Turk Arch Otorhinolaryngol 2024; 62(2): 48-57

Evaluating the Performance of ChatGPT, Gemini, and Bing Compared with Resident Surgeons in the Otorhinolaryngology In-service Training Examination

Original Investigation

Utku Mete

Bursa Uludağ University Faculty of Medicine, Department of Otorhinolaryngology, Bursa, Türkiye

Abstract

Objective: Large language models (LLMs) are used in various fields for their ability to produce human-like text. They are particularly useful in medical education, aiding clinical management skills and exam preparation for residents. To evaluate and compare the performance of ChatGPT (GPT-4), Gemini, and Bing with each other and with otorhinolaryngology residents in answering in-service training exam questions and provide insights into the usefulness of these models in medical education and healthcare.

Methods: Eight otorhinolaryngology in-service training exams were used for comparison. 316 questions were prepared from the Resident Training Textbook of the Turkish Society of Otorhinolaryngology Head and Neck Surgery. These questions were presented to the three artificial intelligence models. The exam results were evaluated to determine the accuracy of both models and residents.

Results: GPT-4 achieved the highest accuracy among the LLMs at 54.75% (GPT-4 vs. Gemini p=0.002, GPT-4 vs. Bing p<0.001), followed by Gemini at 40.50% and Bing at 37.00% (Gemini vs. Bing p=0.327). However, senior residents outperformed all LLMs and other residents with an accuracy rate of 75.5% (p<0.001). The LLMs could only compete with junior residents. GPT-4 and Gemini performed similarly to juniors, whose accuracy level was 46.90% (p=0.058 and p=0.120, respectively). However, juniors still outperformed Bing (p=0.019).

Conclusion: The LLMs currently have limitations in achieving the same medical accuracy as senior and mid-level residents. However, they outperform in specific subspecialties, indicating the potential usefulness in certain medical fields.

Keywords: Artificial intelligence, ChatGPT, otorhinolaryngology, otorhinolaryngology in-service examination, resident education

Introduction

The emergence of artificial intelligence (AI) has drawn a lot of attention towards large language models (LLMs), a member of natural language processing tools. These models are highly proficient in processing and generating text resembling human speech. They are created using advanced deep-learning techniques and comprehensive datasets from the internet (1).

LLMs are sophisticated AI systems designed to understand, interpret, and generate human language in a meaningful and contextually relevant way. They can identify patterns, comprehend context, and link various pieces of information, abilities that make them capable of

ORCID IDs of the authors:

U.M 0000-0003-0902-8061

Cite this article as: *Mete U. Evaluating the Performance of ChatGPT, Gemini, and Bing Compared with Resident Surgeons in the Otorhinolaryngology in-service Training Exam.*
Turk Arch Otorhinolaryngol. 2024; 62(2): 48-57

Corresponding Author:
Utku Mete;
drutkumete@gmail.com
utkumete@uludag.edu.tr

Received Date: 2024-04-29
Accepted Date: 2024-05-07

DOI: 10.4274/tao.2024.3.5

©Copyright 2024 by Turkish Otorhinolaryngology- Head and Neck Surgery Society / Turkish Archives of Otorhinolaryngology is published by Galenos Publishing House.
Licenced under Creative Commons Attribution- NonCommercial 4.0 International (CC BY-NC 4.0).

providing insightful responses and advice on a vast range of subjects (2).

These models, including OpenAI's ChatGPT, Google's Gemini, and Microsoft's Bing, are trained using machine learning, where they learn to predict and generate text based on the patterns they observe in the training data. This allows them to perform language-related tasks like translation, summarization, and question-answering (3).

A variety of studies have explored the effectiveness of LLMs in some exams. Specific research has been conducted on other professions, such as the University of Minnesota Law School, the Bar Exam, the Wharton Master of Business Administration, and accounting exams, even without fine-tuning the pre-trained model (4-7). In the medical field, some studies have assessed the performance of ChatGPT in the United States Medical Licensing Examination (8, 9). Further, some studies compared the performance of different LLMs in various exams like Ophthalmic Knowledge Assessment Program and Neurosurgery Oral Board (10, 11).

Some studies in the field of otorhinolaryngology explored the effectiveness of ChatGPT. One study examined the success rate of ChatGPT and found that it could pass the Royal College of Physicians and Surgeons of Canada Otorhinolaryngology Board Exam (12). Another study explored the usefulness of ChatGPT in the board preparation process by examining quiz skills in various otolaryngologic subspecialties (13).

This study aimed to evaluate the performance of LLMs, namely ChatGPT, Gemini, Bing, and resident surgeons, in the otorhinolaryngology in-service examination (ORLITE). The study also intended to compare the accuracy of each model in various otorhinolaryngology topics. We believe that exploring the potential effectiveness of AI in medical education can assist medical students and potentially improve their exam performances.

