Original Paper

# Comparison of ChatGPT and DeepSeek on a Standardized Audiologist Qualification Examination in Chinese: Observational Study

Beier Qi[1], MS; Yan Zheng[1], BS; Yuanyuan Wang[1], BS; Li Xu[2], MD, PhD

[1]Beijing Tongren Hospital, Capital Medical University, Key Laboratory of Otolaryngology - Head and Neck Surgery (Capital Medical University), Ministry of Education, Beijing, China, Beijing, China

[2]Department of Hearing, Speech and Language Sciences, Ohio University, Athens, OH, United States

**Corresponding Author:**
Li Xu, MD, PhD
Department of Hearing, Speech and Language Sciences
Ohio University
20 E Union St
Athens, OH, 45701
United States
Phone: 1 740 593 0310
Fax: 1 740 593 0287
Email: xul@ohio.edu

## Abstract

**Background:** Generative artificial intelligence (GenAI), exemplified by ChatGPT and DeepSeek, is rapidly advancing and reshaping human-computer interaction with its growing reasoning capabilities and broad applications across fields such as medicine and education.

**Objective:** This study aimed to evaluate the performance of 2 GenAI models (ie, GPT-4-turbo and DeepSeek-R1) on a standardized audiologist qualification examination in Chinese and to explore their potential applicability in audiology education and clinical training.

**Methods:** The 2024 Taiwan Audiologist Qualification Examination, comprising 300 multiple-choice questions across 6 subject areas (ie, basic hearing science, behavioral audiology, electrophysiological audiology, principles and practice of hearing devices, health and rehabilitation of the auditory and balance systems, and hearing and speech communication disorders [including professional ethics]), was used to assess the performance of the 2 GenAI models. The complete answering process and reasoning paths of the models were recorded, and performance was analyzed by overall accuracy, subject-specific scores, and question-type scores. Statistical comparisons were performed at the item level using the McNemar test.

**Results:** ChatGPT and DeepSeek achieved overall accuracies of 80.3% (241/300) and 79.3% (238/300), respectively, which are higher than the passing criterion of the Taiwan Audiologist Qualification Examination (ie, 60% correct answers). The accuracies for the 6 subject areas were 88% (44/50), 70% (35/50), 86% (43/50), 76% (38/50), 82% (41/50), and 80% (40/50) for ChatGPT and 82% (41/50), 72% (36/50), 78% (39/50), 80% (40/50), 80% (40/50), and 84% (41/50) for DeepSeek. No significant differences were found between the two models at the item level (overall *P*=.79), with a small effect size (accuracy difference=+1%, Cohen *h*=0.02, odds ratio 0.90, 95% CI 0.53–1.52) and substantial agreement (κ=0.71). ChatGPT scored highest in basic hearing science (88%), whereas DeepSeek performed the best in hearing and speech communication disorders (84%). Both models scored lowest in behavioral audiology (ChatGPT: 70% and DeepSeek: 72%). Question-type analysis revealed that both models performed well on reverse logic questions (ChatGPT: 79/95, 83%; DeepSeek: 80/95, 84%) but performed moderately on complex multiple-choice questions (ChatGPT: 9/17, 53%; DeepSeek: 11/17, 65%). However, both models performed poorly on graph-based questions (ChatGPT: 2/11, 18%; DeepSeek: 4/11, 36%).

**Conclusions:** Both GenAI models demonstrated strong professional knowledge and stable reasoning ability, meeting the basic requirements of clinical audiologists and suggesting their potential as supportive tools in audiology education. However, the presence of errors underscores the need for cautious use under educator supervision. Future research should explore their performance in open-ended, real-world clinical scenarios to assess practical applicability and limitations.

*(JMIR Form Res 2025;9:e79534)* doi: 10.2196/79534

**KEYWORDS**

generative artificial intelligence; artificial intelligence; AI; ChatGPT; DeepSeek; audiology; medical education

## Introduction

Artificial intelligence (AI), first conceptualized in 1956 [1], has made remarkable progress over the past decade in simulating human cognition and behavior, driving breakthroughs across multiple fields [2]. By constructing brain-inspired computing architectures, AI facilitates automated analysis and decision-making in complex tasks [3]. In recent years, generative AI (GenAI, a subfield of AI that uses generative models to produce text, images, videos, or other forms of data), a leading subfield of AI, has rapidly advanced, driven by algorithmic innovations and the exponential growth of data. Powered by large-scale language models (a language model trained with self-supervised machine learning on a vast amount of text, designed for natural language processing tasks, especially language generation), GenAI has demonstrated broad application potential across medicine, education, scientific research, commerce, and industry [4]. As an evolving form of AI equipped with logical reasoning capabilities, GenAI is fundamentally reshaping the way humans interact with computers.

In the field of education, the leading models of GenAI such as ChatGPT (OpenAI), Gemini Ultra (Google), Meta Llama 3, Anthropic Claude 3, and DeepSeek (DeepSeek) can not only efficiently respond to professional questions and generate a clear problem-solving process but also simulate the thinking process of experts to assist students in academic exploration, problem-solving, and decision-making training [4-6]. GenAI's exceptional generative capabilities offer new avenues for scholarly research, increased learning, classroom improvement, and knowledge sharing, and the interest in how these tools can benefit health education is also growing. Recent studies have shown that GenAI performs well in general medical examinations, such as the United States Medical Licensing Examination, the National Pharmacy Licensing Examination, and the National Nurse Licensing Examination, and has also demonstrated expert-level reasoning ability and diagnostic accuracy in the OTO Chautauqua Otolaryngology specialty examination [7-10]. As the quality and quantity of training data largely determine GenAI's ability to understand and generate text, its outstanding performance in English contexts may not necessarily extend directly to non-English applications. Existing studies have also found that GenAI's accuracy varies across different medical examinations and fields [11,12], with particularly inconsistent performance in examinations conducted in non–English-speaking regions [13,14]. Therefore, it is essential to further explore GenAI's performance in processing non-English information, especially its application potential in Chinese, the most widely spoken first language (12.3%) and the second most spoken language (13.8%) in the world [15].

