Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank

Rohaid Ali, MD[‡*], Oliver Y. Tang, MD[‡*], Ian D. Connolly, MD, MS[§*], Jared S. Fridley, MD[‡], John H. Shin, MD[§], Patricia L. Zadnik Sullivan, MD[‡], Deus Cielo, MD[‡], Adetokunbo A. Oyelese, MD, PhD[‡], Curtis E. Doberstein, MD[‡], Albert E. Telfeian, MD, PhD[‡], Ziya L. Gokaslan, MD[‡], Wael F. Asaad, MD, PhD[‡‖¶#]

[‡]Department of Neurosurgery, The Warren Alpert Medical School of Brown University, Providence, Rhode Island, USA; [§]Department of Neurosurgery, Massachusetts General Hospital, Boston, Massachusetts, USA; [‖]Norman Prince Neurosciences Institute, Rhode Island Hospital, Providence, Rhode Island, USA; [¶]Department of Neuroscience, Brown University, Providence, Rhode Island, USA; [#]Carney Institute for Brain Science, Brown University, Providence, Rhode Island, USA

*Rohaid Ali, Oliver Y. Tang, and Ian D. Connolly contributed equally to this work.

Correspondence: Rohaid Ali, MD, Department of Neurosurgery, Rhode Island Hospital, LPG Neurosurgery, 593 Eddy St, APC6, Providence, RI 02903, USA. Email: RAli@lifespan.org

Received, March 16, 2023; **Accepted,** April 09, 2023; **Published Online,** June 12, 2023.

BACKGROUND AND OBJECTIVES: General large language models (LLMs), such as ChatGPT (GPT-3.5), have demonstrated the capability to pass multiple-choice medical board examinations. However, comparative accuracy of different LLMs and LLM performance on assessments of predominantly higher-order management questions is poorly understood. We aimed to assess the performance of 3 LLMs (GPT-3.5, GPT-4, and Google Bard) on a question bank designed specifically for neurosurgery oral boards examination preparation.

METHODS: The 149-question Self-Assessment Neurosurgery Examination Indications Examination was used to query LLM accuracy. Questions were inputted in a single best answer, multiple-choice format. χ^2, Fisher exact, and univariable logistic regression tests assessed differences in performance by question characteristics.

RESULTS: On a question bank with predominantly higher-order questions (85.2%), ChatGPT (GPT-3.5) and GPT-4 answered 62.4% (95% CI: 54.1%-70.1%) and 82.6% (95% CI: 75.2%-88.1%) of questions correctly, respectively. By contrast, Bard scored 44.2% (66/149, 95% CI: 36.2%-52.6%). GPT-3.5 and GPT-4 demonstrated significantly higher scores than Bard (both $P < .01$), and GPT-4 outperformed GPT-3.5 ($P = .023$). Among 6 subspecialties, GPT-4 had significantly higher accuracy in the Spine category relative to GPT-3.5 and in 4 categories relative to Bard (all $P < .01$). Incorporation of higher-order problem solving was associated with lower question accuracy for GPT-3.5 (odds ratio [OR] = 0.80, $P = .042$) and Bard (OR = 0.76, $P = .014$), but not GPT-4 (OR = 0.86, $P = .085$). GPT-4's performance on imaging-related questions surpassed GPT-3.5's (68.6% vs 47.1%, $P = .044$) and was comparable with Bard's (68.6% vs 66.7%, $P = 1.000$). However, GPT-4 demonstrated significantly lower rates of "hallucination" on imaging-related questions than both GPT-3.5 (2.3% vs 57.1%, $P < .001$) and Bard (2.3% vs 27.3%, $P = .002$). Lack of question text description for questions predicted significantly higher odds of hallucination for GPT-3.5 (OR = 1.45, $P = .012$) and Bard (OR = 2.09, $P < .001$).

CONCLUSION: On a question bank of predominantly higher-order management case scenarios for neurosurgery oral boards preparation, GPT-4 achieved a score of 82.6%, outperforming ChatGPT and Google Bard.

KEY WORDS: Neurosurgery, Medical education, Surgical education, Residency education, Artificial intelligence, Large language models, ChatGPT

Neurosurgery 93:1090–1098, 2023 https://doi.org/10.1227/neu.0000000000002551

Growing interest has surrounded the ability of artificial intelligence (AI) to guide clinical decision-making and care, especially given the recent documentation of the ability of general large language models (LLMs), such as ChatGPT (OpenAI), to pass graduate-level and certification examinations in fields including medicine,[1] law,[2] and business.[3] In a prior analysis, we found that ChatGPT achieved a passing score of 73.4% on a 500-question module emulating the neurosurgery written board examinations, with lower accuracy on questions that were lengthier, incorporated higher-order problem solving, or involved imaging.[4] Although ChatGPT has been available for public use since November 2022 as a "GPT-3.5" release, OpenAI released an

ABBREVIATIONS: AI, artificial intelligence; **LLM,** large language model; **SANS,** Self-Assessment Neurosurgery Examination.

updated model, GPT-4, on March 14, 2023. Similar to its predecessor, GPT-4 was trained using both supervised and unsupervised learning techniques on a large corpus of Internet text data, followed by fine-tuning through reinforcement learning with human feedback. GPT-4 has achieved passing scores in over 25 standardized examinations, including scoring in the 90th percentile of a simulated bar examination, compared with GPT-3.5 scoring in the 10th percentile.[5] Evidence of performance improvements of more than 20% on all 3 US Medical Licensing Examinations has also been documented.[6] GPT-4 additionally has introduced multimodal capabilities, including the ability to evaluate image inputs, which have yet to be released for public use.