Methods

Study Design

This cross-sectional observational study compared the responses of LLMs, namely, ChatGPT, Google Gemini, Microsoft Bing, and the otorhinolaryngology residents in ORLITE. The study did not require ethics committee approval as it relied solely on the question database of the university clinic, which was derived from publicly available online medical textbooks.

Development and Implementation of the ORLITE

ORLITE is an exam designed to assess the periodic competencies of residents specializing in otorhinolaryngology at a tertiary-level university hospital. The questions for the ORLITE are based on the Resident Training Textbook, which is available on the Turkish Society of Otorhinolaryngology-Head and Neck Surgery website (https://www.kbb.org.tr). The exam content is created collaboratively by five experienced and board-certified faculty members, ensuring a consensus-driven approach to the exam content. It is conducted four times in a single academic year, and each consists of 40 questions, including multiple-choice, multiple-select, free-response, and image-based questions. These questions include topics in general otorhinolaryngology, otology/neurotology, rhinology, head and neck surgery, and laryngology. They are prepared in the Turkish language. Each correct answer is rewarded with 2.5 points, with no negative marks for incorrect choices. Residents are given 40 minutes to complete the exam.

Selection of the Questions and Querying Process

Three-hundred-and-twenty questions retrieved from eight ORLITE sessions applied over the last two academic years, 2021-2022 and 2022-2023, were reviewed. Four of these questions were excluded due to incomprehensible image-based content, leaving 316 multiple-choice, multiple-selection, short-answer, and image-based questions to be presented to different LLMs separately. The querying process was conducted from February 15 to February 18, 2024, by an otorhinolaryngology specialist using the website of each model. Each question was asked individually, and the page was refreshed each time to prevent the relevant LLM from establishing connections with previous questions and forming memory. Before each question, the models were prompted with the following message: "Hello, you are a physician currently undergoing training in otorhinolaryngology. You will be answering questions related to the resident training exam conducted at the otorhinolaryngology department of the university clinic. You are only required to indicate the correct option. Are you ready?" The generated responses were marked as correct versus incorrect and recorded.

Description of LLMs and the Resident Surgeons

During the study, various chatbots capable of producing human-like responses were tested. The chatbots used in the study included the subscription-based, paid version ChatGPT, an upgraded version of ChatGPT 3.5 developed by OpenAI and released in March 2023. In addition, the Gemini, a product of Google DeepMind introduced in December 2023, and Bing Chat, which is reported to utilize ChatGPT architecture and was made available to Edge users by Microsoft in February 2023, were also employed. The 22 human participants of the study were resident surgeons who had studied in a department of the otorhinolaryngology clinic of the university hospital in the years indicated. They were divided into three groups based on their five-year specialization training periods: the first 1.5 years as junior (3rd resident), the following two years as mid-level (2nd resident), and the final 1.5 years as senior (1st resident).

There were 7 junior residents, 10 mid-level residents, and 5 senior residents in their respective groups. The success of their exams was then categorized in terms of points.

Statistical Analysis

Statistical analyses were performed to evaluate the overall success rate for each chatbot model and resident, which were calculated as the percentage of correct answers. Independent sample t-tests were applied to compare the accuracy values of chatbots and residents. Descriptive statistics such as mean and standard deviation were preferred to evaluate the performance of each exam. A significance level of $\alpha=0.05$ was set. Statistical data analysis was performed using IBM SPSS 28.0 (IBM Corp. Released 2021. IBM SPSS Statistics for Windows, Version 28.0. Armonk, NY: IBM Corp.) Statistical Software package.

Results

Evaluation of the Performance of LLMs versus Residents in the ORLITE

In the ORLITE, ChatGPT outperformed Gemini and Bing with an accuracy of 54.75%, establishing itself as the leading model ($p=0.002$ and $p<0.001$, respectively). Gemini and Bing achieved similar accuracies of 40.50% and 37.00%, respectively, with a non-significant difference ($p=0.327$). These results highlight the exceptional comprehension and logical reasoning abilities of ChatGPT compared to other models.

The results revealed that senior residents had the highest accuracy rate of 75.50%, outperforming all LLMs and other residents ($p<0.001$). Mid-level residents showed superior results with an accuracy of 63.45% compared to Gemini ($p<0.001$) and Bing ($p<0.001$), though ChatGPT approached their performance levels ($p=0.013$). LLMs were found to be competitive only with junior residents. Junior residents achieved a success rate of 46.90%, outperforming Gemini and Bing ($p=0.019$), but could not pass ChatGPT. However, the differences in performance between junior residents and ChatGPT and between junior residents and Gemini were not statistically significant ($p=0.058$ and $p=0.120$, respectively). Gemini and Bing showed the lowest accuracy scores among them all. Table 1 summarizes, and Figure 1 illustrates, the performance of LLMs compared to residents.

