

ChatGPT Generated Otorhinolaryngology Multiple-Choice Questions: Quality, Psychometric Properties, and Suitability for Assessments

OTO Open
2024, Vol. 8(3):e70018
© 2024 The Author(s). OTO Open
published by Wiley Periodicals LLC
on behalf of American Academy of
Otolaryngology-Head and Neck
Surgery Foundation.
DOI: 10.1002/oto2.70018
<http://oto-open.org>

WILEY

Cecilia Lotto, MD^{1,2,3*} , Sean C. Sheppard, MD^{1*} ,
Wilma Anschuetz, MD⁴ , Daniel Stricker, PhD⁴ ,
Giulia Molinari, MD^{2,3} , Sören Huwendiek, MD, PhD⁴ , and
Lukas Anschuetz, MD^{1,5,6} 

Abstract

Objective. To explore Chat Generative Pretrained Transformer's (ChatGPT's) capability to create multiple-choice questions about otorhinolaryngology (ORL).

Study Design. Experimental question generation and exam simulation.

Setting. Tertiary academic center.

Methods. ChatGPT 3.5 was prompted: "Can you please create a challenging 20-question multiple-choice questionnaire about clinical cases in otolaryngology, offering five answer options?" The generated questionnaire was sent to medical students, residents, and consultants. Questions were investigated regarding quality criteria. Answers were anonymized and the resulting data was analyzed in terms of difficulty and internal consistency.

Results. ChatGPT 3.5 generated 20 exam questions of which 1 question was considered off-topic, 3 questions had a false answer, and 3 questions had multiple correct answers. Subspecialty theme repartition was as follows: 5 questions were on otology, 5 about rhinology, and 10 questions addressed head and neck. The qualities of focus and relevance were good while the vignette and distractor qualities were low. The level of difficulty was suitable for undergraduate medical students ($n = 24$), but too easy for residents ($n = 30$) or consultants ($n = 10$) in ORL. Cronbach's α was highest (.69) with 15 selected questions using students' results.

Conclusion. ChatGPT 3.5 is able to generate grammatically correct simple ORL multiple choice questions for a medical student level. However, the overall quality of the questions was average, needing thorough review and revision by a medical expert to ensure suitability in future exams.

Keywords

artificial intelligence, ChatGPT, exam, large language model, multiple choice question, otolaryngology

Received July 11, 2024; accepted September 7, 2024.

Chat Generative Pretrained Transformer (ChatGPT) is an artificial intelligence (AI) chatbot based on a large language model (LLM) able to converse with human-like text to its user, released in 2022. Trained initially with large data of text, it is fine-tuned with a method called "Reinforcement Learning with Human Feedback" contributing to its ongoing improved models performance.¹ In health care, ChatGPT has been employed for intelligent question-answering, providing information about diseases, symptoms, medications, and treatments to the patients. Moreover, it has demonstrated efficacy in generating medical documentation, such as patient clinical letters, radiology reports, medical notes, and discharge summaries.^{2,3} AI models like ChatGPT also

¹Department of Otorhinolaryngology, Head and Neck Surgery, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland

²Department of Otolaryngology, Head and Neck Surgery, IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy

³Department of Medical and Surgical Sciences, Alma Mater Studiorum-University of Bologna, Bologna, Italy

⁴Institute for Medical Education, University of Bern, Bern, Switzerland

⁵Department of Otorhinolaryngology, Head and Neck Surgery, CHUV, University of Lausanne, Lausanne, Switzerland

⁶The Sense Innovation and Research Center, Lausanne, Switzerland

*These authors contributed equally to this article as co-first authors.

Corresponding Author:

Lukas Anschuetz, MD, Department of Otorhinolaryngology, Head and Neck Surgery, CHUV, University of Lausanne, Rue du Bugnon 46, 1011 Lausanne, Switzerland.

Email: anschuetz.lukas@gmail.com

offer potential in research and data analysis by examining vast amounts of medical literature. They may assist researchers in identifying patterns and predicting outcomes based on historical data, potentially enhancing efficiency and accuracy for health care providers.^{4,5}

In the field of medical education, studies have challenged ChatGPT's performance in answering exam questions in different fields, passing the Medical College Admission Test, bar exam, or the United States Medical Licensing Examination (USMLE).⁶⁻¹⁰ For instance, Gilson et al assessed its performance on the USMLE questions, showing the model's capability to achieve a sufficient score for a third-year medical student. This demonstrates a potential application of ChatGPT as an interactive medical learning tool.⁷ In the field of otorhinolaryngology (ORL), Hoch and et al tested the accuracy of ChatGPT in answering quiz questions for board certification. ChatGPT answered correctly to 57% of the questions but showed lack of knowledge or need for further refinement, particularly in areas such as “legal aspect” or “oral cavity and pharynx.”¹¹

In contrast, the aim of this study is to explore ChatGPT's capability to create assessment questions. Typically, experts in the field write multiple-choice questions for exams, incurring significant costs for universities. Therefore, we investigate whether AI could support this process. Specifically, we evaluate the quality of generated questions and their psychometric properties after exposing them to consultants, residents, and medical students.