In response to the popularity of ChatGPT and GPT-4, various leading software companies have introduced their own language models, showcasing remarkable advancements in artificial intelligence. One such example is the Bard chatbot, developed by Google's parent company, Alphabet Inc. Launched on March 21, 2023, Bard has garnered considerable attention as Google's foray into the chatbot domain, sparking intriguing discussions about the future of search technology.

A key distinction between Bard and ChatGPT and GPT-4 lies in Bard's ability to access and incorporate information from the internet in real time when generating responses. This contrasts with ChatGPT and GPT-4, which rely on prior training data up until September 2021 and do not have webcrawling capabilities incorporated presently. Incorporating real-time web search capabilities, Bard, in theory, could offer users more current and contextually pertinent information. However, direct comparisons between the 2 models are only just beginning to be undertaken. Notably, there have been no head-to-head comparisons of Bard and ChatGPT within the context of any clinical board examination.

The performance of LLMs such as GPT-4 on open-ended oral medical examinations is less understood. In the setting of neurosurgery, the American Board of Neurological Surgery (ABNS) oral board examination is composed of three 45-minute sessions and are most commonly taken 2–3 years after residency graduation, in contrast to the written board examination that is intended for earlier-stage trainees.[7] The oral board examination is widely considered to be the more rigorous and difficult assessment. Although the first-time pass rate for the American Board of Neurological Surgery written board examination has exceeded 96% since 2018, the pass rate for the oral board examination has ranged between 81% and 90% during the same period.[8]

The goals of this study were to (1) assess the performance of 3 LLMs on a question bank with higher-order questions more representative of oral board topics and (2) elucidate differences in accuracy and performance by question characteristics between LLMs.

METHODS

This study assessed the performance of 3 LLMs: ChatGPT, GPT-4, and Google Bard. Performance of LLMs on the neurosurgery oral board examination was proxied using the Self-Assessment Neurosurgery Examination (SANS) Indications Examination, a 149-question module designed specifically for oral boards preparation, focusing on surgical indications, diagnostic testing, and interventional decision-making. Notably, this module is written in a multiple-choice format, which differs from the open-ended approach of the oral boards. As described earlier,[4,9] question characteristics, including subspecialty, word length, and incorporation of higher-order problem solving, were independently collected by 2 authors (RA and OYT), with disagreements adjudicated by a third (IDC). All classification was blinded, without investigator knowledge of any LLM's answers to the questions.

Questions were assessed in a single best answer multiple-choice format, with the question stem reproduced verbatim (Figure 1). Because multimodal input has yet to be incorporated into any of the 3 LLMs, questions with imaging data had only the text portion of the question stem input for evaluation. Questions that an LLM opted not to answer, such as because of citing insufficient contextual data, were classified as incorrect. In addition, for questions with imaging, we tracked responses with confabulations or "hallucinations," which were defined as scenarios where an LLM asserted inaccurate facts or contextual data that it falsely believed were correct in its answer. This phenomenon has been well-documented by OpenAI among ChatGPT and, to a lesser degree, GPT-4.[5] Data were collected on 3/12/23 for ChatGPT, 3/14/23 (day of release) for GPT-4, and 4/1/23 for Google Bard.

All analyses and visualizations were performed using R version 4.1.2 (Foundation for Statistical Computing) and the Matplotlib package on Python (Python Software Foundation), respectively. Associations between category-level performances were queried using linear regression. Differences in performance were assessed using χ^2, Fisher exact, and univariable logistic regression tests. For all analyses, statistical significance was assessed at $P < .05$. This study was conducted in accordance with Strengthening the Reporting of Observational Studies in Epidemiology reporting guidelines.

Data Availability Statement

Owing to the proprietary nature of the data set used for this study (Self-Assessment Neurosurgery Examination: Indications Examination), the authors are unable to post the raw data used for the analysis. However, the authors are able to share any collected data (eg. word count, question classification, responses, etc.) on request to other investigators who have access to this self-assessment examination.

Code Availability Statement

Code used for this study's analyses was adapted from earlier scripts deposited in a public GitHub repository (https://github.com/oliverytang/chatgpt_neurosurgery).

RESULTS

Performance of ChatGPT, GPT-4, and Bard on Indications Question Bank

On the 149-question question bank focused on oral boards topics, GPT-3.5 (ChatGPT) and GPT-4 achieved scores of 62.4% (93 of 149, 95% CI: 54.1%-70.1%) and 82.6% (123 of 149, 95% CI: 75.2%-88.1%), respectively (Table). GPT-3.5 performed significantly worse on the oral boards question bank, relative to GPT-3.5's performance on a previously reported 500-question written boards question bank (62.4% vs 73.4%, $P = .013$).

FIGURE 1. *Representative indications question bank examination responses by ChatGPT (GPT-3.5) and GPT-4. Images are attributed to ChatGPT (OpenAI; San Francisco, CA) and GPT-4 (OpenAI), which can be found at https://chat.openai.com/.* **A**, *A first-order question on the most common first-line treatment option for a vein of Galen malformation causing heart failure.* **B**, *A higher-order question on expectant management for a severe traumatic brain injury patient with a poor prognosis. Higher-order questions were classified as those involving multiple problem-solving steps to arrive at the correct answer, such as identifying an unknown diagnosis, followed by choosing the best course of management (as opposed to just identifying the diagnosis in a first-order question). Notably, in its answer rationale, ChatGPT demonstrated an ability to declare futility.* **C**, *A higher-order question on management of breast cancer spinal metastases answered incorrectly by both GPT-3.5 and GPT-4. In this scenario, ChatGPT recommended fractionated radiotherapy for a nonradiosensitive lesion causing myelopathy.* **D**, *A higher-order question on the management of presumed vestibular schwannoma in a frail octogenarian answered incorrectly by both GPT-3.5 and GPT-4. CT, computed tomography.*

GPT-4's 82.6% score surpassed GPT-3.5's scores on both the written ($P < .001$) and oral boards ($P = .023$) question banks (Figures 2 and 3). GPT-4 correctly answered all 93 questions that GPT-3.5 did and demonstrated significantly better performance in the Spine subspecialty, specifically (90.5% vs 64.3%, $P = .008$).

