1-1-2023

Evaluating the Current Ability of ChatGPT to Assist in Professional Otolaryngology Education.

Habib G. Zalzal
Children's National Hospital

Jenhao Cheng
Children's National Hospital

Rahul K Shah
Children's National Hospital

Original Research

AMERICAN ACADEMY OF OTOLARYNGOLOGY–HEAD AND NECK SURGERY FOUNDATION

OTO Open
2023, Vol. 7(4):e94
© 2023 The Authors. OTO Open published by Wiley Periodicals LLC on behalf of American Academy of Otolaryngology–Head and Neck Surgery Foundation.
DOI: 10.1002/oto2.94
http://oto-open.org
WILEY

Evaluating the Current Ability of ChatGPT to Assist in Professional Otolaryngology Education

Habib G. Zalzal, MD[1], Jenhao Cheng, PhD[2], and Rahul K. Shah, MD[1]

Abstract

Objective. To quantify ChatGPT's concordance with expert Otolaryngologists when posed with high-level questions that require blending rote memorization and critical thinking.

Study Design. Cross-sectional survey.

Setting. OpenAI's ChatGPT-3.5 Platform.

Methods. Two board-certified otolaryngologists (HZ, RS) input 2 sets of 30 text-based questions (open-ended and single-answer multiple-choice) into the ChatGPT-3.5 model. Responses were rated on a scale (correct, partially correct, incorrect) by each Otolaryngologist working simultaneously with the AI model. Interrater agreement percentage was based on binomial distribution for calculating the 95% confidence intervals and performing significance tests. Statistical significance was defined as $P < .05$ for 2-sided tests.

Results. In testing open-ended questions, the ChatGPT model had 56.7% of initially answering questions with complete accuracy, and 86.7% chance of answer with some accuracy (corrected agreement = 80.1%; $P < .001$). For repeat questions, ChatGPT improved to 73.3% with complete accuracy and 96.7% with some accuracy (corrected agreement = 88.8%; $P < .001$). For multiple-choice questions, the ChatGPT model performed substantially worse (43.3% correct).

Conclusion. ChatGPT currently does not provide reliably accurate responses to sophisticated questions in Otolaryngology. Professional societies must be aware of the potential of this tool and prevent unscrupulous use during test-taking situations and consider guidelines for clinical scenarios. Expert clinical oversight is still necessary for myriad use cases (eg, hallucination).

Keywords

artificial intelligence, ChatGPT, continuing medical education, large language model, machine learning, maintenance of certification, OpenAI

Received August 28, 2023; accepted November 4, 2023.

The lay public, medical community, and academicians are enraptured with the explosion of artificial intelligence (AI) and the potential applications to their respective workflows and industries. In November 2022, the Microsoft Corporation® introduced Chat Generative Pre-trained Transformer (ChatGPT) as an AI tool for public use. ChatGPT is based on the concept of large language models (LLM)—deep neural network models trained on vast amounts of publicly available data, natural language understanding, and generation. Journal articles, textbook chapters, and question banks are some of the sources utilized by ChatGPT for its training in academic medicine. AI uses pattern-recognition of its training data to generate new responses, rather than searching from a knowledge bank, which leads to a wide variation of answers even within a single session. As physician-scientists, it is likely that some of our own original works prior to 2021 have been used to develop LLM; we are certain that in the future all academic works will be potentially used in developing and refining LLMs.

Since the public unveiling of ChatGPT, the intersection of AI and medicine has become the topic of much conversation. This manuscript focuses on the professional uses of ChatGPT in healthcare, specifically as it pertains to physician certification and board-style questioning. The use cases for LLM are myriad and cover the spectrum of basic medical knowledge regurgitation to intricate explanations in response to patient queries. There are pilot projects and large-scale evaluations of

[1] Division of Otolaryngology–Head and Neck Surgery, Children's National Hospital, Washington, District of Columbia, USA
[2] Quality, Safety, Analytics, Children's National Hospital, Washington, District of Columbia, USA

