was reasoned that as an active learning programme, ChatGPT should be provided the opportunity to learn from these mistakes as was intended by its pattern recognition algorithm. While not reflective of real-world testing conditions, this was done with the intention of improving the accuracy of the programme.

As question stems are multiple choice and accompanied by lettered options (eg, answer A, B, C, etc), if ChatGPT provided a written answer without referencing a specific alphabetic answer choice, the answer was read in full to determine if it was reflective of any of the multiple-choice options and scored accordingly. As these are multiple-choice questions, ChatGPT was asked to provide a single best answer.

Image-based questions from the question bank were deemed ineligible and not included in the assessment as ChatGPT only accepts text input.

The percentage of questions correctly answered was compiled and compared with an expected percentage of correct answers. For each question, BV provides a summary of the percentage of learners who answered that specific question correctly. The expected percentage was the average of the percentage of learners (neurology residents) who correctly answered each question averaged over all eligible questions. The expected number of correct answers by learners was calculated by multiplying the expected percentage of correct answers by the number of eligible questions. BV does not report the experience level of the neurology residents answering that specific question (ie, Post Graduate Year).

The overall performance percentile of ChatGPT's first attempt was calculated by the question bank, as compared with other BV learners, and recorded. The percentages of questions correctly answered by ChatGPT on the first attempt and over three attempts were compared with the performance quartiles as provided by BV. Percentages correct were also compiled by subject matter (ie, neuro-oncology, critical care, etc). This analysis aimed to provide an alternate way to compare ChatGPT's performance in relation to other learners using the question bank.

RESULTS

Of 560 questions (14 question blocks of 40 questions each), 509 questions were eligible (ie, not image-based) for the study. Of these 509 eligible questions, ChatGPT correctly answered 335 questions on the first attempt and 383 over three attempts, translating to a percentage of correct questions of 65.8% and 75.3%, respectively (table 1).

The worst performing blocks resulted in a 50% and 62.2% total correct, respectively, between the first attempt and third attempt (table 1). The highest performing blocks scored 92.1% (SD=12.8%) with a single attempt and 97.2% (SD=9.6%) with three iterations/attempts. The expected percentage of correct answers based on BV learners is 72.63%, translating to an expected 369.7 correct answers. For 51 questions, ChatGPT either refused to answer the question, provided an answer that was not in the multiple choice, or replied that there were multiple correct answers. In one question pertaining to the populations of depression and suicide, ChatGPT refused to answer and referred us to a suicide hotline.

Table 1 Cumulative performance of ChatGPT versus users over 14 question blocks (40 questions each)

# of eligible questions (excluding picture questions)	509
Total correct answers (one attempt/iteration)	335 (65.82%)
Total correct answers (three attempts/iteration)	383 (75.25%)
Expected # of correct answers by users	369.70 (72.63%)

When tabulating the quartiles of the reported percentage of learners choosing the correct answers, the first, second and third quartiles (corresponding to the 25th, 50th and 75th percentile) were 61%, 75% and 85%, respectively (table 2). When comparing the overall percentage correct of ChatGPT, the first attempt falls just above the 25th percentile, whereas the run over three attempts falls just above the 50th percentile. This is comparable with what was reported by the official Performance Tracker offered by the commercial question bank, which rated ChatGPT at the 26th percentile for its first iteration.

A t-test was performed to evaluate for any statistical difference between the % of correct responses given by ChatGPT versus users (tables 3 and 4).

In terms of scores by subject matter, the highest performing subjects by per cent correct were pain (100%), epilepsy & seizures (85%) and genetic (82%) while the lowest performing subjects were imaging/diagnostic studies (27%), critical care (41%) and cranial nerves (48%) (table 5).

DISCUSSION

In this study, we evaluated ChatGPT's performance on a neurology board-like test using the BV online question bank. BV provides commercially available question banks for a variety of examinations and medical specialties including neurology. It is accredited by the Accreditation Council for Continuing Medical Education to provide continuing medical education to physicians.[16] A third-party survey found that BV users had a 95% pass rate on the Neurology Board Exam compared with the national average of 89% and that 70% of respondents thought BV helped improve their Neurology Board Exam score.[17]

The results of this study demonstrate reasonable performance on the first attempt/iteration, which further improved with subsequent attempts. ChatGPT's performance falls within the expected performance

Table 2 Minimum, maximum and SD of percentages correct by question block (40 question blocks) between ChatGPT and users

	% Correct by ChatGPT (first attempt/iteration)	% Correct by ChatGPT (third attempt/iteration)	Expected % correct by users
Min	50.00	62.16	69.83
Max	92.11	97.22	78.00
SD	12.78	9.62	2.33

ranges as established by neurology learners using the BV question bank. Our results highlight ChatGPT's ability to interpret and answer appropriately to clinical questions and vignettes. Previous work assessing ChatGPT's performance on various medical exams has been written about. For practice questions simulating the USMLE Step exams, ChatGPT correctly answered between 55.8% and 61.3% of the questions and on the practice questions for the American Board of Neurological Surgery board exam questions, ChatGPT answered 53.2% correctly on the first attempt.[10 18] The results of our study showing 65.82% correct answers on the first attempt is slightly higher in comparison, which could be a result of the difference in availability of specialty-specific materials in ChatGPT's database that can be used and potential improvement or patches to the existing ChatGPT model or database. ChatGPT performed statistically worse than the users after one attempt at answering the questions. This difference was not present after ChatGPT was given three attempts at answering the question.