Materials and Methods

The evaluation of ChatGPT's ability to generate questions was conducted between July 4 and 6, 2023, using the latest free version of ChatGPT (ChatGPT 3.5; OpenAI). Initially we asked ChatGPT, “Can you make a difficult multiple-choice quiz about otolaryngology?.” However, the questions created were considered overly mnemonic and theoretical by our research group, and inappropriate for medical school students or ORL residents. We then rephrased the request to: “Can you please create a multiple-choice quiz about otolaryngological clinical cases?.” Finally, we modified the question to: “Can you please create a challenging 20-question multiple-choice questionnaire about clinical cases in otolaryngology, offering five answer options?.” The questions formulated based on this request were chosen for further analysis. Additionally, we asked ChatGPT to provide the correct answers to the newly created test. Questions and answers are listed in **Table 1**, and correct answers according to ChatGPT are highlighted in bold.

Questions generated by ChatGPT were incorporated into a Google Forms document and distributed electronically to final year medical students and ORL medical doctors of varying experience levels, including residents and consultants. No formal ethics committee approval is

required according to the Kantonale Ethikkommission Bern for this study not involving medical data. Participants were recruited from Switzerland and Italy. The questions were answered from July to September 2023. Data collected through Google Forms, was exported to a Microsoft Excel Datable (Microsoft Corporation) for information processing and encoding of open variables. After collecting the filled questionnaires, the data analysis was conducted by the authors as follows.

Quality Analysis

The questions and answers quality and validity were evaluated by separately by a senior ORL (L.A.) and medical education expert (W.A.) based on the formal quality criteria of Krebs, as well as on Case and Swanson.^{12,13} Five criteria were evaluated as follows: (1) *Vignette*: the question is written in clear and understandable language, and provides sufficient information to evaluate an authentic case. The wording is authentic, professional, unambiguous, and clear. (2) *Focus*: The questions should be application-oriented, directly related to the clinical scenario presented, and maintain a clear focus. (3) *Distractors*: The answer options should be plausible and relevant to the main symptoms described and formulated homogeneously. They should not provide any hints or cues that would easily reveal the correct answer. (4) *Relevance*: The content of the question is of high relevance for the specialty. The knowledge of the single item is important for evaluating the competence of the candidate. (5) *Correct answer*: With the given information in the vignette, the depicted answer is clearly the best option. There is no second answer nearly as correct as the correct one. Each criterion was evaluated on a 5-point Likert evaluation scale shown in **Table 2**. In total, the questions could reach 5 points as lowest total score and 25 points as highest total score.

Quantitative Analysis

Group Differences

After collecting the participants' responses, the participants' total score was calculated as the sum of the points on all questions. This total score then served as the dependent variable in a single factor analysis of variance (ANOVA) with current position of the participants as a grouping factor with 3 levels (students, residents, and consultants).

For a second analysis, 3 subscores regarding topic of the questions (otology, rhinology, head and neck) were formed as the sum of the corresponding items. The percentage of correct answers for each participant's subscore (total of 3 subscores for every subspeciality) was then used as the dependent variable in a second ANOVA. Subscore served as within subject effect with 3 levels and current position of the