Audiology is the scientific study of hearing, balance, and related disorders. Audiologists are health care professionals who deliver patient-centered care focused on preventing, identifying, diagnosing, and treating hearing and balance disorders using evidence-based approaches. These conditions can have wide-ranging effects (ie, medical, psychological, physical, social, educational, and occupational). By offering expert and personalized care, audiologists help reduce the impact of these disorders, ultimately enhancing patients' overall well-being and quality of life. In the United States, the doctor of audiology program is typically a 3- or 4-year graduate program. Doctor of audiology students are recommended to take the Praxis Examination in Audiology before their final year of their graduate study. Passing the Praxis Examination is a requirement for the American Speech-Language-Hearing Association Certificate of Clinical Competence in Audiology and state professional licensure in almost all 50 states in the United States [16]. The Educational Testing Service is responsible for administering the Praxis Examination, and the questions in the examination are not published. In Taiwan, an equivalent examination to the Praxis Examination, titled the Taiwan Audiologist Qualification Examination (TAQE), is published online every year [17]. The TAQE is written in traditional Chinese.

Previous studies have begun to explore the feasibility of GenAI in audiology education. Wang et al [18] evaluated the performance of ChatGPT (GPT-4) on the 2023 version of TAQE, showing an overall accuracy of 75%. This study suggests that GenAI possesses a certain level of professional knowledge in audiology and has the potential to support teaching and information acquisition. However, its abilities are still limited by the accuracy and scope of its knowledge, and its feasibility needs further verification.

In this study, the 2024 version of the TAQE was used to compare the performance of GPT-4-turbo and DeepSeek-R1 in audiology professional examinations. We aimed to gain insights into the potential of GenAI models in understanding audiology questions written in Chinese and to assess their applicability in audiology education and clinical practice. At the model level, compared with GPT-4 used by Wang et al [18], GPT-4-turbo demonstrates enhanced inference efficiency and cost-effectiveness, making it more suitable for practical applications. On the other hand, DeepSeek, a Chinese-developed GenAI, is optimized for Chinese language understanding and generation, thereby enhancing its adaptability and effectiveness in Chinese contexts. In addition, it adopts a reinforcement learning framework and incorporates a series of distilled models, which maintain performance while enabling efficient deployment in resource-constrained environments [19].

On the basis of this background, we propose 2 research hypotheses:

- The primary hypothesis is that GenAI models (GPT-4-turbo and DeepSeek-R1) are capable of achieving passing-level performance on the TAQE, thereby reaching the knowledge structure and reasoning ability required of clinical

audiologists and demonstrating their potential applicability in audiology education and clinical training.
- The secondary hypothesis is that compared with the general-purpose model (GPT-4-turbo), DeepSeek-R1 is expected to achieve superior performance in the TAQE, reflecting the advantages of language-specific optimization.

By comparing ChatGPT and DeepSeek, this study not only highlights the technological strengths of both models but also enables a comprehensive evaluation of these 2 leading GenAI systems in the context of professional examinations written in Chinese, thereby shedding light on whether language-specific models (eg, DeepSeek) are superior to general models (eg, ChatGPT).

# Methods

## Ethical Considerations

This study does not involve human participants and is therefore classified as a non–human subject study. As such, it does not require institutional review board approval or informed consent. All data used in this research were derived from publicly available question banks, and the analysis was conducted exclusively using AI models. The study strictly adheres to data use guidelines and ensures full transparency in model implementation and use.

## AI Models

Two GenAI models were used in this study: GPT-4-turbo (released in April 2024) and DeepSeek-R1 (released in May 2024). GPT-4-turbo offers enhanced efficiency, a 128K-token context, and multimodal input [20]. DeepSeek-R1 provides optimized code and mathematical reasoning, a 128K-token context, and open weights for reproducibility. Although newer versions are available, these versions were the latest available GenAI models at the time of this research.