By contrast, Bard provided correct answers for only 44.2% (66 of 149, 95% CI: 36.2%-52.6%) of questions, returning incorrect answers for 45.0% (67 of 149) and opting entirely out of answering 10.7% (16/149) of questions. Of the 16 questions, Bard declined to answer 2 questions that involved imaging while

TABLE. Performance of ChatGPT (GPT-3.5), GPT-4, and Google Bard on Oral Board Examination-Focused Question Bank

					P value		
Question category	**Questions**	**GPT-3.5 performance**	**GPT-4 performance**	**Google Bard performance**	**3.5 vs 4**	**3.5 vs Bard**	**4 vs Bard**
Overall	149	93/149 (62.4%)	123/149 (82.6%)	66/149 (44.2%)	<.001	0.002	<0.001
General	33/149 (22.1%)	20/33 (60.6%)	27/33 (81.8%)	11/33 (33.3%)	.102	0.048	<0.001
Functional	8/149 (5.4%)	7/8 (87.5%)	7/8 (87.5%)	5/8 (62.5%)	1.000	0.569	0.569
Peripheral nerve	20/149 (13.4%)	11/20 (55.0%)	15/20 (75.0%)	4/20 (20.0%)	.320	0.048	0.001
Spine	42/149 (28.2%)	27/42 (64.3%)	38/42 (90.5%)	27/42 (64.3%)	.008	1.000	0.008
Tumor	21/149 (14.1%)	11/21 (52.4%)	15/21 (71.4%)	5/21 (23.8%)	.341	0.111	0.004
Vascular	25/149 (16.8%)	17/25 (68.0%)	21/25 (84.0%)	14/25 (56.0%)	.321	0.561	0.062

Performance of ChatGPT on oral board examination-focused question bank. Overall performance and subspecialty breakdown are reported. Differences in performance between LLMs were queried using χ^2 and Fisher exact tests.

14 were solely text-based (Figure 4A-4C). There were no instances of GPT-3.5 or GPT-4 declining to answer a solely text-based question. Both GPT-3.5 (62.4% vs 44.2%, $P = .002$) and GPT-4 (82.6% vs 44.2%, $P < .001$) had superior performance on the Indications Examination, relative to Bard. Bard had significantly lower scores on the General and Peripheral Nerve categories compared with both GPT-3.5 (both $P < .05$) and GPT-4 (both $P < .002$). In addition, GPT-4 alone outperformed Bard on Spine ($P < .008$) and Tumor ($P = .004$) questions. Of the 66 questions answered correctly by Bard, GPT-3.5 and GPT-4 responded correctly on 77.3% (51 of 66) and 97.0% (64 of 66), respectively.

Question Characteristics and LLM Accuracy

Higher-order questions were significantly more represented in the Indications question bank, relative to the prior written boards question bank (85.2% vs 7.4%, $P < .001$). Although higher-order problem solving was predictive of lower question accuracy for GPT-3.5 (odds ratio [OR] = 0.80, $P = .042$) and Bard (OR = 0.76, $P = .014$), this association was not significant for GPT-4 (OR = 0.86, $P = .085$). Notably, GPT-4 was able to answer higher-order questions involving challenging concepts such as declaring medical futility (Figure 1B) but continued to struggle in other scenarios, such as incorporating disease-specific management considerations (Figure 1C) or factoring in patient-level characteristics such as frailty (Figure 1D). The question length did not predict performance by GPT-3.5, GPT-4, or Bard.

Performance on Imaging-Based Questions

Fifty-one questions (34.2%) incorporated imaging into the question stem. Because of multimodal input presently being unavailable for public use, only question text was provided to these models. Both GPT-3.5 and GPT-4 opted to answer 44 of these questions (86.3%) while declining to answer 7 (13.7%) because of insufficient context. By contrast, Bard returned an answer for 96.1% (49 of 51) of all imaging-based questions. GPT-4's performance on imaging-related questions surpassed GPT-3.5's (68.6% vs 47.1%, $P = .044$) and was comparable with Bard's (68.6% vs 66.7%, $P = 1.000$).

On the 44 imaging-related questions attempted by GPT-3.5 and GPT-4, hallucinations were exhibited in 27.3% (12 of 44) and 2.3% (1 of 44) of answers, respectively. Bard demonstrated evidence of hallucinations in 57.1% (28/49) of the questions that it attempted to answer, including 47.1% (8 of 17) of correct answers and 62.5% (20 of 32) of incorrect answers. For example, in questions where a specific area of the image is labeled, Bard would assert which portion of the image was labeled in its answer explanation, without any imaging input or context clues in the question stem text to inform this (Figure 4D-4F). Both GPT-3.5 (27.3% vs 57.1%, $P = .006$) and GPT-4 (2.3% vs 57.1%, $P < .001$) had significantly lower rates of hallucination than Bard. Moreover, GPT-4 had a lower hallucination rate than GPT-3.5 (2.3% vs 27.3%, $P = .002$). A lack of text description for imaging in the question stem was associated with a significantly higher likelihood of hallucination for GPT-3.5 (OR = 1.45, $P = .012$) and Bard (OR = 2.09, $P < .001$), but not GPT-4 (OR = 1.03, $P = .547$).

