KBB Uygulamaları 2026;14(1):46-54

Original Article / *Özgün Makale*

Accuracy and consistency of ChatGPT responses as an educational tool in the vestibular system

ChatGPT yanıtlarının eğitim aracı olarak vestibüler sistemdeki doğruluğu ve tutarlılığı

Gulce Kirazli[1], Sumeyye Kapusizoglu[2]

[1]*Department of Audiology, Faculty of Health Sciences, Ege University, İzmir, Türkiye*
[2]*Department of Audiology, Institute of Health Sciences, Ege University, İzmir, Türkiye*

ABSTRACT

Objectives: The aim of this study was to compare the accuracy and consistency of ChatGPT's responses in Turkish and English, specifically within the vestibular field.

Materials and Methods: Based on a review of the current literature, a total of 42 questions were created in three subcategories: vestibular system anatomy and physiology, vestibular diagnosis and tests, and vestibular rehabilitation and treatment. These questions were presented to ChatGPT 3.5 in both Turkish and English. The accuracy of the responses was evaluated by nine experts using a five-point Likert scale. One week later, the same questions were presented in a different order, and the consistency of the responses was assessed by the authors using the same five-point Likert scale.

Results: English responses (3.57±0.34 and 3.97±0.47, respectively) were significantly more accurate than Turkish responses (3.33±0.27 and 3.64±0.36, respectively) in the first and third subcategories. Both inter-rater reliability and test-retest reliability were found to be high for both languages.

Conclusion: Our study findings suggest that ChatGPT may serve as a complementary educational tool; however, the reliability and accuracy of its content in specific areas such as the vestibular system require further validation. Additionally, language choice, such as Turkish versus English, may influence performance in certain subcategories.

Keywords: Audiology, ChatGPT, vestibular system.

ÖZ

Amaç: Bu çalışmada, ChatGPT'nin özellikle vestibüler alandaki Türkçe ve İngilizce yanıtlarının doğruluğu ve tutarlılığı karşılaştırıldı.

Gereç ve Yöntemler: Güncel literatür taranarak vestibüler sistem anatomisi ve fizyolojisi, vestibüler tanı ve testler, vestibüler rehabilitasyon ve tedavi olmak üzere üç alt kategoride toplam 42 soru oluşturuldu. Sorular ChatGPT 3.5'e Türkçe ve İngilizce olarak yöneltildi. Yanıtların doğruluğu dokuz uzman tarafından beş puanlık Likert ölçeği ile değerlendirildi. Bir hafta sonra sorular farklı sırayla tekrar sorularak yanıtların tutarlılığı yazarlar tarafından aynı beş puanlık Likert ölçeği kullanılarak analiz edildi.

Bulgular: İngilizce yanıtlar birinci ve üçüncü alt kategorilerde (sırasıyla 3.57±0.34 ve 3.97±0.47) Türkçe yanıtlara (sırasıyla 3.33±0.27 ve 3.64±0.36) kıyasla anlamlı derecede daha doğruydu. Her iki dil için de değerlendiriciler arası güvenirlik ve test-tekrar test güvenirliği yüksekti.

Sonuç: Çalışma sonuçlarımız, ChatGPT'nin tamamlayıcı bir eğitim aracı olarak kullanılabileceğini göstermektedir; ancak, ChatGPT tarafından vestibüler sistem gibi belirli konularda üretilen içeriğin güvenilirliği ve doğruluğunun doğrulanması gerekmektedir. Ayrıca Türkçe veya İngilizce gibi dil seçimi belirli alt kategorilerdeki performansı etkileyebilir.

Anahtar sözcükler: Odyoloji, ChatGPT, vestibüler sistem.

Received: September 02, 2025
Accepted: January 13, 2026
Published online: January 28, 2026

Correspondence: Sumeyye Kapusizoglu, MSc.
E-mail: sumeyyekapusizoglu@gmail.com

Doi: https://doi.org/10.5606/kbbu.2026.87699

Citation:
Kirazli G, Kapusizoglu S. Accuracy and consistency of ChatGPT responses as an educational tool in the vestibular system. KBB Uygulamaları 2026;14(1):46-54. https://doi.org/10.5606/kbbu.2026.87699.