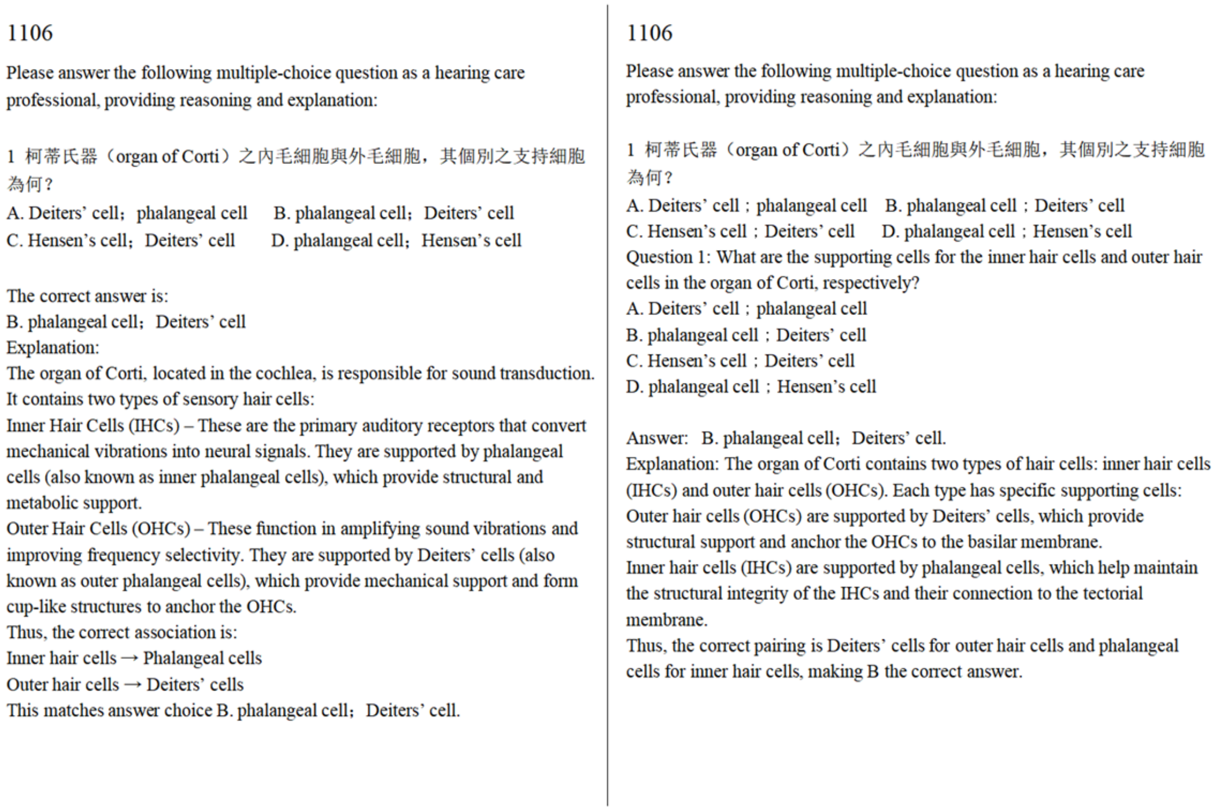
## Test Materials

The test questions were from the officially released version of the 2024 TAQE [17]. The TAQE is a professional licensing examination for audiologists and covers six subject areas: (1) basic hearing science, (2) behavioral audiology, (3) electrophysiological audiology, (4) principles and practice of hearing devices, (5) health and rehabilitation of the auditory and balance systems, and (6) hearing and speech communication disorders (including professional ethics), with 50 questions in each subject area, for a total of 300 questions. All questions were in a multiple-choice format. The examination was written in traditional Chinese. All questions and answer sheets were downloaded from the official website [17] as a PDF document.

## Procedure

We manually entered questions and options into the dialogue interface and recorded the model’s original responses, explanations, and answers. A fixed prompt, “Please answer the following multiple-choice question as a hearing care professional, providing reasoning and explanation,” was used to regulate the response style of the models and ensure the outputs were professionally relevant and interpretable (Figure 1). To ensure consistency and minimize potential human bias in the input process, all prompts were predetermined and entered by copy-and-paste rather than manual typing.

**Figure 1.** Examples of prompts and responses. Left: ChatGPT; right: DeepSeek. This example shows the first question from the basic auditory science examination. All Taiwan Audiologist Qualification Examination questions are multiple choice. The original examination questions are primarily in traditional Chinese and use bilingual (Chinese and English) terminology to express professional concepts.



1106

Please answer the following multiple-choice question as a hearing care professional, providing reasoning and explanation:

1 柯蒂氏器（organ of Corti）之內毛細胞與外毛細胞，其個別之支持細胞為何？

A. Deiters' cell；phalangeal cell B. phalangeal cell；Deiters' cell

C. Hensen's cell；Deiters' cell D. phalangeal cell；Hensen's cell

The correct answer is:

B. phalangeal cell；Deiters' cell

Explanation:

The organ of Corti, located in the cochlea, is responsible for sound transduction. It contains two types of sensory hair cells:

Inner Hair Cells (IHCs) – These are the primary auditory receptors that convert mechanical vibrations into neural signals. They are supported by phalangeal cells (also known as inner phalangeal cells), which provide structural and metabolic support.

Outer Hair Cells (OHCs) – These function in amplifying sound vibrations and improving frequency selectivity. They are supported by Deiters' cells (also known as outer phalangeal cells), which provide mechanical support and form cup-like structures to anchor the OHCs.

Thus, the correct association is:

Inner hair cells → Phalangeal cells

Outer hair cells → Deiters' cells

This matches answer choice B. phalangeal cell；Deiters' cell.

1106

Please answer the following multiple-choice question as a hearing care professional, providing reasoning and explanation:

1 柯蒂氏器（organ of Corti）之內毛細胞與外毛細胞，其個別之支持細胞為何？

A. Deiters' cell；phalangeal cell B. phalangeal cell；Deiters' cell

C. Hensen's cell；Deiters' cell D. phalangeal cell；Hensen's cell

Question 1: What are the supporting cells for the inner hair cells and outer hair cells in the organ of Corti, respectively?

A. Deiters' cell；phalangeal cell

B. phalangeal cell；Deiters' cell

C. Hensen's cell；Deiters' cell

D. phalangeal cell；Hensen's cell

Answer: B. phalangeal cell；Deiters' cell.

Explanation: The organ of Corti contains two types of hair cells: inner hair cells (IHCs) and outer hair cells (OHCs). Each type has specific supporting cells:

Outer hair cells (OHCs) are supported by Deiters' cells, which provide structural support and anchor the OHCs to the basilar membrane.

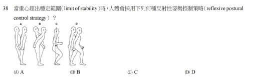
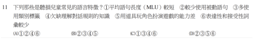
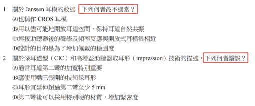
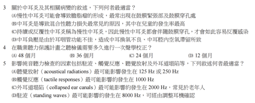
Inner hair cells (IHCs) are supported by phalangeal cells, which help maintain the structural integrity of the IHCs and their connection to the tectorial membrane.

Thus, the correct pairing is Deiters' cells for outer hair cells and phalangeal cells for inner hair cells, making B the correct answer.