DISCUSSION

In summary, GPT-4 achieved a score of 82.6% on a question bank of predominantly higher-order diagnostic and management multiple-choice questions designed for oral boards preparation in the field of neurosurgery. This study represents the first of its kind to focus on predominantly higher-order questions in a medical subspecialty domain and the first in neurosurgery to compare the performance of multiple LLMs. GPT-4 demonstrated improved performance in question categories for which GPT-3.5 exhibited

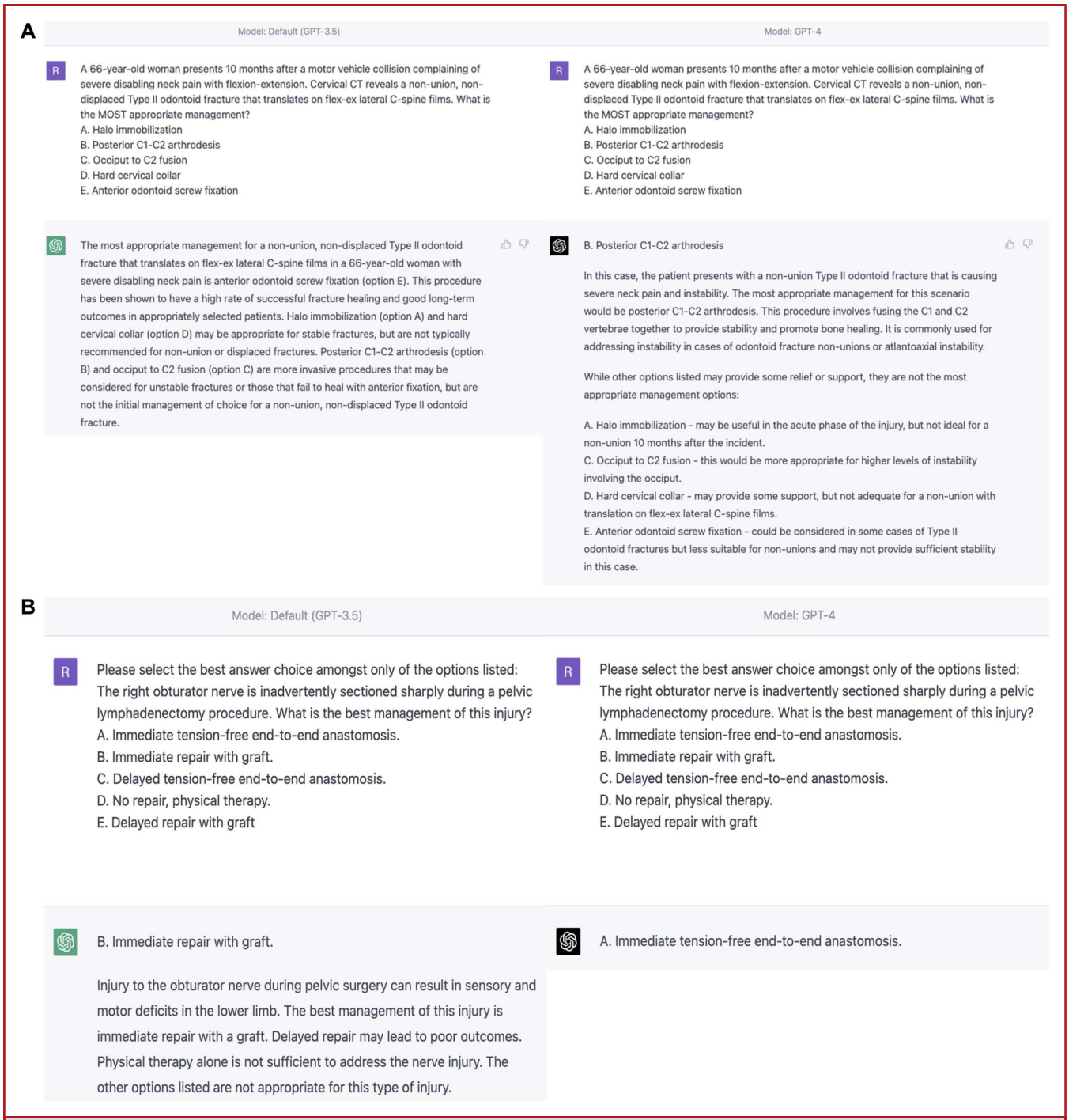

FIGURE 2. *Contrast in question answers Bbtween ChatGPT (GPT-3.5) and GPT-4. Four representative questions across different subspecialties answered incorrectly by ChatGPT (GPT-3.5) but correctly by GPT-4. Images are attributed to ChatGPT (OpenAI) and GPT-4 (OpenAI), which can be found at https://chat.openai.com/.* **A**, *A question on the management of a type II odontoid fracture. In this case, GPT-4 correctly recognized that anterior odontoid screw fixation was a less appropriate treatment modality for a chronic fracture exhibiting nonunion, compared with posterior C1-C2 fusion.* **B**, *A question on the management of an intraoperative transection of the obturator nerve.* **C**, *A question on the management of a severe traumatic brain injury patient with elevated intracranial pressure. In this case, only GPT-4 recognized the patient's hyperthermia and recommended inducing normothermia as the next best step of management.* **D**, *A question on best imaging modality to differentiate cholesteatoma recurrence and encephalocele. CT, computed tomography.*