© 2026 Praxis of Otorhinolaryngology. This is an open access article published under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium, provided the original work is properly cited. http://creativecommons.org/licenses/by-nc/4.0/

Artificial Intelligence (AI) is a technology that, using algorithms and computer systems, attempts to perform activities that human intelligence can perform.[1] ChatGPT is an AI model created by OpenAI that can produce human-like texts.[2] Launched in November 2022, it has emerged as one of the fastest growing web applications.[3] Various scientific studies have employed ChatGPT and other natural language processing (NLP) models. These involve literature review, language barriers, medical report composition, medical decision-making, and more.[4] It also serves the role of providing information to students during their medical education.[5]

ChatGPT's capabilities in regards to the otolaryngology board certification exam, as well as its potential as a patient information source and educational tool for ear, nose, and throat (ENT) specialists have been evaluated in the literature.[1,6-8] Hoch et al.[1] suggested that, although ChatGPT held substantial promise as a supportive tool for examination in the field of otolaryngology, it should be used with caution as an educational tool and evaluated for errors in some otolaryngology areas. In addition, while ChatGPT has reached a massive database, it may lack the domain expertise for advanced or focused subjects in otolaryngology.[7] Revercomb et al.[6] found in their study that ChatGPT provided accurate answers to 51% of the questions, with accuracy varying by subject. They suggested that ChatGPT can be used as a resource for board exams in otolaryngology with caution regarding the validity of its responses.

Despite ChatGPT's current benefits and unrealized potential, ensuring the precision and reliability of the data the AI system provides remains a significant challenge. High precision and accuracy are essential in medical education since, even a little mistake can have a major impact on patient safety.[5] To the best of our knowledge, ChatGPT's performance in specific areas related to audiology and neurotology has not been published in the literature yet.

In the present study, we aimed to evaluate and compare the accuracy and temporal consistency of ChatGPT's responses in Turkish and English, positioning it as a supportive educational resource for audiologists on topics related to the vestibular system.

MATERIALS AND METHODS

Study design

An initial Google Forms survey was conducted to identify audiologists' professional areas and their primary informational needs regarding ChatGPT. Eligible participants were required to hold a bachelor's or graduate degree in audiology and to be actively working in the field. The survey, distributed via social media and messaging platforms, was completed by 105 audiologists (65 females, 40 males). Among the participants, 63.8% held a Bachelor's degree, 8.6% a Master's degree, and 8.6% a Doctor of Philosophy (PhD), while 14.3% were Master's students and 4.8% were PhD candidates. Survey responses indicated that the area most in need of ChatGPT-based resources was Vestibular Audiology (32.4%), followed by Auditory Perception and Rehabilitation (28.6%), Pediatric Audiology (15.2%), Hearing Devices and Implantable Technologies (15.2%), additional responses for Auditory Perception and Rehabilitation (4.8%), and Tinnitus/Misophonia/Hyperacusis (2.9%).

A total of 42 questions, including 23 open-ended and 19 single-choice items, were created in the field of the vestibular system by reviewing Turkish and English academic articles, as well as audiology textbooks. Each single-choice question contained four options, with one designated as the correct answer. The questions were divided into three subcategories: 14 on vestibular system anatomy and physiology (7 open-ended, 7 single-choice), 14 on vestibular diagnosis and testing (7 open-ended, 7 single-choice), and 14 on vestibular rehabilitation and treatment (9 open-ended, 5 single-choice). All questions were developed by the study authors in both Turkish and English. A professional translator verified the translations for accuracy and made necessary adjustments. In addition, expert feedback was obtained from two neuro-otology faculty members, who provided revisions before finalization. The finalized set of questions was submitted to the ChatGPT-3.5 AI model on January 2, 2024, in both Turkish and English. Before each question was asked, ChatGPT was given the following clarification prompt: "Imagine you are an audiologist. I would like to ask you some questions about the vestibular system. Can you answer these questions in detail?"