Abstract is to be presented at the ENT Annual Meeting this October 2023 in Nashville, Tennessee

Corresponding Author:
Habib G. Zalzal, MD, Assistant Professor, Division of Otolaryngology, Children's National Medical Center, 111 Michigan Avenue, NW Suite 3W-800, Washington, DC 20010, USA.
Email: hzalzal@cnmc.org

the use of artificial intelligence tools and platforms to help answer patient inquiries, review records, interpret images, guide procedural planning, and so forth. However, there exists legitimate concern that the current AI tools are not refined nor have been validated by professionals, which can lead to medical misinformation.

Medical education represents an important use case where adoption, learning, and spread (generalizability) intersect rapidly. Huh et al suggest that it is not a question of if ChatGPT can contribute to medical education, rather how and when.[1] The ability of ChatGPT to pass the United States Medical Licensing Examination (USLME) created much discussion in the medical community.[2–4] While the AI model performed adequately in certain areas, Huh demonstrated difficulty for ChatGPT on a subject-specific examination, parasitology.[1] Mbakwe attributes this to a "lack of domain-specific training."[4] Mbakwe notes that passing USMLE is based on rote memorization, while ChatGPT needs to demonstrate its ability to use human-like reasoning and thought processes.[4]

The current study attempts to explore ChatGPTs critical thinking at the level of a board-certified Otolaryngologist. The training model of ChatGPT is beyond the scope of this manuscript. However, there remains a need to evaluate ChatGPT—specifically, its ability to assist in the education of post-residency trained and board-certified Otolaryngologists. This current evaluation aims to test 2 hypotheses regarding ChatGPT's ability to utilize LLM for medical education.

Hypothesis 1. the null hypothesis is that ChatGPT cannot be used as an accompanying education tool for attending otolaryngologists when encountering clinical scenarios.

Hypothesis 2. the null hypothesis is that ChatGPT cannot answer open-ended and single-answer multiple-choice based board-level Otolaryngology questions.

We aim to quantify ChatGPT's concordance with expert Otolaryngologists when posed with high-level questions that require blending rote memorization and critical thinking. For ChatGPT to play a significant role in medical education, this AI tool needs to be validated in an expanded fashion past rote recall to critical reasoning and thinking.[2]

Methods

The current report is deemed exempt by Children's National Hospitals institutional review board.

Question Selection

In creating a set of questions, we utilized open-ended questions slightly modified from publicly available, previously validated online databanks as reasonable questions an Otolaryngologist should be able to answer. This was done to cover a diverse range of topics within the field of Otolaryngology: scientific knowledge and applications, critical thinking, problem solving, reading comprehension, and logical reasoning. In creating questions, we used 5 representative questions from each of the following 6 categories: surgical anatomy, otology, head and neck/malignancy, airway/voice, rhinology, and fundamentals. If questions were incorrectly answered by the AI model or had low rater reliability, these questions were revisited at the end of the question period in the same ChatGPT session. Questions were then resubmitted to the AI model to determine if the model changed its answer compared to previous initial inquiry, and these results were also recorded.

For the second test, a different set of thirty single-answer multiple-choice-based questions were randomly selected by the authors as representative of a requisite knowledge base.[5] Answers to these multiple-choice questions were known prior to involving the AI model, such that we could assess the accuracy of the AI answers to the conventional answer. ChatGPT does not answer question simply by choosing the right multiple-choice question; rather it provides a narrative. Questions were graded based on whether the AI model selected the correct multiple-choice answer, regardless of whether the model understood the question or explained the correct answer despite selecting an incorrect multiple-choice answer. The authors (HZ, RS) were then responsible for interpreting the answers for both tests.

To replicate the test-taking environments for the second test, we exclusively selected single-answer multiple-choice questions. Multiple-choice questions were incorporated by copying and pasting text-only content directly into the AI model, and the model would explain the answer followed by the reasoning behind this answer with each question. Images, radiographs, audiograms, and other visually-based items were excluded for this trial.