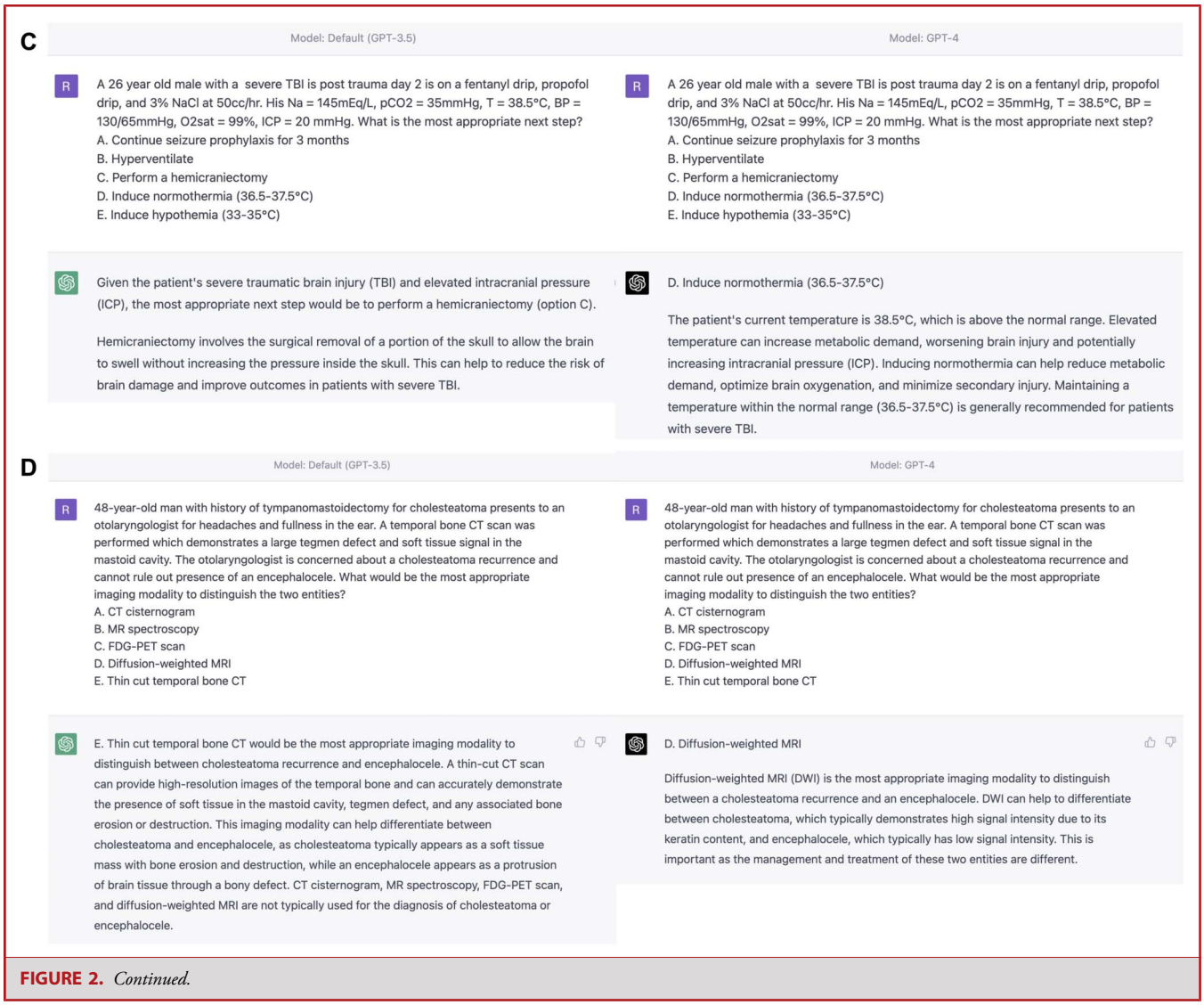

FIGURE 2. *Continued.*

lower accuracy, such as incorporating higher-order problem solving or using context clues alone to answer imaging-related questions. In addition, this study revealed that GPT-4 outperformed Google Bard in all categories, underscoring the critical need for neurosurgeons to stay up-to-date on emerging LLMs and their varying levels of performance for potential application. GPT-3.5's comparatively worse performance on the Indications question bank (62.4%) relative to its written boards results (73.4%) is possibly attributable to the greater representation of higher-order questions in the former.

Because LLMs are increasingly challenged with more open-ended queries, the issues raised by a phenomenon known as "hallucinations" has become more pressing. Hallucinations describe instances in which LLMs assertively incorporate erroneous details into their generated responses. Notably, from a neurological viewpoint, such behavior might more properly be described as "confabulation." This is particularly concerning in the context of neurosurgery, a high-stakes specialty where the consequences of hallucinations could potentially lead to catastrophic mistakes. Therefore, to develop trust in such systems, we must rigorously validate their performance on increasingly higher-order and open-ended scenarios. Recognizing the importance of addressing hallucinations, we developed methods to quantify them, which are essential to further understanding and to ensure that LLMs such as GPT-4 can be safely and effectively integrated into clinical decision-making processes. Our findings demonstrate GPT-4's improved capability to correctly answer higher-order management-focused evaluations of neurosurgical knowledge, reduced rates of hallucination, and an ability to navigate challenging topics such as futility. However, the potential ethical and legal implications of using LLMs in clinical practice must be carefully considered.