For graph-based questions, both the graph and the text were provided. Although the original test questions were in traditional Chinese, certain scientific terms (eg, the organ of Corti) were expressed in both Chinese and English. Both GenAI models were instructed to respond in English to minimize the impact of language differences. Additionally, the source of the test questions was concealed before testing to prevent the models from drawing on specific training data or external knowledge sources. ChatGPT completed all tests first (February 3-13, 2025), followed by DeepSeek (February 14-24, 2025).

To facilitate detailed error and reasoning analysis, the original questions were classified into 4 categories: graph based, complex, reverse logic, and standard (Table 1).

**Table 1.** Description of question categories.

| Category | Definition in this study | Questions, n (%) | Example |
|---|---|---|---|
| Graph-based questions | Graph in the question or answer | 11 (3.7) | <br>38 當重心超出穩定範圍（limit of stability）時，人體會採用下列何種反射性姿勢控制策略（reflexive postural control strategy）？<br>A B C D<br>(A) A (B) B (C) C (D) D |
| Complex multiple-choice questions | Each option consists of multiple knowledge points | 17 (5.7) | <br>11 下列那些是聽損兒童常見的語言特徵？①平均語句長度（MLU）較短 ②較少使用被動語句 ③多使用類別標籤 ④欠缺理解對話規則的知識 ⑤用道具玩角色扮演遊戲的能力差 ⑥表達性和接受性詞彙較少<br>(A)①②④⑥ (B)②③④⑤ (C)①③④⑥ (D)②③⑤⑥ |
| Reverse logic questions | Negative expressions in the questions, such as "wrong," "avoid," "inappropriate," or "not including" | 95 (31.7) | <br>1 關於 Janssen 耳模的敘述，下列何者最不適當？<br>(A)也稱作 CROS 耳模<br>(B)用以儘可能地開放耳道空間，保持耳道自然共振<br>(C)連接助聽器後的聲學及頻率反應與開放式耳模很相近<br>(D)設計的目的是為了增加佩戴的穩固度<br>2 關於深耳道型（CIC）和高增益助聽器取耳形（impression）技術的描述，下列何者錯誤？<br>(A)通常耳道第二彎的加寬特別重要<br>(B)應使用嘴巴張開的技術採耳形<br>(C)耳形宜延伸超過第二彎至少 5 mm<br>(D)第二彎後可以採用特別硬的材質，增加緊密度 |
| Standard multiple-choice questions | Others | 177 (59) | <br>3 關於中耳炎及其相關病變的敘述，下列何者最適當？<br>(A)慢性中耳炎可能會導致膽脂瘤的形成，最常出現在鼓膜緊張部及鼓膜穿孔處<br>(B)中耳炎是導致混合性聽力損失最常見的原因，其中在兒童的發生率最高<br>(C)持續或反覆性中耳炎稱為慢性中耳炎，因此慢性中耳炎都會伴隨鼓膜穿孔，才會如此容易反覆感染<br>(D)中耳負壓是由於耳咽管功能不佳，造成中耳換氣不良、中耳腔內空氣滯留所致<br>4 在職業聽力保護計畫之聽檢儀需要多久進行一次聲學校正？<br>(A) 48 個月 (B) 36 個月 (C) 24 個月 (D) 12 個月<br>5 影響純音聽力檢查的因素包括駐波、觸覺反應、聽覺放射及外耳道塌陷等，下列敘述何者最適當？<br>(A)聽覺放射（acoustical radiations）最可能影響發生在 125 Hz 或 250 Hz<br>(B)觸覺反應（tactile responses）最可能影響約發生在 1000 Hz<br>(C)外耳道塌陷（collapsed ear canals）最可能影響約發生在 2000 Hz，常見於老年人<br>(D)駐波（standing waves）最可能影響約發生在 8000 Hz，可經由調整耳機確認 |

This study categorized the original questions into 4 categories, namely, diagram-based questions, complex multiple-choice questions, reverse logic questions, and standard multiple-choice questions, for detailed analysis. Table 1 provides definitions, percentages, and representative examples to illustrate the visual and conceptual differences between the different question types.

### Statistical Methods

Descriptive statistics were performed overall and by subject areas. Model performance was compared at the item level. The McNemar test was used to assess whether the 2 models differed in accuracy on the same items. Effect sizes were expressed as absolute differences in accuracy, odds ratios, and Cohen *h*. Kappa tests were used to test consistency. Statistical significance was set at *P*<.05. All analyses were performed using R (version 4.2.1).

## Results

Both GenAI models achieved test scores exceeding the qualification examination's passing criterion of 60% accuracy, demonstrating overall strong performance. ChatGPT attained an overall score of 80%, with subject-specific scores ranging from 70% to 88%. DeepSeek achieved an overall score of 79%, with subject-specific scores ranging from 72% to 84%, as shown in Figure 2. McNemar test, applied at the item level within each of the 6 subject areas (50 questions each) and in the pooled analysis across all 300 questions, showed no significant differences in accuracy between the 2 models (*P*=.34-.99; overall *P*=.79). The overall effect size was small (accuracy difference=+1%; Cohen *h*=0.02; odds ratio 0.90, 95% CI 0.53–1.52). Cohen κ indicated substantial agreement between the 2 models' responses (κ=0.71, range 0.56-0.78), as presented in Table 2. These results indicate that the 2 models exhibited substantial response consistency despite minor fluctuations, suggesting highly similar overall capabilities in the audiologist qualification examination.

**Figure 2.** Performance on the 2024 Taiwan Audiologist Qualification Examination by ChatGPT and DeepSeek in the 6 subject areas (ie, basic hearing science, behavioral audiology, electrophysiological audiology, principles and practice of hearing devices, health and rehabilitation of the auditory and balance systems, and hearing and speech communication disorders [including professional ethics]) and the overall scores.