Table 1. ChatGPT-Generated Multiple-Choice Questions

Questions	Answers
1. A 45-year-old male presents with a hoarse voice, difficulty swallowing, and a neck mass. Flexible laryngoscopy reveals a vocal cord lesion. Which condition is most likely?	1: Vocal cord polyp 2: Vocal cord paralysis 3: Laryngeal papilloma 4: Laryngeal cancer 5: Laryngopharyngeal reflux
2. A 7-year-old child presents with recurrent episodes of ear pain, fever, and hearing loss. Otoscopy reveals a bulging and erythematous tympanic membrane. Which condition is most likely?	1: Acute otitis media 2: Otitis externa 3: Otitis interna 4: Serous otitis media 5: Cholesteatoma
3. A 55-year-old female presents with a painless, slow-growing neck mass in the anterior triangle. Fine needle aspiration cytology reveals papillary structures with psammoma bodies. Which condition is most likely?	1: Thyroid adenoma 2: Thyroid carcinoma 3: Parotid tumor 4: Lymphoma 5: Branchial cleft cyst
4. A 25-year-old male presents with recurrent nasal congestion, sneezing, and clear nasal discharge. Physical examination reveals pale and boggy nasal mucosa. Which condition is most likely?	1: Allergic rhinitis 2: Vasomotor rhinitis 3: Chronic sinusitis 4: Nasal polyps 5: Foreign body in the nose
5. A 40-year-old female presents with sudden-onset severe vertigo, nausea, and hearing loss. Examination reveals horizontal nystagmus. Which condition is most likely?	1: Benign paroxysmal positional vertigo 2: Meniere's disease 3: Vestibular neuritis 4: Labyrinthitis 5: Acoustic neuroma
6. A 6-year-old child presents with inspiratory stridor, barking cough, and hoarse voice. The symptoms worsen at night. Which condition is most likely?	1: Croup 2: Epiglottitis 3: Retropharyngeal abscess 4: Laryngomalacia 5: Tracheitis
7. A 50-year-old male smoker presents with a persistent sore throat, odynophagia, and an enlarged cervical lymph node. Biopsy reveals squamous cell carcinoma. Which region is most likely affected?	1: Oropharynx 2: Nasopharynx 3: Larynx 4: Hypopharynx 5: Paranasal sinuses
8. A 35-year-old female presents with recurrent episodes of severe facial pain in the region of the maxillary teeth. The pain worsens when bending forward. Which condition is most likely?	1: Sinusitis 2: Temporomandibular joint disorder 3: Trigeminal neuralgia 4: Dental caries 5: Glossopharyngeal neuralgia
9. A 60-year-old male presents with a painless, progressively enlarging lump in the submandibular region. Examination reveals a firm, nontender mass. Which condition is most likely?	1: Submandibular salivary gland stone 2: Submandibular salivary gland tumor 3: Lymphadenopathy 4: Parotid gland tumor 5: Sublingual gland tumor
10. A 70-year-old male presents with a chronic cough, weight loss, and a history of smoking. Flexible bronchoscopy reveals a lesion in the right main bronchus. Which condition is most likely?	1: Lung cancer 2: Tuberculosis 3: Bronchiectasis 4: Bronchial asthma 5: Pulmonary embolism
11. A 30-year-old female presents with recurrent episodes of epistaxis (nosebleeds) and telangiectasias on the lips and oral mucosa. Which condition is most likely?	1: Hereditary hemorrhagic telangiectasia (Osler-Weber-Rendu syndrome) 2: Wegener's granulomatosis 3: Nasopharyngeal angiofibroma 4: Juvenile nasopharyngeal angiofibroma 5: Nasal polyps
12. A 40-year-old male presents with sudden-onset severe pain, redness, and swelling of the pinna after trauma. Examination reveals a hematoma between the perichondrium and cartilage. Which condition is most likely?	1: Auricular cellulitis 2: Perichondritis 3: Otitis externa 4: Cauliflower ear 5: Otitis media

(continued)

Table 1. (continued)

Questions	Answers
13. A 50-year-old female presents with unilateral facial weakness and inability to close her left eye. Examination reveals decreased taste sensation on the anterior two-thirds of the tongue. Which condition is most likely?	1: Bell's palsy 2: Ramsay Hunt syndrome 3: Cerebrovascular accident (stroke) 4: Trigeminal neuralgia 5: Temporomandibular joint disorder
14. A 65-year-old male presents with a painless lump in the midline of the neck, just above the thyroid cartilage. The mass moves up with swallowing and protrusion of the tongue. Which condition is most likely?	1: Thyroglossal duct cyst 2: Thyroid nodule 3: Parotid tumor 4: Branchial cleft cyst 5: Laryngocele
15. A 20-year-old male presents with acute-onset severe odynophagia, drooling, and a muffled “hot potato” voice. Examination reveals a swollen, erythematous uvula. Which condition is most likely?	1: Peritonsillar abscess 2: Retropharyngeal abscess 3: Epiglottitis 4: Tonsillitis 5: Laryngitis
16. A 45-year-old female presents with nasal congestion, posterior nasal drip, and a persistent cough. Examination reveals a deviated nasal septum. Which condition is most likely?	1: Allergic rhinitis 2: Chronic sinusitis 3: Nasal polyps 4: Nasal septal deviation 5: Nasal foreign body
17. A 60-year-old male presents with hearing loss, tinnitus, and vertigo. Audiometry reveals sensorineural hearing loss and an elevated speech discrimination score. Which condition is most likely?	1: Meniere's disease 2: Presbycusis 3: Noise-induced hearing loss 4: Otosclerosis 5: Labyrinthitis
18. A 55-year-old female presents with a painless, slowly enlarging mass in the parotid region. Examination reveals a firm, nontender mass with facial nerve involvement. Which condition is most likely?	1: Pleomorphic adenoma 2: Warthin tumor 3: Mucoepidermoid carcinoma 4: Adenoid cystic carcinoma 5: Parotitis
19. A 70-year-old male presents with sudden-onset severe dyspnea, stridor, and cyanosis. Physical examination reveals a swollen and erythematous epiglottis. Which condition is most likely?	1: Epiglottitis 2: Laryngomalacia 3: Retropharyngeal abscess 4: Croup 5: Tracheitis
20. A 25-year-old female presents with nasal obstruction, hyposmia, and recurrent sinus infections. Imaging reveals opacification of the paranasal sinuses. Which condition is most likely?	1: Chronic sinusitis 2: Allergic rhinitis 3: Nasal polyps 4: Nasal septal deviation 5: Foreign body in the nose

Correct answers according to ChatGPT are highlighted in bold.