T-tests were used to analyzea the accuracy differences between the LLMs among themselves and when compared with residents. Statistically, the significance level is considered as $p<0.05$.

Comparison of the Accuracy Rates of LLMs and Residents in ORLITE Per Examination

We analyzed the accuracies of eight ORLITE exams and recorded the results. The standard deviations for each model and residents were as follows: ChatGPT (7.60%), Gemini (7.30%), Bing (7.00%), senior resident (5.80%), mid-level resident (3.80%), and junior resident (7.90%). The results were more consistent among senior and mid-level residents. Figure 2 illustrates the performance of LLMs and residents per the ORLITE exam.

Investigation of the Accuracies of LLMs in Subspecialties of Otorhinolaryngology

We evaluated the performance of LLMs across various subspecialties within otorhinolaryngology, including general Ear Nose Throat (ENT), otology, rhinology, laryngology, and head and neck surgery. ChatGPT demonstrated the highest accuracy across most subspecialties, with notable performance in head and neck and rhinology, achieving accuracy rates of 59.40% and 55.60%, respectively. Gemini, on the other hand, showed consistent but more moderate success, with its highest accuracy, same as ChatGPT in head and neck and rhinology at 42.20% and 42.90%, respectively. While generally lower in all subspecialties than ChatGPT and Gemini, Bing was competitive in laryngology, where it

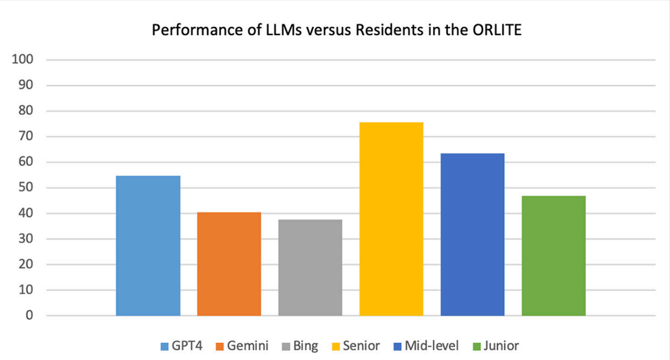

Figure 1. Performance of LLMs in comparison with residents
LLM: Large language model

Table 1. Comparative accuracy and statistical significance of LLMs and residents

	Accuracy (%)	ChatGPT p-value	Gemini p-value	Bing p-value
ChatGPT	54.75	-	0.002	<0.001
Gemini	40.50	<0.001	-	0.327
Bing	37.00	<0.001	0.327	-
1st Resident	75.50	<0.001	<0.001	<0.001
2nd Resident	63.45	0.013	<0.001	<0.001
3rd Resident	46.90	0.058	0.120	0.019

T-tests were used to analyze the accuracy differences between the LLMs and residents. Statistically, the significance level is considered as $p<0.05$

LLM: Large language model

nearly matched ChatGPT with an accuracy rate of 48.10%. Figure 3 demonstrates the accuracy rates of each model in different otorhinolaryngology fields.

Examples of ORLITE questions and corresponding responses provided by ChatGPT, Gemini, and Bing, are presented in Table 2.

Discussion

Three AI models were used in the study: OpenAI ChatGPT Plus (ChatGPT), Google Gemini, and Microsoft Bing. ChatGPT can be accessed through a subscription option, while the others are free to access on their website. It is worth noting that Bing prefers the ChatGPT infrastructure model, but Gemini works with a different type called Palm2 (14, 15). This study is the first of its kind to evaluate the performance of AI models in otorhinolaryngology exams and to compare them with humans specializing in the field.

In previous literature, there have been inconsistent results in studies comparing the performance of different LLMs in various medical fields. For instance, one study on neurosurgery oral board exam preparation question bank concluded that ChatGPT was more effective than Gemini in responding to advanced knowledge queries, and another study on answering frequently asked questions about lung cancer found that ChatGPT was more accurate than Google Gemini (11, 16). However, a study on the Royal College of Ophthalmologists fellowship exams presented a contrasting perspective. It showed that Bing Chat outperformed other AI systems, including the lowest-ranked ChatGPT (17).

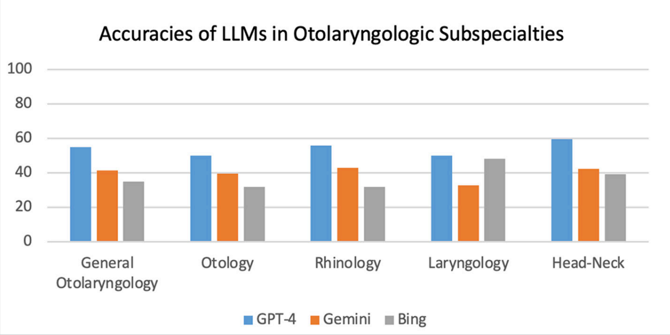

Figure 3. Accuracies of LLMs in subspecialties of otorhinolaryngology
LLM: Large language model

Previously, there have been studies examining the performance of ChatGPT in otorhinolaryngology and with various methodologies. Kuşcu et al. (18) conducted a study on the performance of ChatGPT in answering frequently asked questions about head and neck cancers. The results showed that the model had a high success rate, with 86.4% of responses being comprehensive and correct. Radulesco et al. (19) investigated the ability of ChatGPT to diagnose rhinological clinical cases accurately. They achieved a 62.5% correct or plausible response rate, and the stability of responses was moderate to high.