Inter-rater reliability was used to assess the agreement on the accuracy of ChatGPT answers in both English and Turkish. Therefore, the Turkish and English responses provided by ChatGPT were evaluated for accuracy using a five-point Likert scale (1 = Strongly disagree, 2 = Disagree, 3 = Neither agree nor disagree, 4 = Agree, 5 = Strongly agree).

Three audiologists with Master's degrees, four audiologists with PhD degrees, and two otolaryngologists scored the responses. Additionally, the average accuracy scores assigned by experts to

ChatGPT's responses in both Turkish and English were compared to examine the differences between the two languages.

Test-retest reliability was used to measure the consistency of ChatGPT-generated responses in both English and Turkish across two different time points. Initially, the order of the options in all single-choice questions and the placement of the open-ended questions were altered. These modified questions, with changes in order and option sequences, were then resubmitted to the ChatGPT-3.5 AI model on January 9, 2024, using the same prompt as before. Test-retest reliability was assessed by evaluating the similarity between ChatGPT's responses at two different times using a five-point Likert scale (1 = Completely different, 2 = Different, 3 = Neither the same nor different, 4 = Similar, 5 = Completely the same).

English Question samples and ChatGPT answers and expert score averages are shown in Table 1.

Ethics committee approval was not required as the study analyzed publicly available published abstracts using an AI-based evaluation framework, without involving human subjects or confidential data.

Statistical analysis

Statistical analysis was performed using the IBM SPSS version 26.0 software (IBM Corp., Armonk, NY, USA). Descriptive statistics were expressed in mean ± standard deviation (SD), median (min-max), or number and frequency. The Mann-Whitney-U test was conducted to compare the accuracy of both Turkish and English responses from the ChatGPT model, and question types. Kruskal Wallis-H test was applied to determine the difference between the scores given by the experts for the accuracy of the responses of ChatGPT. Intraclass correlation coefficient (ICC) statistics were used for the inter-rater reliability and test-retest reliability. A p value of <0.05 was considered statistically significant.

RESULTS

Expert demographics

Turkish and English responses obtained from ChatGPT were evaluated for accuracy by nine experts in otolaryngology and audiology. In the study, 78% of the experts are female. The average age of the experts was 30.22±4.44 (range, 27 to 68) years. Two of the experts are otolaryngology professors, four are doctor audiologists and three are audiologists who hold a Master's degree. 33% of the experts work as research assistants, 33% as assistant professors, and 22% as audiologists in the hospital.

Accuracy in ChatGPT's Turkish and English responses

Turkish and English responses from ChatGPT were evaluated for accuracy by nine experts using a five-point Likert Scale. The average of the scores given by the experts to the accuracy of both Turkish and English responses by ChatGPT and the comparison of the scores are in Table 2.

There was a statistically significant difference between the average accuracy scores given by experts for Turkish and English responses for Vestibular System Anatomy and Physiology and Vestibular Treatment and Rehabilitation subcategories and total (overall) score. The ChatGPT's English responses were significantly higher ($p<0.05$).

The inter-rater reliability of ChatGPT's responses

To assess whether the Turkish and English scores obtained from one expert differed from that of another for each category, the ICC was employed. Moreover, the Kruskal-Wallis H test was applied to examine whether there were differences in the scores given by experts to the Turkish and English responses obtained from ChatGPT.

The reliability coefficient among experts in each category varied between 0.892 and 0.927, indicating a high degree of reliability in Turkish responses. There was no statistically significant difference among the mean scores given by experts for Turkish responses ($p>0.005$).

The reliability coefficient among the experts in each category varied between 0.879 and 0.980, indicating a high degree of reliability in English responses. There was also no statistically significant difference among the mean scores given by experts for English responses ($p>0.005$).

The test-retest reliability of ChatGPT's responses over time

The consistency between responses obtained from ChatGPT at different times was evaluated by two audiologists/authors in the study using a five-point Likert Scale. Turkish and English responses obtained from ChatGPT at various times were assessed for inter-rater reliability using the ICC between the two observers (Table 3, 4).