Abbreviation: ChatGPT, Chat Generative Pretrained Transformer.

Table 2. Quality Criteria Used to Assess the Questions

Vignette	Contains enough information for evaluation of an authentic clinical case. Clear language is used.	1/2/3/4/5
Focus	Application-oriented and focused question.	1/2/3/4/5
Distractors	Suitable answer options for the main symptom(s), without cues to the correct answer.	1/2/3/4/5
Relevance	Content is of high relevance for the depicted subject.	1/2/3/4/5
Correct answer	With the given information in the vignette, the depicted answer is clearly the best option.	1/2/3/4/5

Each of the 5 formal aspects were evaluated on a 5-point Likert scale: 1—Not at all / 2—Rather not / 3—Neither nor / 4—Rather yes / 5—Absolutely.

participants served as between subject effect resulting in a 2-factorial split-plot design.

Psychometric Analysis

The internal consistency (Cronbach's α) of the total set of 20 items for all participants was calculated as well as the

difficulty and discrimination index of every single item. The discrimination index for every question is calculated by the percentage of correct answers in the upper versus the lower performance group. This calculation was repeated for students and the rest of the participants separately in an iterative way to determine the internal

consistency (Cronbach's α) with the minimum set of questions, with discrimination index of 0.1 and above. As a result, we developed a set of questions suitable for the entire group, including the students and other participants.

Other Variables

In order to control for other variables, participants' age, their current position, the number of years in ORL, and test duration were used in a stepwise multiple linear regression to predict the total score achieved in the test. A P value of less than .05 was considered significant, post hoc tests were corrected for multiple testing where appropriate.

Statistical Analysis

All data were exported to the Statistical Package for Social Sciences, version 28 (IBM Corp). The quality of the items for their formal structure and correctness was analyzed in a descriptive way. Mean and standard deviations over all items were calculated for the 5 different quality measurements.

Results

ChatGPT-3.5 generated 20 exam questions, reported in Table 1, of which 1 question (Q10) was considered off-topic (pneumology) but included in the subsequent analysis. Subspecialty theme repartition was as follows: 5 questions were on otology, 5 about rhinology, and 10 questions addressed head and neck. The questions were answered by a total of 64 participants, of which 24 (37.5%) were medical students, 30 (46.9%) were ORL residents, and 10 (15.6%) were ORL consultants.

Regarding the answer key, 3 questions (Q1 [new solution 4], Q13 [new solution 1], Q18 [new solution 3]) had an incorrect answer key and were corrected for further analysis by consent of the senior authors (L.A., W.A., G.M.) of the study. Three questions (Q8, Q16, Q20) had multiple correct answers (Q8 [correct 3 and additional 1], Q16 [correct 4 and additional 2], Q20 [correct 1 and additional 3]). Both answers were considered as correct in the quantitative analysis. Overall, we report a correct and unambiguous answer key in 70% of the questions and a mean score of 3.8/5 ($SD \pm 1.83$) for the correct answer quality. ChatGPT delivered the correct answer for 13 of 20 questions as the first answer on position 1 (65%) as reported in Table 1.

Quality Analysis

Formal quality criteria of the questions are reported in Figure 1. Overall correct answer quality scored 3.75/5 ($SD \pm 1.81$). Focus quality and topic relevance showed an over-average score of respectively 4.7/5 ($SD \pm 0.95$) and 4.7/5 ($SD \pm 0.95$). Vignette quality and distractor quality performed worse with a score of 3.1 ($SD \pm 1.24$) and 2.6/5 ($SD \pm 1.2$).

Quantitative Analysis

After quality analysis and correction of the solutions, the answers provided by the participants were analyzed with regard to differences in performance between students, residents and consultants. In a second step, an attempt was made to identify questions suitable for inclusion in an undergraduate examination and questions suitable for postgraduates. The age of the participants was correlated with years in ORL ($r = .901$, $P < .001$) and the position ($r = .813$, $P < .001$). The stepwise multiple regression