The performance of an LLM is mainly influenced by the model's architecture, the amount of diverse training data, the duration of training, and the allocation of resources. To improve the effectiveness of the model, it is crucial to optimize techniques and fine-tune hyperparameters during training. In fields like medicine, where knowledge bases are rapidly evolving, up-to-date and relevant training data are vital. Customization through additional training for specific tasks or sectors can help optimize performance. Additionally, the linguistic and cultural diversity of the training data affects the model's effectiveness across different languages and cultural contexts (20).

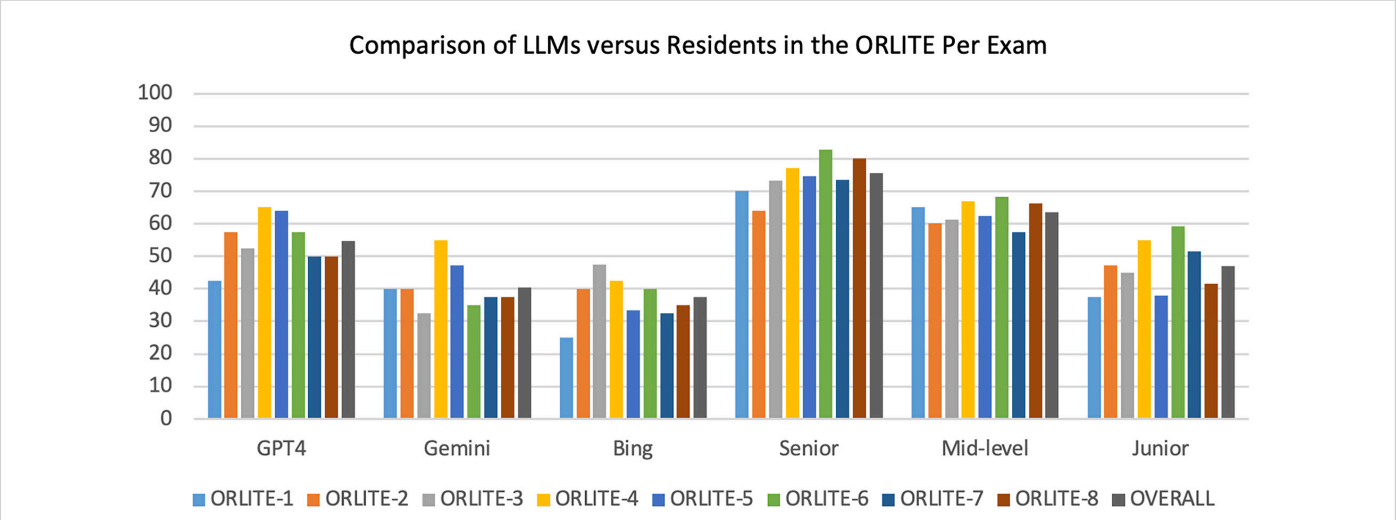

Figure 2. Comparison of LLMs versus residents in the ORLITE Per Exam
LLM: Large language model, ORLITE: Otorhinolaryngology in-service exam

Table 2. Examples of ORLITE questions and corresponding answers provided by the three LLMs