In selecting our AI model, the legacy GPT-3.5 model of ChatGPT (May 24 version; OpenAI) was accessed for this study as it is freely accessible to the public.[6] Each question was asked in a serial manner. The ChatGPT session was not reset between questions. This methodology was utilized in both testing scenarios (open-ended and single-answer multiple-choice).

Inter-Rater Reliability (IRR)

Each question was independently evaluated by authors HZ and RS ("response raters") simultaneously in the same ChatGPT session. Response raters were given 60 seconds to evaluate the answer for accuracy using the following options: (1) Agree with response, (2) Partially correct response, and (3) Response inaccurate. Both rater responses were utilized to create a scoring system (the sum of the response raters score) as to whether the AI model was completely accurate in its answer description (a score of 2) versus completely inaccurate (a score of 6). IRR was then assessed by looking at the consistency in responses between the raters, with an overall aim to use both "consistent" and "good" responses to further evaluate the accuracy. Agreement percentage was first used to check whether 2 raters are consistent with each other and an option to weigh in half point for partial consistent responses such as 1 versus 2 and 2 versus 3.

Statistical Analysis

Data were managed using institutional cloud services. Statistical evaluation was performed using R Statistical Software® version 4.3.0 (R Foundation for Statistical Computing 2023) & Microsoft Excel 365® (Microsoft Corporation 2023). We analyzed the inter-rater agreement percentage based on binomial distribution for calculating the 95% confidence intervals and performing significance tests. Statistical significance was defined as $P < .05$ for 2-sided tests. Gwet's AC2 coefficient was used to evaluate the agreement level between 2 raters on categorical (or ordinal) ratings, but with the agreement portion due to chance corrected.[7] The most widely used threshold to consider a very strong agreement level is above 0.8.[8,9]

Results

Test Environment #1: Open-Ended Questioning

ChatGPT was queried at 10:00 am EST on June 14, 2023 with the questions listed in **Table 1** alongside the accuracy scores for each rater. **Table 2** shows the thirteen questions that were revisited after initially inaccurate responses by the AI model (per one or both response raters).

Looking at ChatGPT accuracy in **Table 3**, 57% (N = 17/30) of questions were initially answered with complete accuracy when assessed by both raters while 13.3% (N = 4/30) had concern for a completely inaccurate answer, meaning that the AI model answered questions with some accuracy about 87% of the time. For IRR in **Table 4**, there were no "totally unmatched" questions (a score of 1 combined with a score of 3), leading to weighted agreement percentage of 88.33%. When corrected for chance, the Gwet AC2 coefficient was 80.14% ($P < .01$), which means a very strong agreement level by both board-certified Otolaryngologists in the answer given by ChatGPT.

For questions revisited after initially low score totals, 38% (N = 5/13) of questions were answered with complete accuracy, while partial accuracy improved or maintained in a total of 92% of questions (N = 12/13). Only one question (question #28) was not answered accurately after both instances. Looking at IRR for revisited questions alongside initially accurate questions (**Table 4**), weighted agreement percentage improved to 91.67%. Gwet AC2 coefficient was 88.77%, once again depicting a strong agreement level between both raters when evaluating ChatGPT responses.

A subsequent test to investigate ChatGPTy's certainty of answers was then conducted. The authors evaluated how the LLM would respond to being asked the same question several times, and whether its answer would be consistent each time. Both authors decided that our most clinically intricate question was #2 from **Table 1**: "Does one perform auricular reconstruction in a patient with absent ossicles and no middle ear space or mastoid pneumatization?" We submitted this query 7 times to the AI system, and found ChatGPT responded with a variation of the correct answer each time (IRR agreement of 100%). After the 6th query, the answer was more definitive and direct rather than explanatory, showing that the system was certain of its response. On the 7th query, the answer response was reworded compared to trials 1 to 6, but the overall answer remained the same and correct.