The ICC coefficient for Turkish responses ranged from 0.956 to 0.983 and 0.899 to 0.985 for English responses, indicating a high degree of consistency. No significant difference was found between the two observers ($p>0.05$).

Table 1

Sample English questions, ChatGP responses, and expert score averages

Question examples	ChatGPT's answers	Average scores given by experts
Vestibular system anatomy and physiology category		
What is the movement that causes the stereocilia to bend towards the kinocilium called?	The movement that causes the stereocilia to bend towards the kinocilium is called "excitation" or "depolarization." When the stereocilia are deflected toward the kinocilium, it leads to an influx of positively charged ions, such as potassium, into the hair cell. This influx of ions depolarizes the hair cell, generating an action potential that travels along the vestibular nerve fibers to the brain. In the vestibular system, hair cells are mechanoreceptors responsible for detecting motion and orientation changes. The bending of stereocilia towards the kinocilium is a key step in the transduction of mechanical stimuli (head movement) into electrical signals that the nervous system can interpret.	4.6
How many vestibular nerve fibers are there in humans? A. Between 40000-50000 B. Between 15000-25000 C. Between 100000-200000 D. Between 1000-2000	The number of vestibular nerve fibers in humans is approximately: B- Between 15,000-25,000	5
Vestibular diagnosis and tests category		
Which of the following is one of the test batteries used to evaluate the vestibular system? A. ABR B. Tympanometry C. TRV Chair D. Audiometer	C- TRV Chair The TRV (rotary test vestibular) chair is one of the test batteries used to evaluate the vestibular system. This chair is often used in conjunction with other tests to assess the function of the vestibular organs, particularly in relation to head movements and rotational stimuli. The TRV chair can aid in diagnosing and evaluating conditions affecting the vestibular system. The other options listed (ABR, Tympanometry, Audiometer) are not specific to vestibular testing; they are more commonly associated with auditory assessments.	4.8
Which disease should be considered in a patient who typically has severe dizziness that starts suddenly and lasts for days, is accompanied by nausea, vomiting and sweating, has no hearing loss, and has an upper respiratory tract infection?	The symptoms described, including sudden severe dizziness lasting for days, accompanied by nausea, vomiting, sweating, and the absence of hearing loss, are indicative of a condition known as vestibular neuritis. Vestibular neuritis is characterized by inflammation of the vestibular nerve, typically following a viral infection, often involving the upper respiratory tract. The sudden onset of severe dizziness, often described as a spinning sensation (vertigo), can be debilitating. Nausea, vomiting, and sweating are common accompanying symptoms. It's important to note that vestibular neuritis is distinct from labyrinthitis. Labyrinthitis involves inflammation of both the vestibular nerve and the cochlear nerve, leading to additional symptoms such as hearing loss or tinnitus, which are not present in vestibular neuritis. If someone is experiencing these symptoms, especially following an upper respiratory tract infection, it is advisable to seek medical attention for a proper diagnosis and appropriate management. Vestibular rehabilitation therapy may be recommended to aid in recovery and improve balance.	5