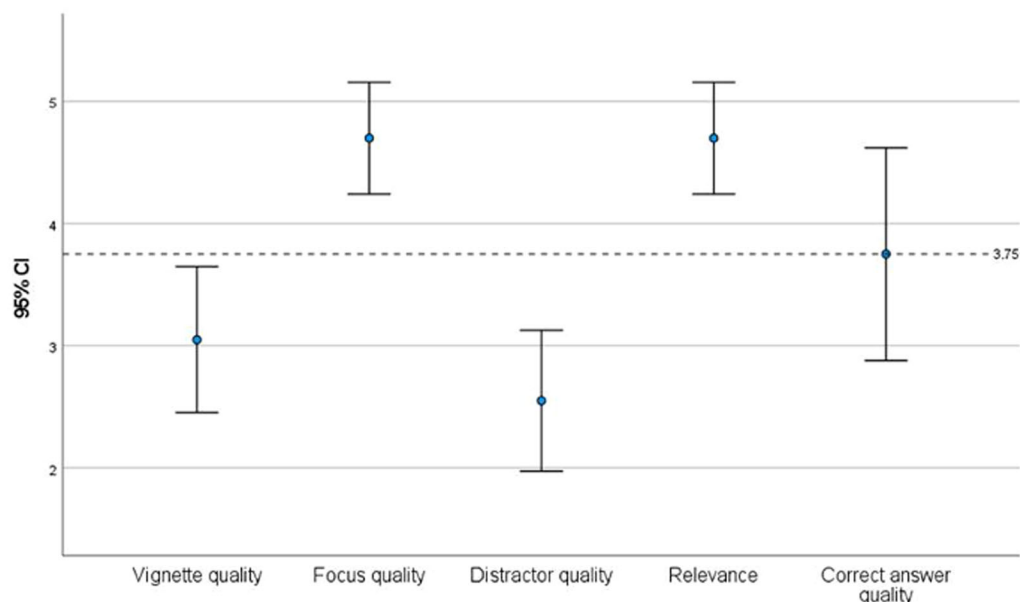


Figure 1. Formal quality assessments. CI, confidence interval.

revealed that current position and number of ORL years of experience predicted 38.6% of the total variance of the total score.

Group Differences

Figure 2 shows the difficulty of each question for all groups separately. Seventy-nine percent (19/24) of the students answered $\geq 50\%$ of all questions correct, while 1 student achieved $\geq 80\%$ correct answers. All residents and consultants answered $\geq 50\%$ correct. Sixty percent (18/30) of residents, and 80% (8/10) consultants answered $\geq 80\%$ correct. The 1-way ANOVA showed a significant effect of group ($F_{2,61} = 26.194$, $P < .001$) whereas students performed significantly worse (mean $P = 59.17\%$ correct, $SE = 2.69$) than residents (mean $P = 79.00\%$, $SE = 1.67$) and consultants (mean $P = 80.00\%$, $SE = 2.69$).

Regarding the secondary analysis per subspecialty (otology, rhinology, head and neck), we observed in the split-plot ANOVA, plotted in **Figure 3**, that the main effect subscore turned out significant ($F_{2,122} = 19.759$, $P < .001$, $\eta_p^2 = 0.245$). This shows that the generated

questions for the 3 subscores differed in terms of overall performance. Subscore *otology* had the least correct answers ($P = 60.63\%$) followed by *head and neck* ($P = 71.41\%$) and *rhinology* ($P = 83.44\%$). The main effect of the education level ("group") was confirmed ($F_{2,61} = 26.194$, $P < .001$, $\eta_p^2 = 0.461$), showing again that the students' performance was worse in all 3 subscores compared to the performance of the residents and the consultants. No difference in performance was observed between residents and consultants. There was no significant interaction between the subscores and the group membership ($F_{4,122} = 0.174$, $P < .951$), confirming that the lower performance of the students compared to the other participants was not attributable to a specific subscore but was similar in all 3 subscores.

Psychometric Analysis

To evaluate consistency of the questions and to determine a minimal set of questions with the maximal consistency, we performed reliability statistics, where a positive trend in the question's discrimination index is interpreted as an adequate exam question. The reliability analysis for all participants revealed that 19/20 questions showed an overall positive corrected item-total correlation, thus confirming the relevance of the items. Eliminating questions 8 and 9 because of their low discrimination index ($r_{it} < 0.1$) led to an increase of the observed Cronbach's α from .631 to .679. Since performance of the students significantly differed from the residents' and consultants' performance, the reliability analysis was repeated based on 2 subgroups, for the students and the rest (residents and consultants) of the participants. The result of this analysis is displayed in **Table 3**. For the students the initially obtained Cronbach's α of .456 improved to .691 by selecting 15 questions that made an appropriate contribution to the overall result. For the rest of the participants 14 questions had to be disregarded in order to get an improved but still low Cronbach's α of .458. In this regard, most questions showed a reliability with positive trend for the subgroup of students, which would be approved as possible implementable exam questions. For residents and consultants, very few questions were considered for a hypothetical test set, which also achieved only low internal consistency.

Discussion

Our findings show that the LLM ChatGPT is capable of generating grammatically on point multiple-choice questions relevant to the topic of general ORL. Detailed evaluation revealed the questions to be of moderate quality as the qualities of focus and relevance were good while the vignette and distractor qualities were low. The level of difficulty was suitable for undergraduate medical students, but too easy for residents or consultants in ORL.