Test Environment #2: Multiple-Choice Questioning

Our next test was to evaluate AI answers to single-answer multiple-choice questions taken from an unpublished question bank. Overall, accuracy of the ChatGPT system in responding to these questions was 43.33% (N = 17). When inaccurate questions were revisited, the AI model continued to select the wrong answer in the multiple-choice format. However, both raters noted the accompanying explanation had some partially correct information despite selection of the incorrect choice by the system.

Discussion

There remains great excitement about the use of LLM in healthcare, and models such as ChatGPT are leading the way showing the potential use-cases for both the clinical and educational aspects of medicine.[10,11] In our research into the role of the AI model as an accompanying educational tool for attending Otolaryngologists, we can reject our first null hypothesis, as the system can be used to help in aiding discussion within our field. For our second hypothesis, we potentially can accept the null hypothesis although LLM correctly can answer open-ended questions in the field of Otolaryngology. There is significant potential in AI answering multiple-choice questions, but at this time, it does not fare as successfully in this task as it does for open-ended questioning.

ChatGPT was nuanced in answering certain questions from our cohort. Specifically, it was able to understand that questions were related to the field of Otolaryngology, catching the subtlety of ossicle versus auricular that is presented in **Table 1**, question #2. While reading comprehension is a common component of human error when taking knowledge-based exams, it appears the LLM deciphers and interprets intricacies in the way a question is asked better than the rater at times.

Table 1. Questions Posed to Chat-GPT and Score by Reach "Reliability Rater (RR)"

	Initial AI inquiry			
Question	RR1	RR2	Total Score (out of 6)	Interrater reliability
1. What is the surgery for dysphagia lusoria?	2	1	3	0.5
2. Does one perform auricular reconstruction in a patient with absent ossicles and no middle ear space or mastoid pneumatization?	1	1	2	1
3. What nerve is injured during thyroidectomy that reduces supraglottic sensation?	1	1	2	1
4. Where is the facial nerve relative to a Workman type 1 branchial cyst?	2	2	4	1
5. What muscle is trapped in an orbital blowout fracture?	3	3	6	1
6. What is the first line treatment for sudden sensorineural hearing loss?	1	1	2	1
7. When does one perform an ENOG during management of facial nerve paralysis in a temporal bone fracture?	3	3	6	1
8. What is the treatment for auditory neuropathy?	1	1	2	1
9. What is the most important predictor of cochlear implant performance relative to hearing improvement?	1	2	3	0.5
10. What is the next best step in diagnosis of vertigo following a normal audiogram?	1	1	2	1
11. In the management of Merkel cell carcinoma, when do you perform radiation therapy after wide local excision?	1	2	3	0.5
12. What are the indications for central neck dissection in management of papillary thyroid carcinoma?	1	1	2	1
13. What is the management of juvenile nasopharyngeal angiofibroma with orbital and intracranial invasion?	1	1	2	1
14. Where does a recurrent thyroglossal duct cyst most likely occur?	2	2	4	1
15. When is surgery preferable to chemoradiation in treatment of squamous cell carcinoma of the tonsil?	2	1	3	0.5
16. What is the best injection therapy for recurrent respiratory papilloma of the larynx?	2	1	3	0.5
17. What is the narrowest point of the airway in a 1-month-old presenting with sleep apnea?	2	1	3	0.5
18. What is the management of velopharyngeal insufficiency after adenoidectomy?	1	1	2	1
19. Why would an EMG be normal of the interarytenoid muscle after vagus nerve sacrifice?	2	2	4	1
20. How does the vocal quality differ between spasmodic dysphonia and muscle tension dysphonia?	1	1	2	1
21. What is the best imaging test for frontal osteoma?	1	1	2	1
22. What is the treatment of a mycetoma in the nasal cavity?	1	1	2	1
23. What is the next best step for managing a CSF leak during sinus surgery?	1	1	2	1
24. How do I treat a child with a subperiosteal abscess of <0.5 cm and has not received antibiotics?	1	1	2	1
25. How do I manage a dehiscent carotid artery during sphenoid surgery?	1	1	2	1
26. What is the management for anaphylaxis to allergy sera in the office?	1	1	2	1
27. What is the first thing to do in an operating room fire while operating in the airway?	2	3	5	0.5
28. What is the sequence for donning and doffing personal protective equipment?	3	3	6	1
29. What is the standard for sterilizing a fiberoptic laryngoscope?	1	1	2	1
30. What is the first thing to do if a patient suffers a vasovagal episode in the office?	1	1	2	1

Each rater graded the AI response with the following scale: 1 = full agreement, 2 = partial agreement, 3 = disagree.