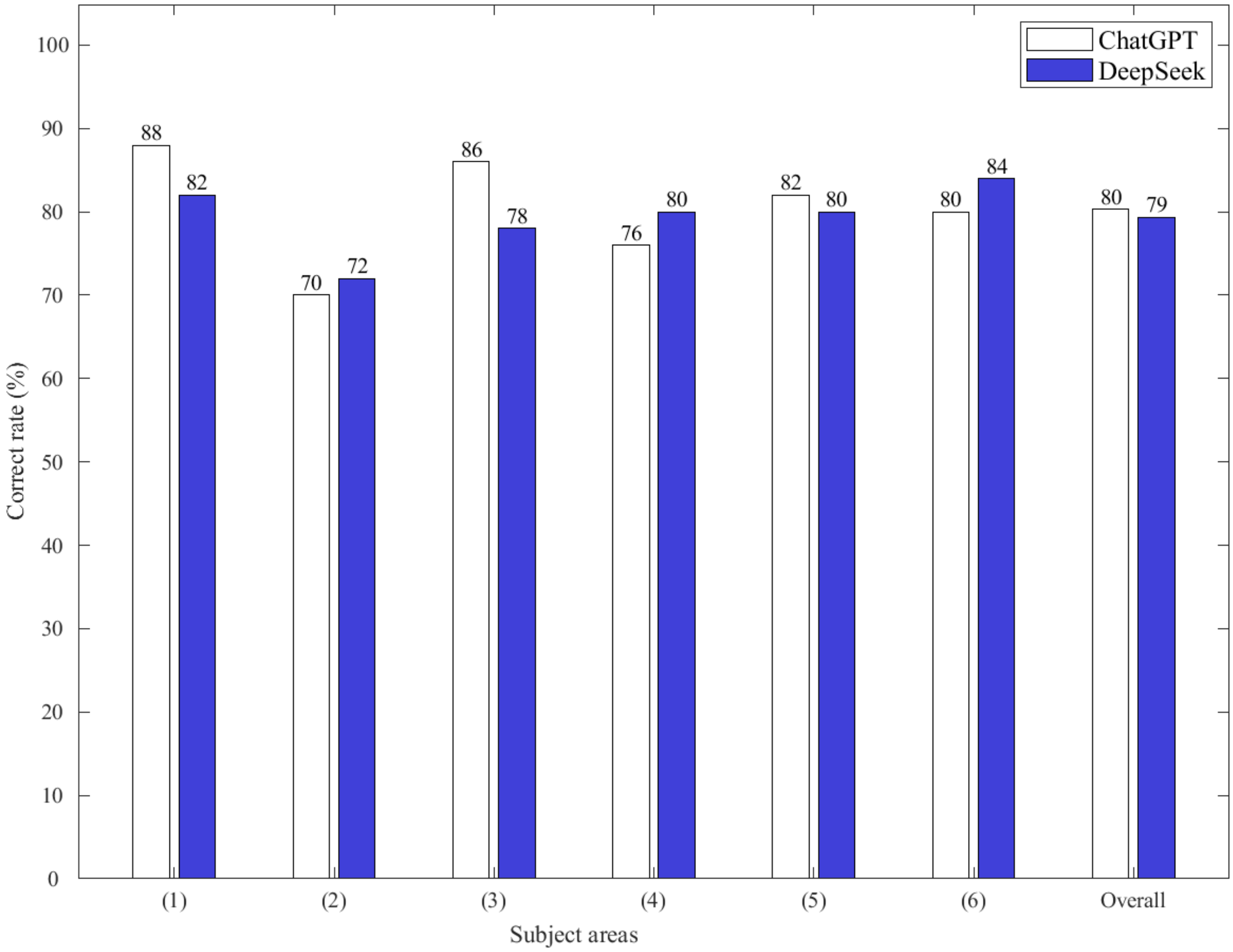


**Table 2.** Model capability comparison between ChatGPT and DeepSeek.

| Subject area | GPT_ACC[a] | DS_ACC[b] | ACC_diff[c] | Cohen *h* | OR[d] | McNemar *P* value | κ |
|---|---|---|---|---|---|---|---|
| S1 | 0.88 | 0.82 | +0.06 | 0.17 | 0.40 (0.08–2.06) | .45 | 0.78 |
| S2 | 0.70 | 0.72 | −0.02 | −0.04 | 1.17 (0.39–3.47) | .99 | 0.56 |
| S3 | 0.86 | 0.78 | +0.08 | 0.21 | 0.43 (0.11–1.66) | .34 | 0.67 |
| S4 | 0.76 | 0.80 | −0.04 | −0.10 | 2.00 (0.37–10.92) | .69 | 0.76 |
| S5 | 0.82 | 0.80 | +0.02 | 0.05 | 0.80 (0.21–2.98) | .99 | 0.72 |
| S6 | 0.80 | 0.84 | −0.04 | −0.10 | 1.50 (0.42–5.32) | .75 | 0.70 |
| Overall | 0.803 | 0.793 | +0.01 | 0.02 | 0.90 (0.53–1.52) | .79 | 0.71 |

[a] GPT_ACC: accuracy for ChatGPT.

[b] DS_ACC: accuracy for DeepSeek.

[c] ACC_diff: accuracy difference between GPT_ACC and DS_ACC.

[d] OR: odds ratio.

Subject-specific analysis showed that ChatGPT achieved its highest score in basic hearing science (88%), whereas DeepSeek performed best in hearing and speech communication disorders (including professional ethics; 84%). However, both models recorded their lowest scores in behavioral audiology, with ChatGPT at 70% and DeepSeek at 72% (Figure 2).