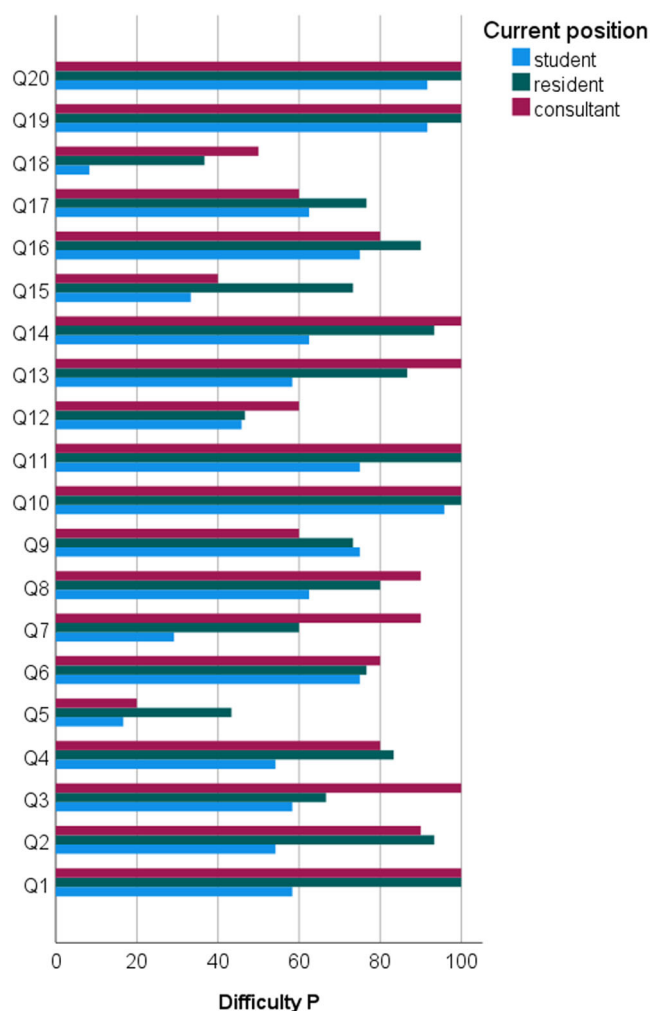


Figure 2. Mean difficulty values as percentage of correct answers displayed for all questions and all 3 groups of participants separately.

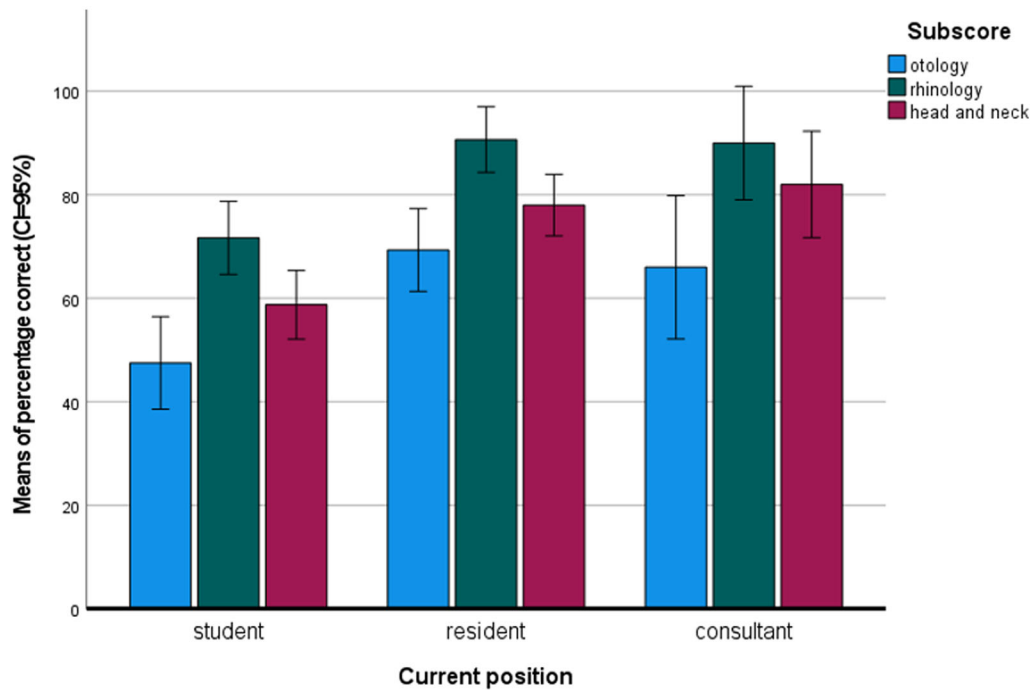


Figure 3. Mean values of percentage correct answers displayed for the 3 groups and all 3 subscores separately. Error bars denote the 95% confidence interval of the means.