There are additionally 2 notable changes to the neurosurgery written and oral boards examination process over the past decade,

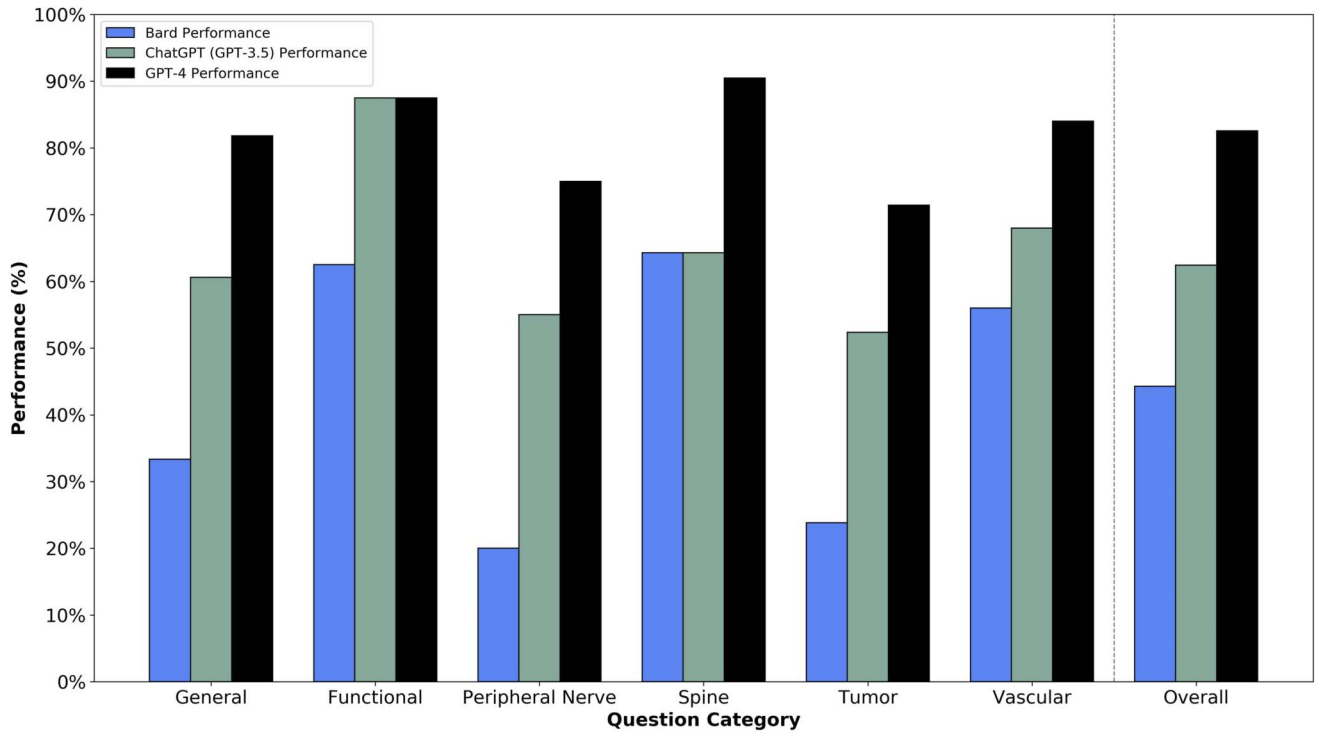

FIGURE 3. *Performance of ChatGPT (GPT-3.5), GPT-4, and Bard on the Indications question bank. Histograms of performance of ChatGPT (GPT-3.5; gray), GPT-4 (black), and Bard (blue) on the Indications question bank, including overall and subspecialty-level performance. Error bars denote 95% confidence interval for scores.*

which warrant discussion in the context of this study's findings. First, the ABNS written (primary) examination has been intentionally modified over the past decade to serve as "mastery" assessments. In close collaboration with the American Association of Neurological Surgeons (AANS), Congress of Neurological Surgeons (CNS), and Society of Neurological Surgeons (SNS) considerable resources have been devoted to ensuring that the content tested on these examinations is as transparent and accessible as possible. This likely influences the availability of written examination content on the Internet, thereby providing an excellent foundation for training LLMs). Similarly, it is also worth noting that the knowledge of oral examination content and commonly tested management concepts accessible online, even in a multiple-choice form such as the SANS Indications Examination, may also enhance the training data available to LLMs. Nevertheless, it is conceivable that the successful performance of LLMs on these more open-access multiple-choice assessments may not fully translate to comparable performance when confronted with more unpredictable and unique real-world situations, which is more reflective of the oral boards testing environment and actual clinical practice. This potential gap in application may potentially already be appreciated by the poorer performance of GPT-3.5 and Bard on higher-order questions.

Second, the oral boards have undergone reorganization into a 3-session format, including 2 sessions with standardized clinical scenarios with a wide range of acceptable responses.

However, the third session is based on the candidate's own cases as a practicing neurosurgeon, and despite the candidate being the definitive authority on the knowledge of how these cases were managed, knowledge gaps and practice concerns can lead to a failing grade on an individual case. This paradox demonstrates the importance of distinguishing possession of knowledge from application of knowledge, especially for complicated cases with individual-level considerations and significant equipoise, which compose a significant portion of neurosurgical practice. When considering the optimal approach for certification of future neurosurgeons, the utility of multiple-choice examinations, which can now be passed by LLMs, warrants further assessment. Although performance on these tests serves as an indicator of possessing foundational knowledge, the oral boards examination highlights the significance of thoroughly probing broader management decisions in an open-ended and verbal format to determine whether that knowledge can be applied appropriately, safely, and compassionately. AI programs will prove to be valuable resources, supplying clinicians with rapidly accessible and reliable information. However, it is the responsibility of the clinician to integrate these data with the unique circumstances of each patient. In other words, although AI algorithms may exhibit remarkable knowledge capabilities, it is ultimately the clinician who must exercise wisdom.