Question-type analysis revealed that both models underperformed on graph-based questions, with DeepSeek scoring 4 out of 11 (36% correct) and ChatGPT scoring 2 out of 11 (18% correct), both below the passing threshold. For complex multiple-choice questions, both approached the passing level, with DeepSeek scoring 11 out of 17 (65% correct) and ChatGPT scoring 9 out of 17 (53% correct). In reverse logic questions, both models performed well at comparable levels, with DeepSeek scoring 80 out of 95 (84% correct) and ChatGPT scoring 79 out of 95 (83% correct).

## Discussion

This study evaluated the performance of 2 GenAI models in a standardized audiologist qualification examination in Chinese. The results showed that (1) both GenAI models passed the qualification examination, indicating that they met the basic requirements of clinical audiologists and possessed the necessary audiological knowledge, and (2) the scores of the GenAI models varied across subjects, with the lowest performance in behavioral audiology. Performance also differed by question type: both models performed poorly on graph-based questions but showed good performance on reverse logic questions and (3) no overall advantage of DeepSeek over ChatGPT was observed, suggesting that language-specific optimization did not translate into a clear superiority in the audiology examination written in Chinese.

This study compared the performance of 2 GenAI models in a real audiologist qualification examination. This examination covers the core areas of audiology, fully reflects the breadth and depth of audiology, and is an important reference for evaluating audiology professional capabilities. The basic requirements for candidates include a degree in audiology or a related field and at least 6 months or 375 hours of clinical practice experience [21]. The passing criterion of the examination has been set at 60% correct, and the pass rate for the 2024 TAQE is 87.61%.

Our test program simulated the process of the candidate's answer without external intervention. The overall accuracy of ChatGPT and DeepSeek were 80% and 79%, respectively. The results show that the 2 GenAI models can stably cope with the standardized (multiple-choice) audiology examination, have good audiology knowledge, and meet the basic standards for clinical audiologists. At the same time, the 2 GenAI models can provide a complete thinking process and clearly show the reasoning chain for solving problems, which makes them potentially useful in audiology education.

In this study, the performance of GenAI on the audiology examination (ie, a non-English examination) was similar to that of other professional examinations in Chinese, including the National Medical Qualification Examination, the National Pharmacist Qualification Examination, and the National Nursing Qualification Examination, with scores ranging from 73% to 84% using GPT-4.0 [10,22-24]. Although these scores are still lower than those of the English version, they are a significant improvement over GPT-3.5, indicating that GenAI is gradually adapting to the non-English examination environment.

In addition, Wang et al [18] evaluated the performance of GPT-4 on the 2023 TAQE, reporting an overall accuracy of 75%. In this study, testing on the 2024 TAQE, GPT-4-turbo (80% correct) and DeepSeek-R1 (79% correct) both outperformed the previous version of ChatGPT. Before concluding that improvements in the GenAI model led to better performance in audiology, we must consider that the difficulty level of the tests may differ between years. We reevaluated GPT-4-turbo on the 2023 TAQE and obtained a score of 76%, which was not significantly different from the score reported by Wang et al [18]. Thus, we do not have enough evidence to support the notion that the performance improvement is likely due to the improved GenAI model itself. To ensure comparability, this study followed the same approach as Wang et al [18], instructing the AI model to generate responses only in English and manually reviewing the accuracy of all outputs. Although no comprehension or accuracy issues attributable to bilingual processing were observed, potential bias arising from cross-lingual input and output should be carefully considered in future AI evaluation studies.

Both models scored lowest on behavioral audiology, likely due to the complexity of the questions that combine theoretical knowledge with clinical case analysis. For example, the basic knowledge questions are as follows:

> *Question 2.* 聽力測試前進行耳鏡檢查之注意事項，下列何者錯誤*?* [Which of the following is incorrect regarding precautions for otoscopic examination before hearing tests?]
>
> *Question 4.* 在職業聽力保護計畫中，聽檢儀需多久進行一次聲學校正*?* [In a hearing conservation program, how often should the audiometer be acoustically calibrated?]

Furthermore, the clinical case–based questions are as follows:

> *Question 18* 針對雙側不對稱感音神經性聽力損失病人，根據語音辨識閾（*SRT*）及舒適聽閾推算遮蔽音設定. [For a patient with bilateral asymmetric sensorineural hearing loss, how should masking levels be set based on the speech recognition threshold (SRT) and the most comfortable listening level?]
>
> *Question 24:* 針對有非器質性聽損嫌疑的患者，從聽力檢查中辨別其表現特徵. [For a patient suspected of nonorganic hearing loss, how can audiological tests be used to identify characteristic patterns?]

Compared to the structured questions relying on memorization or definitions, GenAI models are prone to reasoning biases when dealing with highly interactive and context-dependent questions. Additionally, the behavioral audiology contained the highest number of graph-based questions (5 graphs for 6 questions), which are the question types where both GenAI models performed the worst. These factors might contribute to the poor performance of both GenAI models in this module.

Meanwhile, the poor performance in graph-based questions is consistent with the findings of Wang et al [18] and may be partly due to the low image quality and resolution of the TAQE materials. These results suggest that current GenAI still relies on high-quality original images for effective graphical processing and also highlight limitations in multimodal processing, particularly in image-text integration and visual semantic understanding. However, the results of this study and Wang et al [18] were both lower than the 90.7% accuracy in the United States Medical Licensing Examination test reported by Yang et al [25]. This discrepancy may be attributed to language differences in the test materials [26,27]. In the first 2 studies, the stems of graph-based questions were presented in Chinese, whereas Yang et al [25] used English. Previous studies have shown that language effects are particularly evident in tasks requiring specific cultural background knowledge or relying on the latest literature and data in other languages [28], which may further account for the differences in model performance. In contrast, both GenAI models performed better on the reverse logic questions, demonstrating their strengths in causal reasoning and decision-support tasks.