Table 3. Results of the Reliability Analysis for All Participants and Separately for Students and the Rest of the Participants

Question	Discrimination indices					
	All participants (N = 64)		Students (N = 24)		Residents, consultants (N = 40)	
	Initial	Final	Initial	Final	Initial	Final
Q1	0.553	0.512	0.417	0.473	na	
Q2	0.262	0.285	−0.215		0.272	0.199
Q3	0.369	0.368	0.538	0.589	0.178	
Q4	0.167	0.163	−0.091		0.007	
Q5	0.176	0.189	0.149	0.156	−0.034	
Q6	0.142	0.157	0.216	0.224	0.133	0.273
Q7	0.36	0.363	0.081	0.243	0.246	0.289
Q8	0.032		−0.2		−0.106	
Q9	−0.106		0.096	0.185	−0.247	
Q10	0.374	0.356	0.496	0.38	na	
Q11	0.146	0.17	−0.237		na	
Q12	0.083	0.136	0.202	0.188	−0.029	
Q13	0.349	0.33	0.191	0.186	0.088	
Q14	0.159	0.189	−0.2		−0.078	
Q15	0.372	0.401	0.415	0.533	0.089	0.177
Q16	0.267	0.267	0.341	0.224	0.084	
Q17	0.148	0.183	0.135	0.24	0.08	0.222
Q18	0.377	0.387	0.336	0.327	0.185	0.221
Q19	0.28	0.284	0.231	0.268	na	
Q20	0.411	0.384	0.486	0.449	na	
Cronbach α	.631	.679	.456	.691	.185	.458

Item discrimination indices are displayed for the initial calculation as well as for the final iteration of the selection of questions. na denotes zero variance items in the initial iteration.

Questions generated by LLMs could be an option to increase the efficacy of the process of writing new questions. Creating examination questions for medical students requires a thorough, multistep quality process with experts from both the medical specialist area and the assessment development experts. The process of generating questions for medical examinations is laborious and is associated with relevant costs to the universities and regulatory authorities.¹⁴ In this regard, a chatbot which almost instantly generates a series of questions seems very time-efficient. However, it is essential that the generated questions meet the quality requirements of the respective exam.

On the quality assessment, clear wording was analyzed, which was mostly the case. Nevertheless, according to this panel of questions, ChatGPT does not seem to be able to use the exact medical terms correctly. There were expressions used inappropriately or imprecisely, which weakened the reliability of the exam questions. The use of a clear and unambiguous language is essential regarding the quality of multiple-choice questions used for assessment purposes.^{13,15} The criterion of focus and application-oriented questioning was met in 19/20 questions. However, all 19 questions contained identical wording for the question phrase (“Which condition is most likely?”). Only 1 question was formulated differently and, in this case, not application-oriented. Although the prompt was formulated in general terms, ChatGPT was not able to reflect the versatility in accordance with everyday clinical practice. At best, this could also be remedied by targeted prompting with the formulation of individual topics but would also increase the effort involved in creating the questions. All the available information matched the answer sought, but the distractors (wrong answers) often could not really be assessed because relevant clinical aspects for assessing the distractors were not available in the vignette. This fact often made it possible to judge the other answer options as incorrect/unattractive very quickly with only basic medical knowledge. Such effects also weaken the reliability of the examination question, as it is not really necessary to know the answer sought, but the correct answer is found by excluding incorrect answers. Accordingly, the quality of the answer options was the greatest weakness of the questions in terms of both form and content. Many correct answers could be solved based on hints, linguistic associations and the simplest interpretations of findings from the vignette, without the need for in-depth medical knowledge. Some distractors were also so absurd that the guess probability for the remaining answers increased significantly.

The prompt was about “clinical cases in otolaryngology,” but no real vignettes or case studies were formulated. The generated questions tested typical textbook knowledge, with 19 out of 20 questions asking the most probable diagnosis. For the most part, the questions tested which clinical picture is most likely given a

constellation of 2 to 3 symptoms and 1 to 2 other aspects. Even if it was sufficient, relevant information was missing in almost all questions, such as the information “since when” the symptoms have been present or the development of the symptoms, which should never be missing in a medical history in everyday life. Furthermore, none of the questions referred to diagnostic steps, management, or therapy. A well-balanced “application-oriented” examination should assess the participant's knowledge while also preparing them for future real-world clinical scenarios.

Despite ChatGPT's impressive ability to mimic human writing, attention must be paid to its limitations. Our results report 1 off-topic question and multiple false given answers. ChatGPT may create convincing text, but it remains unknown how the AI model reasons to find answers.¹⁰ Moreover, the model relies on the training data which is not updated in real time and itself subject to bias.¹⁶ This can lead to outdated examination material without regard to the latest research or guidelines. Fortunately, it does not affect the questions that are based on strong evidence. In addition, concerns may arise on the confidentiality of exam questions when generated by an open LLM. Finally, we observed that the accuracy of ChatGPT's responses is dependent on the specificity and clarity of the prompt provided. Contextual information, such as the specifying role (eg, “as a university teacher”) or the medical specialty (eg, “otolaryngology”) should be included. A more detailed target audience, for example, “final year medical students” could have improved the accuracy of the level of the questions. The type of questions, “multiple-choice questions offering five answer options,” was specified but not that only 1 answer should be correct. ChatGPT also offers the ability to incorporate resources like textbooks or clinical guidelines directly within the prompt, which can help generate questions that are more aligned with specific subject matter.

