Feasibility of Multimodal Artificial Intelligence using Generative Pre-Trained Transformer 4-Vision for the Classification of Middle Ear Disease

Masao Noda, Hidekane Yoshimura, Takuya Okubo, Ryota Koshu, Yuki Uchiyama, Akihiro Nomura, Makoto Ito, Yutaka Takumi

Submitted to: JMIR AI
on: March 13, 2024

Table of Contents

Feasibility of Multimodal Artificial Intelligence using Generative Pre-Trained Transformer 4-Vision for the Classification of Middle Ear Disease

Masao Noda[1] MD, PhD, MBA; Hidekane Yoshimura[2] MD, PhD; Takuya Okubo[2] MD; Ryota Koshu[1] MD; Yuki Uchiyama[2] MD; Akihiro Nomura[3] MD, PhD; Makoto Ito[1] MD, PhD; Yutaka Takumi[2] MD, PhD

[1]Department of Otolaryngology and Head and Neck Surgery Jichi medical university Shimotsuke JP
[2]Department of Otolaryngology ? Head and Neck Surgery Shinshu University Matsumoto JP
[3]College of Transdisciplinary Sciences for Innovation Kanazawa University Kanazawa JP

Corresponding Author:
Masao Noda MD, PhD, MBA
Department of Otolaryngology and Head and Neck Surgery
Jichi medical university
Shimotsuke, Yakushiji ??????
Shimotsuke
JP

Abstract

Background: The integration of artificial intelligence (AI), particularly deep learning models, has transformed the landscape of medical technology, especially in the field of diagnosis utilizing imaging and physiological data. In otolaryngology, AI has shown promise in image classification for middle ear diseases. However, existing models often lack patient-specific data and clinical context, limiting their universal applicability. The emergence of Generative Pre-trained Transformer 4 Vision (GPT-4V) has enabled a multimodal diagnostic approach, integrating language processing with image analysis.

Objective: In this study, we investigated the effectiveness of GPT-4V in diagnosing middle ear diseases by integrating patient-specific data with otoscopic images of the tympanic membrane.

Methods: The study design was divided into two phases: (1) establishing a model with appropriate prompts and (2) validating the ability of the optimal prompt model to classify images. 305 otoscopic images of four middle ear disease (acute otitis media (AOM), middle ear cholesteatoma (Chole), chronic otitis media (COM), and otitis media with effusion (OME)) were obtained from patients who visited Shinshu University or Jichi Medical University between April 2010 and December 2023. The optimized GPT-4V settings were established using prompts and patients' data, and the model with the optimal prompt created was used to verify the diagnostic accuracy of GPT-4V on 190 images. To compare the diagnostic accuracy of GPT-4V with that of physicians, 30 clinicians completed a web-based questionnaire consisting of 190 images.

Results: The multimodal AI approach achieved an accuracy of 82.1%, which is superior to that of certified pediatricians, 70.6%, but trailing behind that of otolaryngologists, more than 95%. The model's disease-specific accuracy rates were 89.19% for AOM, 76.5% for COM, 79.3% for cholesteatoma, and 85.7% for OME, which highlights the need for disease-specific optimization. Comparisons with physicians revealed promising results, suggesting the potential of GPT-4V to augment clinical decision-making.

Conclusions: Despite its advantages, challenges such as data privacy and ethical considerations must be addressed. Overall, this study underscores the potential of multimodal AI for enhancing diagnostic accuracy and improving patient care in otolaryngology. Further research is warranted to optimize and validate this approach in diverse clinical settings

(JMIR Preprints 13/03/2024:58342)
DOI: https://doi.org/10.2196/preprints.58342

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

✓ **Only make the preprint title and abstract visible.**
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Feasibility of Multimodal Artificial Intelligence using Generative Pre-Trained Transformer 4-Vision for the Classification of Middle Ear Disease

Abstract

Background: The integration of artificial intelligence (AI), particularly deep learning models, has transformed the landscape of medical technology, especially in the field of diagnosis utilizing imaging and physiological data. In otolaryngology, AI has shown promise in image classification for middle ear diseases. However, existing models often lack patient-specific data and clinical context, limiting their universal applicability. The emergence of Generative Pre-trained Transformer 4 Vision (GPT-4V) has enabled a multimodal diagnostic approach, integrating language processing with image analysis.