DeepSeek performed on par with ChatGPT in the audiologist qualification examination written in Chinese (79% vs 80%). As a GenAI model optimized for the Chinese language, DeepSeek-R1 significantly increased its coverage of Chinese in the training corpus and adopted alignment strategies specifically tailored for Chinese during fine-tuning, making it more targeted in Chinese language applications [29].

One would expect that DeepSeek might demonstrate superior performance in professional tasks in Chinese contexts. Our results do not support such a conjecture. It is possible that a vast majority of literature and textbooks on audiology are written in English. The amount of text information on audiology written in Chinese is only a fraction of that in English. Therefore, optimized processing ability in Chinese would not benefit the overall performance of the GenAI in the field of audiology.

This study conducted a preliminary exploration of the application of GenAI in a standardized audiology test written in Chinese. A few limitations were noted. First, all test questions were multiple-choice questions, lacking open-ended questions. In actual clinical practice, audiologists need to not only master discrete knowledge points but also develop integrative reasoning skills to solve complex problems. Therefore, future evaluations should include open-ended questions to assess the ability of GenAI to handle actual clinical scenarios. Second, due to limited access to standardized audiology test materials, this study only examined the TAQE, whose content and format may differ from those used in other countries or regions. In addition, the official TAQE website only reports the annual number of examinees who passed and the overall passing rate, without providing category-specific performance data. These factors may limit the generalizability of the findings in a broader international context and prevent us from further comparing the performance differences between GenAI and human test takers across specific domains. Future studies with access to more granular category–level results would enable a more comprehensive evaluation of similarities and differences between GenAI and human candidates. Third, GenAI is a continuously evolving and updating tool. The versions used in this study may have been updated by the time of publication, so the results may not fully reflect the state-of-the-art performance of the GenAI models.

In conclusion, ChatGPT and DeepSeek demonstrated good professional knowledge and problem-solving ability in the standardized audiologist qualification examination. They also provide a complete reasoning process, enabling teachers and students to quickly grasp key concepts and thought processes, demonstrating their potential as supplementary tools for audiology education. Furthermore, AI's superior ability to process massive data and accumulate knowledge allows it to be updated in real time, thus keeping pace with the evolving new technologies and methods in audiology, highlighting its value as a supportive tool for audiology education. However, GenAI responses still contain a certain error rate, and thus, their use in teaching requires careful judgment and guidance from educators. In addition, the performance of GenAI in open-ended tasks has not yet been fully validated, and its ability to foster humanistic qualities such as empathy and communication remains limited. Future research should prioritize incorporating open-ended tasks based on real-world scenarios to systematically evaluate the potential and limitations of generative AI in audiology education. Moreover, this study primarily focused on general-purpose models without fine-tuning or domain adaptation. It is reasonable to expect that domain-specific training could further enhance performance, which provides an important direction for future work.

## Acknowledgments

This study was supported in part by the China Scholarship Council.

## Data Availability

All data generated or analyzed during this study are included in this published paper and Multimedia Appendices 1-3.

## Authors' Contributions

LX led the study design, supervised the research process, and was responsible for manuscript revision. BQ assisted in the study design, performed data analysis, and drafted the manuscript. YW and YZ were responsible for data collection.

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

ChatGPT transcripts.
[PDF File (Adobe PDF File), 1105 KB-Multimedia Appendix 1]

## Multimedia Appendix 2

DeepSeek transcripts.
[PDF File (Adobe PDF File), 1204 KB-Multimedia Appendix 2]

## Multimedia Appendix 3

Keys to the 2024 Taiwan Audiologist Qualification Examination.
[PDF File (Adobe PDF File), 260 KB-Multimedia Appendix 3]

## References

1. Haleem A, Javaid M, Khan I. Current status and applications of artificial intelligence (AI) in medical field: an overview. Curr Med Res Pract. Nov 2019;9(6):231-237. [FREE Full text] [doi: 10.1016/j.cmrp.2019.11.005]
2. Holzinger A, Keiblinger K, Holub P, Zatloukal K, Müller H. AI for life: trends in artificial intelligence for biotechnology. New Biotechnol. May 2023;74:16-24. [FREE Full text] [doi: 10.1016/j.nbt.2023.02.001]
3. Montejo-Ráez A, Jiménez-Zafra SM. Current approaches and applications in natural language processing. Appl Sci. May 11, 2022;12(10):4859. [doi: 10.3390/app12104859]
4. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. J Med Syst. Feb 17, 2024;48(1):22. [FREE Full text] [doi: 10.1007/s10916-024-02045-3] [Medline: 38366043]
5. Zhang K, Meng X, Yan X, Ji J, Liu J, Xu H, et al. Revolutionizing health care: the transformative impact of large language models in medicine. J Med Internet Res. Jan 07, 2025;27:e59069. [FREE Full text] [doi: 10.2196/59069] [Medline: 39773666]
6. Peng L, Luo G, Zhou S, Chen J, Xu Z, Sun J, et al. An in-depth evaluation of federated learning on biomedical natural language processing for information extraction. NPJ Digit Med. May 15, 2024;7(1):127. [FREE Full text] [doi: 10.1038/s41746-024-01126-4] [Medline: 38750290]
7. Mihalache A, Huang RS, Popovic MM, Muni RH. ChatGPT-4: an assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. Med Teach. Mar 2024;46(3):366-372. [doi: 10.1080/0142159X.2023.2249588] [Medline: 37839017]
8. Jin HK, Lee HE, Kim E. Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis. BMC Med Educ. Sep 16, 2024;24(1):1013. [FREE Full text] [doi: 10.1186/s12909-024-05944-8] [Medline: 39285377]
9. Ramchandani R, Guo E, Mostowy M, Kreutz J, Sahlollbey N, Carr MM, et al. Comparison of ChatGPT-4, Copilot, Bard and Gemini Ultra on an otolaryngology question bank. Clin Otolaryngol. Jul 2025;50(4):704-711. [doi: 10.1111/coa.14302] [Medline: 40081900]