In this study, we did not include a control group and used a limited set of 20 questions which can be considered a limitation. The prompting could probably also be further refined and improved in order to improve the quality of the questions. We believe, that this would be an interesting topic for future investigations. This study's strengths are its large number of participants of different educational levels and the realistic qualitative and statistical analysis as it would have been a real final grade exam at medical school. This interdisciplinary project included ORL specialists as well as medical education experts allowing a comprehensive investigation of the topic. The described observations revealed important limitations in the quality of the AI-generated questions, making them not suitable for use in a real assessment situation. In addition, from our point of view, a huge amount of work would have to be invested into the questions in order to bring them on an acceptable level. Nevertheless, the questions reached suitable statistical numbers and were able to reasonably discriminate between students and ORL specialists.

Conclusion

ChatGPT 3.5 can generate grammatically correct simple ORL multiple-choice questions for a medical student level. However, the overall quality of the clinical cases was average, especially regarding clinical vignette and realism. To be used in an assessment setting, a thorough review and extensive revision by a medical expert to bring them on a suitable level would be strongly advised.

Author Contributions

Cecilia Lotto, conception and design, data acquisition and analysis, manuscript drafting, revision and approval of the final version; **Sean C. Sheppard**, data acquisition and analysis, manuscript drafting, revision and approval of the final version; **Wilma Anschuetz**, conception and design, data acquisition and analysis, manuscript drafting, revision and approval of the final version; **Daniel Stricker**, data acquisition and analysis, manuscript drafting, revision and approval of the final version; **Giulia Molinari**, data acquisition and analysis, manuscript drafting, revision and approval of the final version; **Sören Huwendiek**, conception and design, manuscript drafting, revision and approval of the final version; **Lukas Anschuetz**, conception and design, data acquisition and analysis, manuscript drafting, revision and approval of the final version.

Disclosures








Competing interests: The authors declare that there is no conflict of interest.

Funding source: No funding was received for this work.

Data Availability Statement

Anonymized version of the data is available upon request.

ORCID iD

Cecilia Lotto  <https://orcid.org/0000-0001-6276-1997>
 Sean C. Sheppard  <https://orcid.org/0000-0001-6427-3102>
 Wilma Anschuetz  <https://orcid.org/0009-0000-9385-546X>
 Daniel Stricker  <http://orcid.org/0000-0002-7722-0293>
 Giulia Molinari  <https://orcid.org/0000-0002-3402-4364>
 Sören Huwendiek  <https://orcid.org/0000-0001-6116-9633>
 Lukas Anschuetz  <https://orcid.org/0000-0001-6927-0664>

References

1. OpenAI. What is ChatGPT? 2024. Accessed July 10, 2024. <https://help.openai.com/en/articles/6783457-what-is-chatgpt>
2. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388:1233-1239.
3. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. *J Med Internet Res*. 2023;25:e48568.
4. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst*. 2023;47:33.
5. Mu Y, He D. The Potential applications and challenges of ChatGPT in the medical field. *Int J Gen Med*. 2024;17:817-826.
6. Mihalache A, Popovic MM, Muni RH. Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment. *JAMA Ophthalmol*. 2023;141:589.
7. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312.
8. Revilla-León M, Barmak BA, Sailer I, Kois JC, Att W. Performance of an artificial intelligence-based chatbot (ChatGPT) answering the European Certification in Implant Dentistry Exam. *Int J Prosthodont*. 2024;37:221-224.
9. Katz DM, Bommarito MJ, Gao S, Arredondo P. GPT-4 passes the bar exam. *Philos Trans R Soc A*. 2024;382:20230254.
10. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. *JMIR Med Educ*. 2019;5:e16048.
11. Hoch CC, Wollenberg B, Lüers J-C, 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 Otorhinolaryngol*. 2023;280:4271-4278.
12. Case SM, Swanson DB. *Constructing Written Test Questions for the Basic and Clinical Sciences*. National Board of Medical Examiners; 1998.
13. Krebs R. *Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung*. Institut für Medizinische Lehre IML, Abteilung für Ausbildungs- und Examensforschung AAE; 2004.
14. Mukurunge E, Nyoni CN, Hugo L. Assessment approaches in undergraduate health professions education: towards the development of feasible assessment approaches for low-resource settings. *BMC Med Educ*. 2024;24:318.
15. Abedi J. Language issues in item development. In: Lane S, Raymond MR, Haladyna TM, eds. *Handbook of Test Development*. Routledge; 2011:391-412.
16. Fang X, Che S, Mao M, Zhang H, Zhao M, Zhao X. Bias of AI-generated content: an examination of news produced by large language models. *Sci Rep*. 2024;14:5224.