Table 1
Continued

Question examples	ChatGPT's answers	Average scores given by experts
Vestibular treatment and rehabilitation category		
Which of the following is not one of the indications for vestibular rehabilitation? A. Residual dizziness after BPPV treatment B. People with psychogenic vertigo. C. Acute stage of acutely developing pathologies D. Bilateral vestibular diseases	C- Acute stage of acutely developing pathologies Vestibular rehabilitation is generally not indicated during the acute stage of acutely developing pathologies, especially when there is an ongoing and significant medical condition that requires immediate attention. During the acute stage, interventions to address the underlying pathology or stabilize the patient may take precedence over vestibular rehabilitation. The other options, A, B, and D, are potential indications for vestibular rehabilitation: - *A- Residual dizziness after BPPV treatment:* Vestibular rehabilitation may be beneficial in addressing persistent symptoms or residual dizziness following the treatment of BPPV. - *B- People with psychogenic vertigo:* Vestibular rehabilitation can be considered for individuals with psychogenic vertigo, as part of a comprehensive approach that addresses both physical and psychological factors contributing to symptoms. - *D- Bilateral vestibular diseases:* Vestibular rehabilitation is often used in cases of bilateral vestibular diseases to improve balance, reduce symptoms, and enhance overall function. It's important to note that the appropriateness of vestibular rehabilitation depends on the specific characteristics of the patient's condition, and individualized assessment by a healthcare professional with expertise in vestibular disorders is crucial for determining the most suitable interventions.	4.2
Which type of BPPV is Semont's liberator maneuver used to treat?	Semont's Liberatory Maneuver is primarily used to treat posterior canal benign paroxysmal positional vertigo (PC-BPPV), also known as canalithiasis of the posterior semicircular canal. In this form of BPPV, otoconial debris (calcium carbonate crystals) becomes dislodged within the posterior semicircular canal and can cause episodes of vertigo triggered by specific head movements. The Semont's Liberatory Maneuver is designed to reposition these displaced particles within the posterior semicircular canal, alleviating symptoms associated with PC-BPPV. It involves a series of quick and controlled movements to guide the particles out of the canal and into an area where they do not provoke vertigo. It's important to note that the choice of the maneuver may depend on individual factors, and healthcare professionals, particularly those specialized in vestibular disorders, should assess and determine the most appropriate maneuver based on the patient's specific condition and symptoms.	4

ABR: Auditory brainstem response; TRV: Thomas Richard-Vitton; BPPV: Benign paroxysmal positional vertigo.

Comparison of ChatGPT's accuracy performance on single choice and open-ended questions

In the Vestibular System Anatomy and Physiology subcategory, the mean accuracy score of ChatGPT`s responses to Turkish single-choice questions was 3.22±0.27, and for open-ended questions, it was 3.44±0.41. For English, the scores were 3.14±0.16 and 4.46±0.40, respectively. In the Vestibular Diagnosis and Tests subcategory, the mean accuracy score of ChatGPT`s responses to Turkish single-choice questions was 3.13±0.53, and for open-ended

Table 2
Comparison of Turkish and English responses accuracy

Scale		Mean±SD	Alpha	Sig.
Vestibular system anatomy and physiology	Turkish	3.33±0.27	0.220	0.003
	English	3.57±0.34	0.606	
Vestibular diagnosis and tests	Turkish	3.16±0.46	0.717	0.094
	English	3.23±0.34	0.193	
Vestibular treatment and rehabilitation	Turkish	3.64±0.36	0.607	0.002
	English	3.97±0.47	0.634	
Total	Turkish	3.37±0.30	0.704	0.006
	English	3.79±0.17	0.542	

SD: Standard deviation; Alpha: Cronbach's Alpha sign; Mann-Whitney-U test *p* value <0.05.

Table 3
Test-retest reliability of ChatGPT's Turkish responses

Scale	Expert	Mean±SD	ICC	Sig.
Vestibular system	1	3.50±1.09	0.956	0.754
Anatomy and physiology	2	3.57±1.22		
Vestibular	1	3.00±1.66	0.983	0.668
Diagnosis and tests	2	3.21±1.67		
Vestibular	1	3.36±1.08	0.958	0.755
Treatment and rehabilitation	2	3.28±1.27		

SD: Standard deviation; ICC: Intraclass correlation coefficient; Sign: Mann-Whitney-U test *p* value.

Table 4
Test-retest reliability of ChatGPT's English responses

Scale	Expert	Mean±SD	ICC	Sig.
Vestibular system	1	2.93±1.54	0.985	0.924
Anatomy and physiology	2	2.93±1.64		
Vestibular	1	3.36±1.08	0.953	0.083
Diagnosis and tests	2	3.86±1.35		
Vestibular	1	3.57±0.85	0.899	0.701
Treatment and rehabilitation	2	3.71±0.91		

SD: Standard deviation; ICC: Intraclass correlation coefficient; Sign: Mann-Whitney-U test *p* value.

questions, it was 3.19±0.47. For English, the scores were 2.63±0.28 and 3.95±0.32, respectively. In the Vestibular Treatment and Rehabilitation subcategory, the mean accuracy score of ChatGPT`s responses to Turkish single-choice questions was 3.56±0.52, and for open-ended questions, it was 3.80±0.14. For English, the scores were 3.89±0.33 and 4.40±0.34, respectively.