Question	**Specialty**	**ChatGPT**	**Gemini**	**Bing**
How long does it take for T lymphocytes involved in wound healing to reach their peak? a) 6 hours b) 24 hours c) 2 days **d) 7 days** e) 14 days	General ENT	**d) 7 days.** T lymphocytes, which play a role in the immune response during wound healing, typically peak activity around 7 days after injury. This is part of the proliferative phase of wound healing, where immune cells are active in clearing infection, promoting tissue growth, and supporting the repair process.	**d) 7 days.** T lymphocytes infiltrate the wound site gradually, not immediately after injury. Their numbers increase throughout the inflammatory phase and peak around 7 days after wounding.	**d) 7 days.** In the healing of wounds, the peak production time of duty T lymphocytes is between the 5th and 7th days. In this period, the inflammation stage in the healing process usually concludes, and the proliferation phase begins right after.
Which of the following is not a complication associated with the drainage of an abscess in the head and neck area? a) Aspiration b) Cranial nerve injury c) Bleeding from major arteries and veins **d) Jugular vein thrombosis** e) Sepsis	General ENT	**d) Jugular vein thrombosis.** This specific condition is generally not a direct complication of abscess drainage but can occur as a complication of infections in the neck, such as Lemierre's syndrome, which is different from the direct mechanical or procedural complications of draining an abscess.	**a) Aspiration.** Aspiration is the intended outcome of draining an abscess. It involves using a needle or surgical incision to remove the pus, the infected material within the abscess.	**a) Aspiration.** The option not related to the complications of head and neck abscess drainage is aspiration.
In stapes surgery, ideally, how many millimeters should the piston be inserted into the vestibule? a) 0.10 mm **b) 0.25 mm** c) 0.50 mm d) 0.75 mm e) 1.00 mm	Otology	**c) 0.50 mm.** This measurement is considered optimal to ensure the piston engages effectively with the inner ear structures without extending too far, which could potentially damage the sensitive structures within the inner ear, such as the cochlea.	**c) 0.50 mm.** This depth helps to balance several factors: Avoiding dislocation of the prosthesis from sudden pressure changes (e.g., sneezing). Maintaining proximity to the inner ear structures for optimal sound transmission. Surgeons aim for this depth to achieve a successful stapedotomy outcome.	**b) 0.25 mm.** There are different opinions regarding the ideal protrusion length of the piston into the vestibule in st apes surgery. However, many studies suggest that the piston should not protrude more than 0.25 mm into the vestibule.
Which of the following is considered a good prognostic factor in the prognosis of sudden hearing loss? a) Being over 60 years old **b) Hearing loss predominantly affects low frequencies** c) Hearing loss present for more than 3 months d) Accompanying vertigo e) Hearing loss having a flat configuration	Otology	**b) Hearing loss predominantly affects low frequencies.** Patients with sudden hearing loss that primarily affects low frequencies tend to have better recovery rates compared to those whose hearing loss affects all frequencies or predominantly the high frequencies. This is possible because damage limited to the low frequencies may indicate a less extensive injury to the cochlea.	**e) Hearing loss having a flat configuration.** A flat audiogram indicates a sensorineural hearing loss that affects all frequencies more or less equally. This suggests a more widespread issue potentially affecting the inner ear fluids or blood flow, which may respond better to treatment compared to a specific frequency loss.	**b) Hearing loss predominantly affects low frequencies.** The flat configuration of hearing loss (i.e., similar loss across all frequencies) is generally considered less favorable for recovery compared to low-frequency or high-frequency predominant hearing loss2. So, b) Hearing loss predominantly affecting low frequencies might be a better prognostic factor

Table 2. Continued

Question	**Specialty**	**ChatGPT**	**Gemini**	**Bing**
In a patient with nasal polyps, asthma, aspirin intolerance, and daily rhinitis symptoms, which treatment would you prefer initially? **a) Intranasal corticosteroids** b) Systemic corticosteroids c) Antibiotics d) Leukotriene antagonists e) Antifungals	Rhinology	**d) Leukotriene antagonists**. For a patient with nasal polyps, asthma, aspirin intolerance, and daily rhinitis symptoms-a combination often referred to as Samter's triad or aspirin-exacerbated respiratory disease (AERD)-the initial preferred treatment would be Leukotriene antagonists. They are particularly effective in this group because they target the pathway involved in both asthma and nasal polyp growth that is exacerbated by aspirin intolerance. This treatment helps reduce inflammation and may improve nasal and asthma symptoms.	**a) Intranasal corticosteroids.** The patient's symptoms point towards a condition called AERD, also known as Samter's triad. This involves nasal polyps, asthma, and aspirin intolerance. Intranasal corticosteroids are the mainstay of treatment for AERD because they effectively reduce inflammation in the nasal passages, improving rhinitis symptoms and potentially reducing polyp size.	**a) Intranasal corticosteroids.** Samter's Triad, also known as AERD, is a chronic condition characterized by the following three components: Asthma, sinus inflammation with recurring nasal polyps, and sensitivity to aspirin and other non-steroidal anti-inflammatory drugs. Intranasal corticosteroids are a reasonable choice for managing sinus inflammation and nasal polyps as initial treatment strategy.
In paranasal sinus neoplasias, which lesion is identified by the bending forward of the posterior wall of the maxillary sinus, known as the Holman-Miller sign in imaging? a) Osteoma b) Ameloblastoma **c) Angiofibroma** d) Papilloma e) Thornwaldt cyst	Rhinology	**c) Angiofibroma.** This sign is particularly associated with juvenile nasopharyngeal angiofibroma (JNA), which is a benign but locally aggressive vascular tumor that predominantly affects adolescent males. The Holman-Miller sign indicates the anterior bowing of the posterior wall of the maxillary sinus due to tumor pressure or growth in the nasopharynx and pterygopalatine fossa.	**c) Angiofibroma.** The Holman-Miller sign is a characteristic finding associated with JNA, a type of benign vascular tumor that frequently arises in the nasopharynx and can extend into the maxillary sinus. The expanding tumor mass causes the posterior wall of the maxillary sinus to bulge forward on imaging studies like lateral skull X-rays or cross-sectional scans.	**c) Angiofibroma.** The Holman-Miller sign, also referred to as the antral sign, is associated with JNA. It manifests as the anterior bowing of the posterior wall of the maxillary antrum on imaging studies.
According to the European Laryngological Society (ELS) classification of supraglottic laryngectomy, what type is the endolaryngeal surgery that involves the removal of the epiglottis, preepiglottic space, and one ventricular band? a) Type I b) Type IIa c) Type IIb d) Type IIIa **e) Type IIIb**	Laryngology	**c) Type IIb.** The ELS classification of supraglottic laryngectomy categorizes different types of endolaryngeal surgery based on the extent of tissue removal. In this case, the surgery described involves the removal of the epiglottis, preepiglottic space, and one ventricular band. The correct answer would be Type IIb.	**e) Type IIIb.** Considering the removal of the epiglottis, pre-epiglottic space, and one ventricular band, Type IIIb best aligns with the ELS classification the correct answer should be Type IIIb.	**e) Type IIIb.** The ELS has proposed a classification system for different types of endoscopic supraglottic partial laryngectomies. The surgery involving removal of the epiglottis, preepiglottic space, and one ventricular band corresponds to Type IIIb in the ELS classification.