Another observation discovered during this process was the hallucinating, or confabulating, answers by the LLM. Hallucination is a term used in AI that refers to the situation where there is a confident response by an AI model that does not seem to be justified by its training data. This event occurred in 13.33% (N = 4) of open-ended questions (Supplemental S1, available online). The responses provided by the model appeared correct, but not direct enough to fully answer the question. This observation is an important factor and can be appreciated in the change in accuracy between Tables 1–3. Specifically, the AI model directly answered 43.3% of multiple-choice questions correctly versus 56.6% of open-ended questions.

Our methodology is similar to the subjective questions posed by four obstetric and gynecologic physicians to ChatGPT.[12] Their questions were a mix of value to the scientific community and public-facing audiences. Further, a recent publication by Hoch et al looked into ChatGPT's quiz skills in multiple subspecialty domains within Otolaryngology.[11] Their research into 2576 multiple-choice questions (2097 single-answer and 479 multiple-answer) found that the system does significantly better in the single-answer environment (57% correct)

Table 2. Questions Posed to Chat-GPT and Score by Reach "Reliability Rater (RR)" With Inclusion of Questions That Required Repeat Inquiry, With Subsequent RR

	Initial AI inquiry			Repeat AI inquiry		Interrater
Question	RR1	RR2	Improvement status	RR1	RR2	reliability
1. What is the surgery for dysphagia lusoria?	2	1	Improved on repeat inquiry	1	1	1
4. Where is the facial nerve relative to a Workman type 1 branchial cyst?	2	2	Improved on repeat inquiry	1	1	1
5. What muscle is trapped in an orbital blowout fracture?	3	3	Improved on repeat inquiry	1	1	1
7. When does one perform an ENOG during management of facial nerve paralysis in a temporal bone fracture?	3	3	Improved on repeat inquiry; discussed ENOG should not be done right after injury	2	1	0.5
9. What is the most important predictor of cochlear implant performance relative to hearing improvement?	1	2	Improved on repeat inquiry	1	1	1
11. In the management of Merkel cell carcinoma, when do you perform radiation therapy after wide local excision?	1	2	No improvement; shorter response & more generic	2	2	1
14. Where does a recurrent thyroglossal duct cyst most likely occur?	2	2	No improvement; similar answer	2	2	1
15. When is surgery preferable to chemoradiation in treatment of squamous cell carcinoma of the tonsil?	2	1	No improvement; answer did not discuss HPV status	2	1	0.5
16. What is the best injection therapy for recurrent respiratory papilloma of the larynx?	2	1	No improvement; answer did not discuss Avastin among options/somewhat outdated	2	1	0.5
17. What is the narrowest point of the airway in a 1-month-old presenting with sleep apnea?	2	1	Improved on repeat inquiry; mentioned cricoid specifically	1	1	1
19. Why would an EMG be normal of the interarytenoid muscle after vagus nerve sacrifice?	2	2	Improved on repeat inquiry; did not explicitly mention bilateral innervation	2	1	0.5
27. What is the first thing to do in an operating room fire while operating in the airway?	2	3	Improved on repeat inquiry; AI rearranged previous answer to emphasize correct one	2	1	0.5
28. What is the sequence for donning and doffing personal protective equipment?	3	3	No improvement; sequence was still incorrect	3	3	1

Table 3. Total Score Summary by Each Score Category (Score 2-6), With a Lower Score Having Better Accuracy