No significant difference was found in terms of question types in the subcategories of Vestibular System Anatomy and Physiology, Vestibular Diagnosis and Tests, and Vestibular Treatment and Rehabilitation in both English and Turkish languages ($p>0.05$).

DISCUSSION

In the present study, we evaluated ChatGPT as a supportive resource for audiologists in the vestibular field, moving beyond previous research that examined AI models in broader medical or educational contexts by focusing specifically on vestibular-related subcategories. To the best of our knowledge, this study represents the first comparison of ChatGPT's responses in Turkish and English within audiology and neurotology. Nielsen et al.[7] reported that ENT physicians assessed ChatGPT as a patient information source in otolaryngology, assigning mean scores of 3.51 for accuracy, 3.71 for relevance, and 3.00 for depth; in our study, ChatGPT's mean accuracy scores in English and Turkish were comparable. Nielsen et al.[7] concluded that these findings suggest AI systems might appropriately understand and respond to user questions. In contrast, Revercomb et al.[6] found that ChatGPT answered only 26.7% of otology-related questions correctly, noting that the 150 randomly selected BoardVitals items may not accurately represent the ENT Board Certification Exam. Another study of German otolaryngology board certification questions reported a 63% accuracy rate in the vestibular domain. Hoch et al.[1] suggested that these results might reflect ChatGPT's capacity for continuous improvement through user interaction, highlighting its potential as an educational tool for subspecialties in otolaryngology, including vestibular system content.

Our study identified the vestibular system as the area of greatest educational resource need in audiology. The ChatGPT demonstrated a certain level of accuracy, particularly in English responses, suggesting its potential as a complementary educational tool, although further validation is required regarding its reliability. The discrepancy between Turkish and English accuracy scores was statistically significant in the categories of the Anatomy and Physiology of the Vestibular System, Treatment and Rehabilitation of the Vestibular System, and the overall accuracy across all items, with English responses proving more accurate. This finding indicates the influence of language on AI model performance. Ando et al.[9] evaluated ChatGPT's responses to medical questions in anesthesiology in both English and Japanese, reporting adequate performance for common questions in English. Their findings emphasized the impact of the originating language on AI performance in specialized medical domains and attributed weaker outcomes in certain languages to insufficient training data. They proposed generating initial responses in English and subsequently translating them to other languages to enhance response quality in multinational medical contexts. Similarly, Ozturk et al.[10] assessed ChatGPT's Turkish and English proficiency in pharmacotherapy and observed a substantial performance gap, attributing the weaker outcomes in Turkish to linguistic limitations that reduce its effectiveness for native speakers. Consistent with these findings, our study also showed greater accuracy in English responses. Furthermore, the limited availability of Turkish educational resources, along with the dominance of medical literature in English, may constrain ChatGPT's ability to accurately process complex Turkish medical terminology. This underscores the importance of developing comprehensive multilingual training datasets to improve AI model effectiveness across different languages and dialects.

Inter-rater reliability is defined as the level of consensus between two or more raters.[11] In our study, high inter-rater reliability was found between the average scores given by nine experts for Turkish and English responses. However, no significant difference was found between the scores of the experts. This finding indicates that ChatGPT is dependable in producing responses in both languages and that the expert assessments of the accuracy of responses are consistent and similar across both languages. Consequently, this high level of agreement among experts mirrors the uniformity of the evaluation criteria and their interpretations, which makes it possible to obtain uniform results. Similar to our study, high inter-rater reliability was found in the studies in the literature in the field of otolaryngology. Lechien et al.[12] asked 25 questions regarding the definition, diagnosis and treatment of Laryngopharyngeal Reflux Disease (LPRD) to ChatGPT-4. In their study, scoring was done with a Likert-type scale by four laryngologists and high inter-rater reliability was obtained. Cheong et al.[13] created patient education materials for obstructive sleep apnea through ChatGPT and Google Bard. The raters evaluated these materials. It was found to be high agreement among rater for both AI tools. The authors concluded that these results showed a reliable potential for ChatGPT for medical and educational applications.