It is interesting to note that ChatGPT was sensitive to questions regarding depression and suicide and referred us to a suicide hotline. This suggests either a naturally grown sensitivity towards certain issues like suicide, or that 'guardrails' are implemented from a top-down approach to the system. If the latter, it suggests that more hard-coded guidance could be implemented to tailor the system towards healthcare, and perhaps even narrowing the scope of the AI to healthcare specialties. Furthermore, it may be possible to have Large Language Models (LLMs) act as guides or tutors alongside question banks or as an interactive chatbot when reviewing medical concepts in online reference materials.

One of the criticisms of ChatGPT is the application's ability to 'hallucinate'. In this situation of hallucination, the answer by ChatGPT is factually incorrect but the answer is provided in a way that is very reasonable and convincing that it almost always 'looks' correct. While

Table 3 Comparison of % correct answers between ChatGPT (one attempt) and users out of 509 eligible questions

% Correct by ChatGPT (one attempt/iteration)	Expected % correct by users	Statistical significance p value
335 (65.82%)	369.70 (72.63%)	0.0014

P value <0.05 is statistically significant.

Table 4 Comparison of % correct answers between ChatGPT (three attempts) and users out of 509 eligible questions

% Correct by ChatGPT (three attempts/iteration)	Expected % correct by users	Statistical significance p value
383 (75.25%)	369.70 (72.63%)	0.295

P value <0.05 is statistically significant.

'hallucinations' are a common criticism of ChatGPT, no 'hallucinations' were detected in our trials per se; this may be due to multiple reasons including posing a limited, constrained scenario (in the form of a question stem) to ChatGPT, and the fact that it may be difficult to discern an incorrect answer from a 'hallucination'. This problem can be difficult to discern and thus, any medical use of ChatGPT must incorporate steps to verify the accuracy of its answers.

ChatGPT is still a relatively young technology at the time of this writing, and we anticipate room for improvement. For example, the integration of plugins could enhance its current abilities; incorporating the WolframAlpha plugin could improve its weaknesses

Table 5 Performance of ChatGPT by subject area

Subject	Score	Correct	Incorrect
Basic neuroscience	66%	27	14
Behavioural, cognitive, psych	77%	62	19
Cerebrovascular	73%	18	7
Child neurology	50%	10	10
Congenital	79%	11	3
Cranial nerves	48%	12	13
Critical care	41%	7	10
Demyelinating disorders	74%	14	5
Epilepsy, seizures	85%	11	2
Ethics	67%	3	2
Genetic	82%	14	3
Headache	82%	14	3
Imaging/diagnostic studies	27%	3	8
Movement disorders	64%	18	10
Neuro-ophthalmology	74%	14	5
Neuro-otology	53%	10	9
Neuroinfectious disease	56%	9	7
Neurologic complications of systemic disease	63%	10	6
Neuromuscular	65%	20	11
Neurotoxicology, nutrition, metabolic	67%	20	10
Oncology	53%	9	8
Pain	100%	20	0
Pharmacology	64%	16	9
Pregnancy	81%	13	2
Sleep	57%	13	10
Trauma	50%	7	7

in mathematics.[19] The WolframAlpha plugin is an additional add on AI tool to enhance ChatGPT performance. With this installation, ChatGPT can be turned into a powerful computational tool in order to perform accurate mathematics, curate knowledge to be more precise, and provides real-time data monitoring. Although the base model lacks visual input and thus is currently unable to elucidate image-based questions, collaborations with visual accessibility companies such as Be My Eyes could potentially yield exciting results.[20] Be My Eyes is a first-ever digital visual assistant that is powered by ChatGPT language model to provide blind people with a powerful new resource to navigate their physical environments, address their activities of daily living needs and gain more independence. Users can send images via the Be My Eyes application which can answer any questions about that image and provides immediate visual assistance for an array of tasks. Once informatically matured, it could have significant implications for the medical field.

There are several limitations associated with this study. We used a commercial question bank without access to the underlying data to verify the provided statistics. Moreover, the study did not involve official (mock) exams provided by the American Academy of Neurology, thus predictions about its performance on the neurology board exams would be speculative. There is also concern about the model's performance in clinical reasoning. ChatGPT, as an LLM, produces structured texts based on probabilities but is also prone to state 'facts' that are untrue without self-awareness. This makes it difficult to be trusted at higher levels of clinical decision making. Last, the comparison between human candidates and ChatGPT may be more complicated. The board exam tests the ability of candidates to draw on their memory bank and must rely on their unaided memory to answer the question in real time. ChatGPT has access to large amounts of online information that it can curate for the best answer to the question. Human candidates allowed freely to use electronic devices to answer questions may perform much better than the results quoted in our paper.