Table 2. Continued

Question	**Specialty**	**ChatGPT**	**Gemini**	**Bing**
Which of the following vocal cord vibration modes is the most efficient and commonly used in daily life? a) Glottal fry **b) Modal register** c) Falsetto d) Resonance e) Articulation	Laryngology	**b) Modal register.** This is the normal vocal register used in daily conversation. It involves the vibration of the vocal folds in a regular and relatively relaxed manner, producing most of the sounds used in speech.	**b) Modal register.** Modal register refers to the normal speaking voice. It produces clear and efficient sound by allowing the vocal cords to vibrate fully at their fundamental frequency. This vibration pattern uses the least amount of energy to create a strong, audible voice.	**b) Modal register.** Modal register is the normal speaking voice that we use for everyday communication. In this mode, the vocal folds vibrate regularly, producing clear and resonant sounds.
Which specific lymphoma of the head and neck region is more commonly associated with Epstein-Barr virus (EBV)? a) Waldeyer's ring lymphoma **b) Sinonasal lymphoma** c) Salivary gland lymphoma d) Thyroid lymphoma e) Laryngeal lymphoma	Head & Neck	**b) Sinonasal lymphoma.** Lymphoma of the head and neck region commonly associated with EBV typically occurs in the nasal cavities (sinonasal). EBV has been associated with various types of lymphomas, such as Burkitt lymphoma, but sinonasal lymphoma in the head and neck region is more distinctly linked with EBV.	**a) Waldeyer's ring lymphoma.** EBV is strongly linked to Burkitt's lymphoma, a type of aggressive B-cell non-Hodgkin lymphoma. Waldeyer's ring, located in the upper pharynx behind the nose and above the tonsils, is a frequent site for Burkitt's lymphoma development, especially in children and young adults. While EBV can be associated with other lymphomas in the head and neck region, the connection is strongest with Burkitt's lymphoma arising in Waldeyer's ring.	**b) Sinonasal lymphoma.** EBV-associated lymphomas are a diverse group of lymphoproliferative disorders linked to EBV infection. Given the options provided, the lymphoma most commonly associated with EBV in the head and neck region is sinonasal lymphoma. Sinonasal lymphomas often exhibit angiocentric and natural killer cell features, and EBV infection is implicated in their pathogenesis.
Which of the following characteristics affect the choice of treatment in toxic solitary nodule? I. thyroid-stimulating hormone (TSH) level II. Nodule size III. Response to antithyroid treatment IV. Pregnancy or lactation V. Presence of compressive symptoms a) I, II, IV, V b) I, III, IV, V **c) II, III, IV, V** d) I, II, III, IV e) I, II, III, V	Head & Neck	**c) II, III, IV, V.** The choice of treatment in toxic solitary nodules is influenced by factors such as TSH levels (I), nodule size (II), response to antithyroid treatment (III), and the presence of compressive symptoms (V). These factors help determine whether medical management, radioactive iodine therapy, or surgical intervention is most appropriate for managing the condition.	**c) II, III, IV, V.** The choice of treatment for a toxic solitary nodule is influenced by several factors: nodule size, response to antithyroid treatment, pregnancy or lactation, and presence of compressive symptoms. A toxic nodule, by definition, suppresses TSH, so the absolute TSH level is not a major factor in choosing treatment.	**b) I, III, IV, V.** The choice of treatment for a toxic solitary nodule (also known as a toxic adenoma) depends on several factors. The correct answer is I, III, IV, V. These factors play a role in determining the most appropriate treatment for toxic solitary nodules. Remember that individual patient characteristics and preferences also influence treatment decisions.