In summation, it can be argued that multiple-choice examinations, even those consisting primarily of higher-order questions,

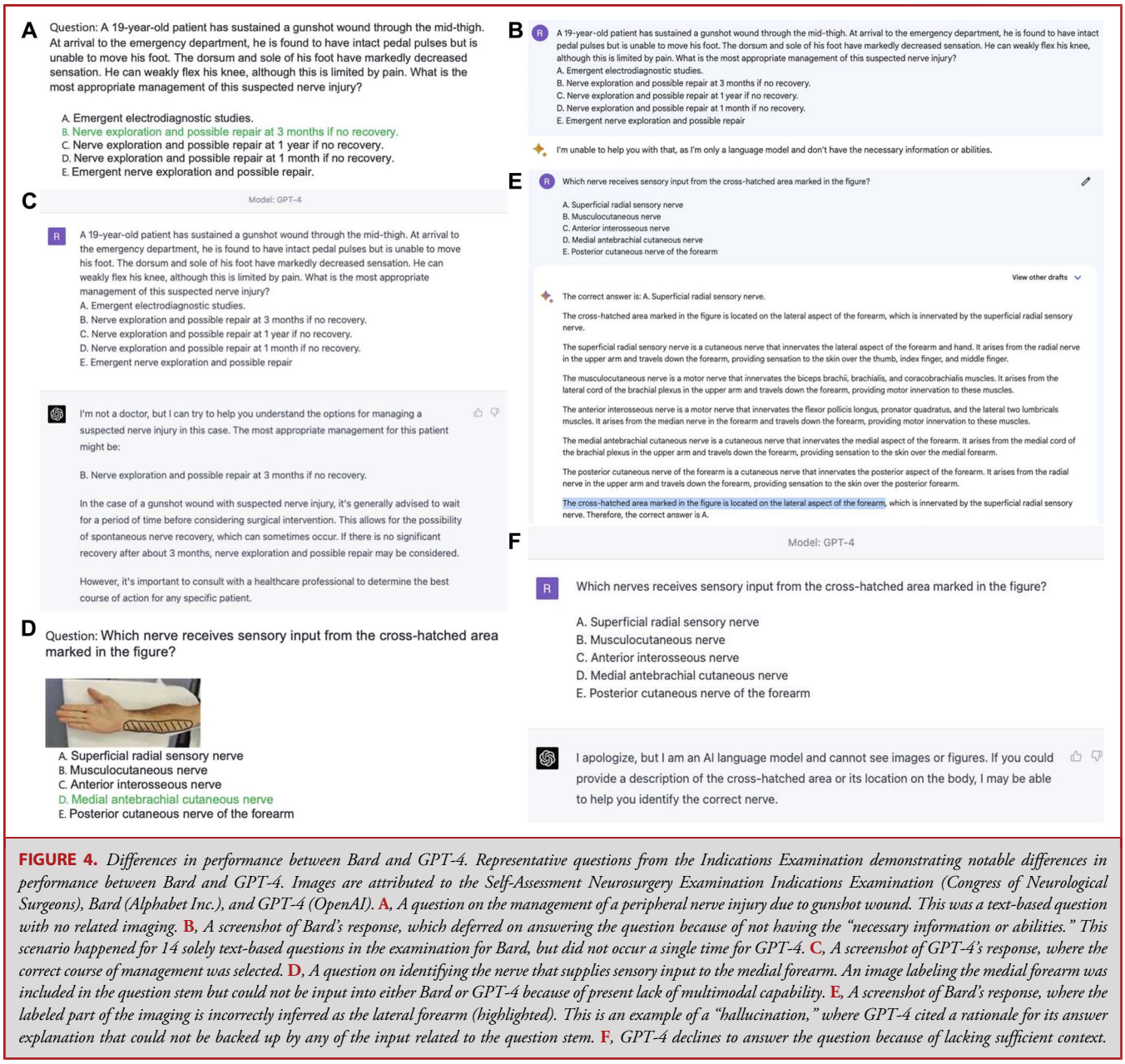

FIGURE 4. *Differences in performance between Bard and GPT-4. Representative questions from the Indications Examination demonstrating notable differences in performance between Bard and GPT-4. Images are attributed to the Self-Assessment Neurosurgery Examination Indications Examination (Congress of Neurological Surgeons), Bard (Alphabet Inc.), and GPT-4 (OpenAI).* **A**, *A question on the management of a peripheral nerve injury due to gunshot wound. This was a text-based question with no related imaging.* **B**, *A screenshot of Bard's response, which deferred on answering the question because of not having the "necessary information or abilities." This scenario happened for 14 solely text-based questions in the examination for Bard, but did not occur a single time for GPT-4.* **C**, *A screenshot of GPT-4's response, where the correct course of management was selected.* **D**, *A question on identifying the nerve that supplies sensory input to the medial forearm. An image labeling the medial forearm was included in the question stem but could not be input into either Bard or GPT-4 because of present lack of multimodal capability.* **E**, *A screenshot of Bard's response, where the labeled part of the imaging is incorrectly inferred as the lateral forearm (highlighted). This is an example of a "hallucination," where GPT-4 cited a rationale for its answer explanation that could not be backed up by any of the input related to the question stem.* **F**, *GPT-4 declines to answer the question because of lacking sufficient context.*

provide only a superficial assessment of a neurosurgeon's expertise in patient management, with limited representation of a neurosurgeon's intuition and decision-making. Accordingly, oral board examination failures are frequently attributed to inappropriate surgical indications and subtle errors in judgment, rather than a lack of factual knowledge. Therefore, it is essential to further evaluate the performance of LLMs in this domain, which will be the subject of future studies. As AI continues to advance, multiple-choice examinations may assume a less prominent role in medical education, with oral boards-style verbal assessments gaining increased importance. Another change that AI may bring to neurosurgical education is the use of LLMs by neurosurgical trainees for boards preparation. For example, with the initial input of a clinical scenario to discuss, an LLM such as GPT-4 may be used as a conversational aid to rehearse the discussion of more challenging topics for the boards or even appreciate new clinical insights or rationales from the responses generated by LLMs.