Test-retest reliability refers to the stability of scores across two sessions conducted under similar conditions.[14] In our study, high test-retest reliability was found between the Turkish and English responses of ChatGPT at different times. Also, no significant difference was found between the two raters. These findings suggest that ChatGPT responses do not change over time. The responses were usually seen as "neither totally the same nor completely different," suggesting some consistency but also a degree of variation. Al-Dujaili et al.[15] evaluated the consistency of ChatGPT's responses to 20 pharmacotherapy-related cases at three different time points. The responses of ChatGPT at three different time points were found to be moderately consistent, indicating that ChatGPT can understand the context and respond consistently to pharmacotherapy issues, thus demonstrating its potential to provide comprehensive pharmaceutical information. Kochanek et al.[16] evaluated the reproducibility of ChatGPT 3.5 and 4.0 responses on objective testing methods of hearing and found the agreement rate between responses on different days to be 75 to 79% for ChatGPT 3.5 and 85 to 88% for ChatGPT 4, with corresponding Cohen's Kappa values of 0.65-0.69 and 0.80-0.85, respectively in their study.

In the current study, we found no statistically significant differences between ChatGPT's average scores for open-ended and single-choice questions in Turkish and English according to the question type. Nonetheless, in all categories, higher scores were achieved on open-ended questions, with this difference being more evident in English responses. This suggests that ChatGPT's performance is constrained by the restrictive nature of single-choice items but improves in open-ended formats requiring flexibility, contextual reasoning, and more comprehensive responses. In alignment with these results, Zalzal et al.[8] reported higher accuracy for ChatGPT in open-ended questions compared to multiple-choice items in otolaryngology. In contrast, Hoch et al.[1] evaluated ChatGPT using multiple-choice and single-choice questions for otolaryngology board certification, finding superior performance in single-choice (63%) relative to multiple-choice questions (34%). Methodologically, our study differed by focusing only on single-choice and open-ended formats, again showing better performance in the latter. Overall, these findings indicate that ChatGPT's information-processing ability remains constrained across structured formats, but is enhanced when providing flexible, context-driven responses. Furthermore, Zalzal et al.[8] reported that ChatGPT's accuracy in multiple-choice tasks could improve with training data beyond 2021, highlighting the need for further research and more extensive updates to optimize performance across question types.

Nonetheless, our study has certain limitations. The questions were prepared by the research team rather than independent experts, and only two question formats were analyzed, which may restrict generalizability. Moreover, ChatGPT's knowledge base is limited to information up to 2021, potentially reducing accuracy in rapidly evolving fields. The study also assessed only version 3.5, while newer versions with enhanced capabilities, such as version 5, were not included. However, evaluating version 3.5, which was publicly accessible at the time, ensured the relevance of our findings for a broad user population. Future research should compare multiple model versions and question formats to better assess performance.

In conclusion, the findings of our study suggest that AI language models such as ChatGPT may have the potential to be a complementary educational tool, but their reliability and accuracy in specific domains such as the Vestibular system still require further validation. Moreover, it is of utmost importance to consider that originating language differences, such as between Turkish and English, as well as different question types, may influence the performance of the model in specific specializations. This finding indicates the need for a thorough evaluation of the consistency and reliability of the answers provided by ChatGPT in different languages; thus, it points out that attention should be paid to the limitations of such AI tools that may arise from language differences in clinical and academic fields.

Data Sharing Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Author Contributions: G.K., S.K.: Idea/concept, design, data collection and/or processing, analysis and/or interpretation, literature review, writing the article, references and fundings; G.K.: Control/supervision, critical review.

Conflict of Interest: The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding: The authors received no financial support for the research and/or authorship of this article.