10. Wu Z, Gan W, Xue Z, Ni Z, Zheng X, Zhang Y. Performance of ChatGPT on nursing licensure examinations in the United States and China: cross-sectional study. JMIR Med Educ. Oct 03, 2024;10:e52746. [FREE Full text] [doi: 10.2196/52746] [Medline: 39363539]
11. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. BJOG. Feb 2024;131(3):378-380. [FREE Full text] [doi: 10.1111/1471-0528.17641] [Medline: 37604703]
12. Alfertshofer M, Hoch CC, Funk PF, Hollmann K, Wollenberg B, Knoedler S, et al. Sailing the seven seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. Ann Biomed Eng. Jun 08, 2024;52(6):1542-1545. [FREE Full text] [doi: 10.1007/s10439-023-03338-3] [Medline: 37553555]
13. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ. Feb 14, 2024;24(1):143. [FREE Full text] [doi: 10.1186/s12909-024-05125-7] [Medline: 38355517]
14. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. J Chin Med Assoc. Jul 01, 2023;86(7):653-658. [doi: 10.1097/JCMA.0000000000000942] [Medline: 37227901]
15. The World Factbook. URL: https://www.cia.gov/the-world-factbook/countries/world/#people-and-society [accessed 2025-11-08]
16. Apply for certification in audiology. American Speech-Language-Hearing Association. URL: https://www.asha.org/certification/audcertification/ [accessed 2025-11-08]
17. Exam question inquiry platform. Ministry of Examination ROC (Taiwan). URL: https://wwwq.moex.gov.tw/exam/wFrmExamQandASearch.aspx [accessed 2024-12-10]
18. Wang S, Mo C, Chen Y, Dai X, Wang H, Shen X. Exploring the performance of ChatGPT-4 in the Taiwan audiologist qualification examination: preliminary observational study highlighting the potential of AI chatbots in hearing care. JMIR Med Educ. Apr 26, 2024;10:e55595. [FREE Full text] [doi: 10.2196/55595] [Medline: 38693697]
19. deepseek-ai / DeepSeek-R1. GitHub. URL: https://github.com/deepseek-ai/DeepSeek-R1 [accessed 2025-06-12]
20. Introducing ChatGPT. Open AI. URL: https://openai.com/blog/chatgpt [accessed 2024-10-01]
21. Regulations for examinations of specialized professionals and technical personnel. Ministry of Examination ROC (Taiwan). URL: https://wwwc.moex.gov.tw/main/ExamLaws/wfrmExamLaws.aspx?kind=3&menu_id=320&laws_id=133 [accessed 2025-11-08]
22. Tong W, Guan Y, Chen J, Huang X, Zhong Y, Zhang C, et al. Artificial intelligence in global health equity: an evaluation and discussion on the application of ChatGPT, in the Chinese National Medical Licensing Examination. Front Med (Lausanne). 2023;10:1237432. [FREE Full text] [doi: 10.3389/fmed.2023.1237432] [Medline: 38020160]
23. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inform. Sep 2023;177:105173. [doi: 10.1016/j.ijmedinf.2023.105173] [Medline: 37549499]
24. Fang C, Wu Y, Fu W, Ling J, Wang Y, Liu X, et al. How does ChatGPT-4 preform on non-English national medical licensing examination? An evaluation in Chinese language. PLOS Digit Health. Dec 1, 2023;2(12):e0000397. [FREE Full text] [doi: 10.1371/journal.pdig.0000397] [Medline: 38039286]
25. Yang Z, Yao Z, Tasmin M, Vashisht P, Jang WS, Wang B, et al. Performance of multimodal GPT-4V on USMLE with image: potential for imaging diagnostic support with explanations. medRxiv. Preprint posted online on October 26, 2023. [FREE Full text] [doi: 10.1101/2023.10.26.23297629]
26. Watari T, Takagi S, Sakaguchi K, Nishizaki Y, Shimizu T, Yamamoto Y, et al. Performance comparison of ChatGPT-4 and Japanese medical residents in the general medicine in-training examination: comparison study. JMIR Med Educ. Dec 06, 2023;9:e52202. [FREE Full text] [doi: 10.2196/52202] [Medline: 38055323]
27. Yu P, Fang C, Liu X, Fu W, Ling J, Yan Z, et al. Performance of ChatGPT on the Chinese postgraduate examination for clinical medicine: survey study. JMIR Med Educ. Feb 09, 2024;10:e48514. [FREE Full text] [doi: 10.2196/48514] [Medline: 38335017]
28. Seghier ML. ChatGPT: not all languages are equal. Nature. Mar 2023;615(7951):216. [doi: 10.1038/d41586-023-00680-3] [Medline: 36882613]
29. DeepSeek. Wikipedia. URL: https://en.wikipedia.org/wiki/DeepSeek?utm_source=chatgpt.com [accessed 2025-06-18]

## Abbreviations

**AI:** artificial intelligence
**GenAI:** generative artificial intelligence
**TAQE:** Taiwan Audiologist Qualification Examination

*Edited by A Mavragani; submitted 23.Jun.2025; peer-reviewed by J Hu, OC Balogun; comments to author 06.Sep.2025; revised version received 26.Sep.2025; accepted 20.Oct.2025; published 28.Nov.2025*

*Please cite as:*
*Qi B, Zheng Y, Wang Y, Xu L*
*Comparison of ChatGPT and DeepSeek on a Standardized Audiologist Qualification Examination in Chinese: Observational Study*
*JMIR Form Res 2025;9:e79534*
*URL: https://formative.jmir.org/2025/1/e79534*
*doi: 10.2196/79534*
*PMID:*