		Initial ratings			Repeated ratings		
Total score	Accuracy level	N	% Share	% Cumulative	N	% Share	% Cumulative
Score 2	Best	17	56.7%	56.7%	22	73.3%	73.3%
Score 3	Moderate	6	20.0%	76.7%	5	16.7%	90.0%
Score 4	Some	3	10.0%	86.7%	2	6.7%	96.7%
Score 5	Least	1	3.3%	90.0%	0	0.0%	96.7%
Score 6	None	3	10.0%	100.0%	1	3.3%	100.0%

% Share is obtained by dividing # Questions by 30 (total number of questions).

Table 4. Interrater Reliability Results in Contingency Table

Initial ratings				Repeated ratings			
	Rater 2				Rater 2		
Rater 1	1	2	3	Rater 1	1	2	3
1	17	2	0	**1**	22	0	0
2	4	3	1	**2**	5	2	0
3	0	0	3	**3**	0	0	1
Observed agreement = 76.7% (60.9%–92.5%, *P* < .001)				Observed agreement = 83.3% (69.5%–97.2%, *P* < .001)			
Weighted agreement = 88.3% (80.4%–96.2%, *P* < .001)				Weighted agreement = 91.7% (84.7%–98.6%, *P* < .001)			
Corrected agreement = 80.1% (64.9%–95.4%, *P* < .001)				Corrected agreement = 88.8% (78.3%–99.2%, *P* < .001)			

Agreement pairs are diagonal and disagreement pairs are off diagonal. Percentages with parentheses are 95% confidence intervals.

versus the multiple-answer questions (34% correct). This differs from our experience into the matter, albeit on a smaller scale, where ChatGPT in our single-answer multiple-choice scenario only answered 43.3% of questions correct. A strength of our study in comparison is the board-certified Otolaryngologist review of AI responses. It is in our belief that if given enough time to utilize deep learning techniques on data beyond 2021 (the time of ChatGPTs training set), that ChatGPT will improve in its ability to answer multiple-choice questions. However, this deserves further investigation as we studied only single-answer questions. Nevertheless, both our study and Hoch et al demonstrated ChatGPT accuracy in the range of 43% to 57%; hence equivocal acceptance of the second hypothesis of this study.[11]

The ability of the AI model to posit new, sophisticated answers was also observed, such as question #15 on squamous cell carcinoma (**Table 1**), where an answer is not directly discernable. For these questions, the response by the AI model is subjectively much slower, meaning the system is likely trying to interpret the question in real-time, is querying its training set, and constructing a logical answer; this led to repetitive and rambling responses at times. In other situations that required simpler responses, the system referenced direct literature, such as the American Thyroid Association guidelines for management of papillary thyroid carcinoma (question #12). Grunebaum et al admittedly noted the lag between the training of the model and the output in their study, and explained this is no different than current clinical practice guidelines and the incorporation of recent evidence and literature.[12] Without oversight by Otolaryngologists, the element of hallucination can lead to adverse intentions if not evaluated from a patient care standpoint with regard to reality and accuracy.

For the vast majority of questions, the LLM recommended discussion with an Otolaryngologist. This was important for both raters from a clinical and public health perspective in the event a patient were to use the system for an otolaryngology-specific condition. The model recognizes which physician one should see in response to the queries. A recent publication by Park et al discussed the potential uses of ChatGPT in a clinical role in our field.[10] Patient education, clinical decision support, and literature summarization are important applications of LLM. Our impression based upon our experience with ChatGPT answering open-ended questions is that AI models have the capacity, with further data training and time, to be potentially of value to board-certified Otolaryngologists.

Limitations

There are several limitations to using ChatGPT as a test-taking adjunct when approaching educational-based examinations. ChatGPT is only accurate as of information through 2021, however, this model has the potential to learn far beyond what the human mind is capable of through its ability to synthesize limitless text and online sources. The model has the ability to learn new facts about medicine, as demonstrated by Kung et al in that ChatGPT improved its score on the United States Medical Licensing Exam (USMLE) over a time-period.[3] Additionally, while the current variation of ChatGPT does not provide direct references or resources for its answer choices, other forms of LLM and AI models across the internet do have this ability.[12] As such, we implore medical education to rapidly pivot, accelerate understanding and adoption of AI into discussions, planning, and learning.