ENT:,

LLMs are known to perform differently in various languages, which is often linked to the amount and quality of training data available. Since a considerable amount of online content is in English, LLMs in English usually have better comprehension capabilities and access to a broader knowledge base. A technical report released on GPT4 revealed that GPT3.5 and PaLM perform 70.1% and 69.3% accuracy in massive multitask language understanding studies, respectively, whereas GPT4 shows 85.5% in English (21). It has been found that Turkish has a similarity rate of 80%, closely matching Italian at 84.1%, German at 83.7%, and Korean at 77%. However, the success rate drops for Nepali to 72.2%, Thai to 71.8%, and Telugu to 62% (22).

Furthermore, the complexity of a language's structure, its grammatical rules, and cultural factors can influence the model's performance (23). Languages with more complex grammatical features, such as gender, case, and tense, may pose greater challenges for LLMs. Nonetheless, technological advancements and the increasing use of multilingual models have significantly improved performance in other languages. This progress can make language models more universally applicable, providing better services to users in different languages.

As the world advances in every manner, educational models continuously evolve from traditional to more technology-based styles (23). Residents in various specialty areas and students in medicine can benefit from using LLMs to enhance their learning and clinical experiences. LLMs are advanced repositories of medical knowledge that provide instant access to a wide range of medical literature and research, making it easier to learn and make decisions based on evidence. They provide personalized education by offering responses tailored to specific queries, allowing residents and students to explore complex medical scenarios. LLMs also aid in developing differential diagnoses by providing conditions based on existing symptoms, which can help with clinical reasoning and decision-making processes (24).

In addition to their primary functions, these models can be used to interpret medical data such as laboratory results and radiographic images. The models can help medical professionals make more informed decisions by providing contextual information and potential implications. LLMs can also be used for language translation in medical contexts. This is particularly useful in understanding medical texts in various languages, promoting a more global medical perspective. LLMs also offer the potential for simulation-based learning, where residents and students can engage in virtual patient scenarios to enhance their diagnostic and therapeutic skills in a safe environment (25).

There are a few limitations to this study that are worth noting. Firstly, we only examined three of the most commonly used LLMs today: ChatGPT, Gemini, and Bing. While these models are widely available, many others could also be included in more comprehensive studies to further our knowledge in this field. Secondly, one potential limitation is that the reproducibility of the chatbots' responses varied with each query. This is something to keep in mind when interpreting the results. Thirdly, the questions used in the study were based on a specific textbook formatted according to the universal medical literature by the Turkish Society of Otorhinolaryngology Head and Neck surgery. It is important to consider that different question formats based on various database sources may produce different outcomes. Lastly, it is worth noting that the questions were designed to comply with the International Test Commission guidelines (26). However, minor discrepancies may occur since they were prepared by a joint commission of five faculty members.

Conclusion

Our study suggests that LLMs currently have limitations in achieving the same medical accuracy as senior resident surgeons. However, that the performance of ChatGPT is comparable to that of mid-level residents-and they excel in specific subspecialties-indicates the potential usefulness in certain medical fields. Meanwhile, Gemini and Bing show promise as valuable resources for education and the initial stages of clinical support, as their accuracy levels are similar to junior surgeons. Nevertheless, the performance of these models varies across different subspecialties, highlighting the need for the development and application of tailored LLMs to meet the requirements of each field.

Acknowledgements

Special thanks to Oğuz Basut MD, Fikret Kasapoğlu MD, Ömer Afşin Özmen MD, Uygar Levent Demir MD, Mustafa Aslıer MD, and Hakan Coşkun MD (retired), who are the faculty members of the Otorhinolaryngology Department at Bursa Uludağ University. They prepared the questions for resident education and contributed to this study by providing data. I am grateful for their permission to use this material for research purposes.

Ethics Committee Approval: The study did not require ethics committee approval as it relied solely on the question database of the university clinic, which was derived from publicly available online medical textbooks.

Informed Consent: The study did not require patient consent as it was based entirely on the university clinic's database of questions obtained from publicly available online medical textbooks.

Conflict of Interest: The authors have no conflicts of interest to declare.

Financial Disclosure: The authors declared that this study has received no financial support.