REFERENCES

1. Hoch CC, Wollenberg B, Lüers JC, Knoedler S, Knoedler L, Frank K, 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;280:4271-8. doi: 10.1007/s00405-023-08051-4.
2. Park I, Joshi AS, Javan R. Potential role of ChatGPT in clinical otolaryngology explained by ChatGPT. Am J Otolaryngol 2023;44:103873. doi: 10.1016/j.amjoto.2023.103873.
3. Frosolini A, Franz L, Benedetti S, Vaira LA, de Filippis C, Gennaro P, et al. Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. Eur Arch Otorhinolaryngol 2023;280:5129-33. doi: 10.1007/s00405-023-08205-4.

4. Gritti MN, AlTurki H, Farid P, Morgan CT. Progression of an Artificial Intelligence Chatbot (ChatGPT) for pediatric cardiology educational knowledge assessment. Pediatr Cardiol 2024;45:309-13. doi: 10.1007/s00246-023-03385-6.
5. Leng L. Challenge, integration, and change: ChatGPT and future anatomical education. Med Educ Online 2024;29:2304973. doi: 10.1080/10872981.2024.2304973.
6. Revercomb L, Patel AM, Choudhry HS, Filimonov A. Performance of ChatGPT in otolaryngology knowledge assessment. Am J Otolaryngol 2024;45:104082. doi: 10.1016/j.amjoto.2023.104082.
7. Nielsen JPS, von Buchwald C, Grønhøj C. Validity of the large language model ChatGPT (GPT4) as a patient information source in otolaryngology by a variety of doctors in a tertiary otorhinolaryngology department. Acta Otolaryngol 2023;143:779-82. doi: 10.1080/00016489.2023.2254809.
8. Zalzal HG, Cheng J, Shah RK. Evaluating the current ability of ChatGPT to assist in professional otolaryngology education. OTO Open 2023;7:e94. doi: 10.1002/oto2.94.
9. Ando K, Sato M, Wakatsuki S, Nagai R, Chino K, Kai H, et al. A comparative study of English and Japanese ChatGPT responses to anaesthesia-related medical questions. BJA Open 2024;10:100296. doi: 10.1016/j.bjao.2024.100296.
10. Ozturk N, Yakak I, Ağ MB, Aksoy N. Is ChatGPT reliable and accurate in answering pharmacotherapy-related inquiries in both Turkish and English? Curr Pharm Teach Learn 2024;16:102101. doi: 10.1016/j.cptl.2024.04.017.
11. Lange RT. Inter-rater reliability. In: Kreutzer JS, DeLuca J, Caplan B, editors. Encyclopedia of Clinical Neuropsychology. New York: Springer; 2011. p. 1467-9.
12. Lechien JR, Carroll TL, Huston MN, Naunheim MR. ChatGPT-4 accuracy for patient education in laryngopharyngeal reflux. Eur Arch Otorhinolaryngol 2024;281:2547-52. doi: 10.1007/s00405-024-08560-w.
13. Cheong RCT, Unadkat S, Mcneillis V, Williamson A, Joseph J, Randhawa P, et al. Artificial intelligence chatbots as sources of patient education material for obstructive sleep apnoea: ChatGPT versus Google Bard. Eur Arch Otorhinolaryngol 2024;281:985-93. doi: 10.1007/s00405-023-08319-9.
14. Nguyen R, Brooks M, Bruno R, Peacock A. Behavioral measures of state impulsivity and their psychometric properties: a systematic review. Pers Individ Differ 2018;135:67-79.
15. Al-Dujaili Z, Omari S, Pillai J, Al Faraj A. Assessing the accuracy and consistency of ChatGPT in clinical pharmacy management: A preliminary analysis with clinical pharmacy experts worldwide. Res Social Adm Pharm 2023;19:1590-4. doi: 10.1016/j.sapharm.2023.08.012.
16. Kochanek K, Skarzynski H, Jedrzejczak WW. Accuracy and repeatability of ChatGPT based on a set of multiple-choice questions on objective tests of hearing. Cureus 2024;16:e59857. doi: 10.7759/cureus.59857.