A limitation of our study methodology is the generalizability of the questions used to evaluate the LLM (**Table 1**). Questions were constructed by academic board-certified otolaryngologists without external validation. If the questions were constructed by a panel of experts for validating purposes, would ChatGPT have yielded different answers? Future studies could consider partnering with organizations skilled in test question creation. For this research, the authors intent was to mirror how practicing otolaryngologists would query ChatGPT, not to aim for near-perfect test construction.

An additional limitation of the LLM is that answer responses are not always reproducible, which affects the internal validity of our study. This was made evident in **Table 2**, where initially incorrect responses were changed by ChatGPT when alternative information was learned while responding to the inquiry. We attempted to limit this bias by viewing the responses during a single session of ChatGPT use, but if this question were asked during a different session on another account, it is unclear if the same response would be received. Another surprising finding was ChatGPT was not able to forget a prior question and its ability to answer sequential medical questions deteriorated over the sequence of inquiries. Hoch et al identified this limitation as well, suggesting that session clearance before each question would significantly impact the accuracy of the responses provided.[11] For open-ended questions, however, this limitation was not consistently noted, likely because the system was not limited in its selection of an answer and would synthesize prior information to respond if necessary. While the model continues to evolve and will likely improve its ability in answering such style of questions, ChatGPT remains inaccurate enough currently to not consistently aid in multiple-choice-based examinations.

Conclusion

The authors consider ChatGPT to be a novel, innovative tool in the educational toolbox for physicians that completed their training, but not a panacea. LLM must be contextualized to be optimized for test-taking situations and during patient care scenarios (open-ended queries). The current ChatGPT as studied in this evaluation has an open-ended accuracy rate of 56.7% for initially answering questions and an 86.7% chance of an answer with some accuracy. For multiple-choice questions, the ChatGPT model performed substantially worse, only selecting the correct answer for 43.3% of questions. Future integration of AI models into our field should be used with caution and oversight prior to application due to its current inaccuracy and unreliability. Governing bodies need to understand the strengths, weaknesses, and potential for intentional deceit posed when rolling-out continuing certification portals and methodologies.

Author Contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Habib G. Zalzal and Rahul K. Shah. Statistical analysis was provided by Jenhao Cheng. The first draft and subsequent revisions of the manuscript were written by Habib G. Zalzal and Rahul K. Shah. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Disclosures

Competing interests: None.

Funding source: None.

Data Availability Statement

On request.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

ORCID iD

Habib G. Zalzal http://orcid.org/0000-0002-0777-977X

References

1. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. *J Educ Eval Health Prof*. 2023;20:21.
2. 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.
3. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digital Health*. 2023;2(2):e0000198.
4. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. *PLOS Digital Health*. 2023;2(2):e0000205.
5. Daily ENT Question Bank; 2019. https://www.daily-ent.com
6. OpenAI. ChatGPT: OpenAI; 2023. https://chat.openai.com/chat
7. Gwet KL. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. 4th ed: Advanced Analytics, LLC; 2014.
8. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.
9. Altman DG. In: Lindsey CCJZJ, ed. *Practical Statistics for Medical Research (Chapman & Hall/CRC Texts in Statistical Science)*. 1st ed. London: Chapman and Hall; 1991.

10. Park I, Joshi AS, Javan R. Potential role of ChatGPT in clinical otolaryngology explained by ChatGPT. *Am J Otolaryngol*. 2023;44(4):103873.

11. Hoch CC, Wollenberg B, Lüers JC, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. *Eur Arch Otrhinolaryngol*. 2023;280: 4271-4278.

12. Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. *Am J Obstet Gynecol*. 2023; 228(6):696-705.