Although this multiple-choice question bank cannot fully capture the dynamic, conversation-based, and open-ended nature of the oral boards, our findings do hint at the potential value of LLMs such as GPT-4 in neurosurgical education and in clinical decision-making. Given a score improvement of more than 20% between 2 AI models

released just 4 months apart, it is critical for neurosurgeons to stay informed and up-to-date about these rapidly evolving tools and their potential applications to clinical practice. Toward this end, the development of methods to quantify and understand hallucinations as well as the validation of LLMs on higher-order and open-ended scenarios is vital for the successful integration of these tools into neurosurgery and other high-stakes medical specialties. Ultimately, the capacity and extent to which LLMs are incorporated into practice will depend heavily on the ability to minimize and recognize hallucinations. In summary, this study represents an important initial benchmark in evaluating LLM performance in higher-order and relatively more open-ended clinical scenarios.

Limitations

This study has several potential imitations. First, as discussed earlier, the use of multiple-choice questions to quantify LLM knowledge for higher-order neurosurgical topics incompletely captures the open-ended nature of the true neurosurgery oral board examination. We aim to conduct more open-ended assessments of LLM neurosurgical knowledge in future assessments. Second, it is possible that certain question characteristics, such as incorporation of higher-order problem-solving, may have been characterized incorrectly or differently by a separate evaluator. However, this study used 3 authors to collect data on question characteristics, and differences between evaluators was minimal (<5%). Third, the accuracy of LLMs when incorporating imaging-related data could not be assessed because of these functions not being presently being publicly available, and LLM performance on imaging requests after multimodal input will be the subject of future studies. Fourth, owing to continual "under-the-hood" improvements to LLMs influenced by factors such as aggregate user input, the performance of LLMs, such as GPT-4, may change gradually and returned answers may differ from how they are presented in this study. To minimize heterogeneity due to these factors, data collection was performed in a 24-hour range for each separate LLM.

CONCLUSION

On a question bank of predominantly higher-order management case scenarios intended for neurosurgery oral boards preparation, GPT-4 achieved a score of 82.6%, outperforming GPT-3.5 and Google Bard. Unlike GPT-3.5 and Bard, higher-order problem solving was not predictive of lower answer accuracy from GPT-4. Finally, GPT-4 exhibited significantly lower rates of hallucinations on imaging-related questions.

Funding

This study did not receive any funding or financial support.

Disclosures

The authors have no personal, financial, or institutional interest in any of the drugs, materials, or devices described in this article.

REFERENCES

1. 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(2):e0000198.
2. Choi JH, Hickman KE, Monahan A, Schwarcz D. ChatGPT goes to law school. *SSRN Electron J*. 2023;23(3).
3. Terwiesch C. Would chat GPT3 get a Wharton MBA? *A Prediction Based on Its Performance in the Operations Management Course* Philadelphia, PA: University of Pennsylvania;2023.
4. Ali R, Tang OY, Connolly I, et al. Letter: Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *medRxiv*. Published online ahead of print January 23, 2023. doi: http://dx.doi.org/10.2139/ssrn.4335905.
5. OpenAI. GPT-4 Technical Report. 2023. Accessed March 14, 2023. https://cdn.openai.com/papers/gpt-4.pdf
6. Nori H, King N, McKinney SM, Carignan D, Horvitz E. *Capabilities of GPT-4 on Medical Challenge Problems*. 2023. Accessed March 20, 2023. https://www.microsoft.com/en-us/research/publication/capabilities-of-gpt-4-on-medical-challenge-problems/
7. Wang MC, Boop FA, Kondziolka D, et al. Continuous improvement in patient safety and quality in neurological surgery: the American Board of Neurological Surgery in the past, present, and future. *J Neurosurg*. 2021;135(2):637-643.
8. The American Board of Neurological Surgery. Frequently asked questions. 2023. Accessed February 28, 2023. https://abns.org/frequently-asked-questions/#faq-general-27b
9. Moran S. *How to Prepare for the USMLE® Step 1. 2020*. 2023. Accessed February 28, 2023. https://blog.amboss.com/us/how-to-prepare-for-the-usmle-step-1

Acknowledgments

We would like to acknowledge and thank the Congress of Neurological Surgeons for their development and dissemination of the mock examination questions used for this study.

COMMENTS

In this timely and interesting Letter, the authors demonstrate the relative differences in performance of 3 LLMs (ChatGPT, GPT-4, and Google Bard) on answering questions from a multiple-choice question bank designed specifically for neurosurgery oral boards examination preparation. Of note is the incremental performance improvement to 82.6% correct achieved by GPT-4 released in March, 2023, outperforming ChatGPT released in November, 2022. These results parallel the demonstrations of AI achieving passing performance on certification exams across a variety of fields. However, there is recognition of its limitations manifested as "hallucinations," which are increasingly associated with open-ended questions.

The ABNS Oral Exam has long served as a momentous final milestone towards ABNS Board Certification, aiming to assess how well candidates demonstrate clinical reasoning, but has evolved in structure to incorporate standardized cases and candidates' own patients. Thus, performance is not based simply on knowing the answers to the test questions, but rather on competent discussion of patient management with demonstration of appropriate clinical judgement and ethics.

The authors' findings confirm what most of us realized in medical school—that super achieving test scorers do not necessarily equate to the best clinicians—likely due to the requisite "higher-order problem solving," incorporation of patient characteristics, and disease-specific management considerations—otherwise known as experience and judgment.

Judy Huang
Baltimore, Maryland, USA