Main Points

- Technology and artificial intelligence are becoming ever-increasingly popular and being integrated into our lives. Artificial intelligence products known as Large Language Models (LLMs), such as chatbots, generate human-like responses and problem-solving skills. Their abilities to solve exam questions such as USMLE, Bar, and MBA have been investigated.
- Although examined in cardiology, ophthalmology, orthopedics, obstetrics, gynecology, and otorhinolaryngology, the comparison of the success of LLMs and human counterparts has not yet been investigated.
- In this study, the performance of LLMs in exams applied for resident training at a University Clinic was compared among themselves and with residents at three levels of seniority, using the Turkish Otorhinolaryngology Head and Neck Society Assistant Training Basic Textbook as reference.
- ChatGPT was more successful than other LLMs in total and across all subspecialties of otorhinolaryngology. Bing showed close success to ChatGPT in the field of laryngology. The senior residents was the most successful, while ChatGPT approached the performance of the mid-level residents. ChatGPT and Gemini achieved results similar to those of the junior residents.
- LLMs are far from senior residents' levels of knowledge, skills, and experience under current conditions. However, they can be preferred for their supportive features in the early years of resident training.

References

1. Gkinko L, Elbanna A. The appropriation of conversational AI in the workplace: a taxonomy of AI chatbot users. Int J Inf Manage. 2023; 69: 102568. [Crossref]
2. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ. 2023; 103: 102274. [Crossref]
3. Adamopoulou E, Moussiades L. Chatbots: history, technology, and applications. Mach Learn with Appl. 2020; 2: 100006. [Crossref]
4. Choi JH, Hickman KE, Monahan AB, Schwarcz DB. ChatGPT goes to law school. SSRN Electron J. Published online January 23, 2023. [Crossref]
5. Katz DM, Bommarito MJ, Gao S, Arredondo P. GPT-4 passes the bar exam. Philos Trans A Math Phys Eng Sci. 2024; 382: 20230254. [Crossref]
6. Terwiesch C. Would Chat GPT3 get a Wharton MBA? A prediction based on its performance in the operations management course. Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania: 2023. [Crossref]
7. Wood DA, Achhpilia MP, Adams MT, Aghazadeh S, Akinyele K, Akpan M, et al. The ChatGPT artificial intelligence chatbot: how well does it answer accounting assessment questions? Issues Account Educ 2023; 38: 81-108. [Crossref]
8. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, 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. [Crossref]
9. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Heal 2023; 2: e0000198. [Crossref]
10. Teebagy S, Colwell L, Wood E, Yaghy A, Faustina M. Improved performance of ChatGPT-4 on the OKAP examination: A comparative study with ChatGPT-3.5. J Acad Ophthalmol. 2023; 15: 184-7. [Crossref]
11. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery. 2023; 93: 1090-8. [Crossref]
12. Long C, Lowe K, Zhang J, Santos AD, Alanazi A, O'Brien D, et al. A novel evaluation model for assessing ChatGPT on otolaryngology–head and neck surgery certification examinations: performance study. JMIR Med Educ. 2024; 10: e49970. [Crossref]
13. 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. [Crossref]
14. Confirmed: the new Bing runs on OpenAI's GPT-4. Bing Search Blog. Accessed February 6, 2024. [Crossref]
15. Google AI PaLM 2 - Google AI. Accessed February 6, 2024. [Crossref]
16. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. 2023; 307: e230922. [Crossref]
17. Raimondi R, Tzoumas N, Salisbury T, Di Simplicio S, Romano MR; North East Trainee Research in Ophthalmology Network (NETRiON). Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. Eye (Lond). 2023; 37: 3530-3. [Crossref]
18. Kuşcu O, Pamuk AE, Sütay Süslü N, Hosal S. Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? Front Oncol. 2023; 13: 13:1256459. [Crossref]
19. Radulesco T, Saibene AM, Michel J, Vaira LA, Lechien JR. ChatGPT-4 performance in rhinology: A clinical case series. Int Forum Allergy Rhinol. 2024; 14: 1123-30. [Crossref]
20. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ. 2019; 5: e16048. [Crossref]
21. Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, et al. Measuring Massive Multitask Language Understanding. ICLR

2021 - 9th Int Conf Learn Represent. Published online September 7, 2020. Accessed February 9, 2024. [Crossref]

22. https://cdn.openai.com/papers/gpt-4.pdf [Crossref]
23. Juhi A, Pipil N, Santra S, Mondal S, Behera JK, Mondal H. The capability of ChatGPT in predicting and explaining common drug-drug interactions. Cureus. 2023; 15: e36272. [Crossref]
24. Sinha RK, Deb Roy A, Kumar N, Mondal H. AApplicability of ChatGPT in assisting to solve higher order problems in pathology. Cureus. 2023; 15: e35237. [Crossref]
25. Mondal H, Marndi G, Behera JK, Mondal S. ChatGPT for teachers: practical examples for utilizing artificial intelligence for educational purposes. Indian J Vasc Endovasc Surg. 2023; 10: 200-5. [Crossref]
26. International Test Commission. The ITC guidelines for translating and adapting tests (Second Edition).; 2017. Accessed January 30, 2024. [Crossref]