Lastly, future studies into ChatGPT can potentially address ChatGPT's ability to simply recall facts versus synthesise information together to perform next-step thinking. Multiple-choice questions test two things (1) the possession of a key piece of information for factual recall and (2) the ability to synthesise those facts to solve a problem. The second ability is considered higher order thinking that requires the learner to think more critically. Ultimately, residency training programmes and curriculum encourage the development of critically thinking physicians who can think critically and sometimes out of the box when faced with clinical patient scenarios. While we did not specifically test ChatGPT's critical thinking skills in this manuscript, the next questions about ChatGPT's higher order thinking can be investigated in the future.

CONCLUSION

ChatGPT is a natural language processing model whose pragmatic use is recently being explored in the field of healthcare. It has successfully taken the USMLE Step exams as well as neurosurgery board-styled exams. We demonstrate that its capabilities extend to the field of neurology. These results add to the growing literature of the capabilities of AI and offer a glimpse into a future potential for a safe and productive collaboration between neurologists and AI.

Contributors TCC contributed to data acquisition/analysis, writing–original draft, writing–review and editing, investigation. EM contributed to data acquisition/analysis, writing–review and editing, investigation. PK contributed to data acquisition/analysis, writing–review and editing, investigation. JD contributed to review and editing. AD contributed to review and editing. DM contributed to review and editing. AW contributed to conceptualization, supervision, writing–review and editing, methodology, investigation and project administration, AW is guarantor of this 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.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iD

Arthur Wang http://orcid.org/0009-0007-4396-2516

REFERENCES

1 Jiang F, Jiang Y, Zhi H, *et al*. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017;2:230–43. 10.1136/svn-2017-000101 Available: https://pubmed.ncbi.nlm.nih.gov/29507784/
2 Suhag A, Kidd J, McGath M, *et al*. CHATGPT: a pioneering approach to complex Prenatal differential diagnosis. *Am J Obstet Gynecol MFM* 2023;5:101029. 10.1016/j.ajogmf.2023.101029 Available: https://doi.org/10.1016/j.ajogmf.2023.101029
3 Rao A, Kim J, Kamineni M, *et al*. Evaluating Chatgpt as an adjunct for radiologic decision-making. *medRxiv* 2023:2023.02.02.23285399.
4 Layard Horsfall H, Palmisciano P, Khan DZ, *et al*. Attitudes of the surgical team toward artificial intelligence in Neurosurgery: International 2-stage cross-sectional survey. *World Neurosurg* 2021;146:e724–30.
5 Chen P-H, Liu Y, Peng L. How to develop machine learning models for Healthcare. *Nat Mater* 2019;18:410–4.
6 Derevianko A, Pizzoli SFM, Pesapane F, *et al*. The use of artificial intelligence (AI) in the Radiology field: what is the state of doctor-patient communication in cancer diagnosis? *Cancers (Basel)* 2023;15:470.
7 Gilson A, Safranek CW, Huang T, *et al*. How does Chatgpt perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312.

8 Hopkins BS, Nguyen VN, Dallas J, *et al*. Chatgpt versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board-style questions. *J Neurosurg* 2023;139:904–11.
9 Kedar S, Khazanchi D. Neurology education in the era of artificial intelligence. *Curr Opin Neurol* 2023;36:51–8.
10 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.
11 McDermott MBA, Wang S, Marinsek N, *et al*. Reproducibility in machine learning for health research: still a ways to go. *Sci Transl Med* 2021;13:eabb1655.
12 Patel UK, Anwar A, Saleem S, *et al*. Artificial intelligence as an emerging technology in the current care of neurological disorders. *J Neurol* 2021;268:1623–42.
13 Pedersen M, Verspoor K, Jenkinson M, *et al*. Artificial intelligence for clinical decision support in neurology. *Brain Commun* 2020;2:fcaa096.
14 Thomas LB, Mastorides SM, Viswanadhan NA, *et al*. Artificial intelligence: review of current and future applications in medicine. *Fed Pract* 2021;38:527–38. 10.12788/fp.0174 Available: https://www.mdedge.com/fedprac/issue/248403/federal-practitioner-3811a
15 Vishnu VY, Vinny PW. The neurologist and artificial intelligence: Titans at crossroads. *Ann Indian Acad Neurol* 2019;22:264–6.
16 Board Vitals. Neurology board review questions and practice tests. 2023. Available: https://www.boardvitals.com/neurology-board-review
17 Board Vitals. Boardvitals neurology board results 2014 / 2015. 2015. Available: https://www.boardvitals.com/neurology-board-results
18 Hopkins BS, Nguyen VN, Dallas J, *et al*. Chatgpt versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board–style questions. *J Neurosurg* 2023;139:904–11.
19 Wolfram Alpha. Wolfram Plugin for Chatgpt. 2023. Available: https://www.wolfram.com/wolfram-plugin-chatgpt/ [Accessed 12 Jun 2023].
20 Be My Eyes. Be my eyes. Available: https://www.bemyeyes.com/ [Accessed 20 Jun 2023].