Objective: In this study, we investigated the effectiveness of GPT-4V in diagnosing middle ear diseases by integrating patient-specific data with otoscopic images of the tympanic membrane.

Methods: The study design was divided into two phases: (1) establishing a model with appropriate prompts and (2) validating the ability of the optimal prompt model to classify images. In total, 305 otoscopic images of 4 middle ear disease (acute otitis media [AOM], middle ear cholesteatoma [Chole], chronic otitis media [COM], and otitis media with effusion [OME]) were obtained from patients who visited Shinshu University or Jichi Medical University between April 2010 and December 2023. The optimized GPT-4V settings were established using prompts and patients' data, and the model created with the optimal prompt was used to verify the diagnostic accuracy of GPT-4V on 190 images. To compare the diagnostic accuracy of GPT-4V with that of physicians, 30 clinicians completed a web-based questionnaire consisting of 190 images.

Results: The multimodal AI approach achieved an accuracy of 82.1%, which is superior to that of certified pediatricians, 70.6%, but trailing behind that of otolaryngologists, more than 95%. The model's disease-specific accuracy rates were 89.19% for AOM, 76.5% for COM, 79.3% for Chole, and 85.7% for OME, which highlights the need for disease-specific optimization. Comparisons with physicians revealed promising results, suggesting the potential of GPT-4V to augment clinical decision-making.

Conclusions: Despite its advantages, challenges such as data privacy and ethical considerations must be addressed. Overall, this study underscores the potential of multimodal AI for enhancing diagnostic accuracy and improving patient care in otolaryngology. Further research is warranted to optimize and validate this approach in diverse clinical settings.

Keywords: Artificial intelligence; Deep learning; Generative AI; Tympanic membrane; Middle ear disease; GPT4-Vision; Otolaryngology

Introduction

The emergence of artificial intelligence (AI) has altered the landscape of medical technology, particularly in diagnosis, which leverages the identification of features based on imaging and physiological data [1-3]. In the field of otolaryngology, AI and deep learning models are being used for imaging; ongoing efforts focus on classifying diseases based on tympanic membrane images of middle ear disease [4-6]. Technological advancements, including deep learning and transfer learning

using pre-trained models, have resulted in an accuracy range of 70–90% in models for analyzing otoscopic images [7]. There have also been advancements in its application, such as implementing smartphone-based point-of-care diagnostics [8]. However, these models rely on trained image data, require large image datasets, and do not consider patient information or clinical context. Consequently, the universality of these models is limited, and their optimal application in clinical practice remains unclear.

Recently, large-scale language-processing models have become available for general use. One such model, the Generative Pre-trained Transformer 4 (GPT-4), has demonstrated specialist-level medical knowledge through its language-processing abilities [9-11]. Since October 2023, GPT-4 Vision (GPT-4V) has gained the ability to evaluate image data, enabling a multimodal diagnostic approach that incorporates both language processing and image analysis [12]. GPT-4V enables the integration of patient information analysis and image-based deep learning models, providing valuable support in diagnosis and treatment, similar to decisions made in a clinical setting [13]. Multimodal AI, which bases diagnosis on multiple pieces of information, has been reported to be more effective than methods that rely on a single type of information. This is demonstrated in various medical applications, including the combination of pathology images with genomic information [14] and their utilization in liver cancer [15] and cervical cancer [16], where imaging information is integrated. In otorhinolaryngology, there have been few reports; however, efforts to incorporate AI for otoscopic images could further improve the quality of care.

In this study, we aimed to investigate the effectiveness of a multimodal approach using GPT-4V to diagnose middle ear disease. This approach was designed to integrate patient-specific data (age, sex, and chief complaint) with tympanic membrane images to assess the accuracy of the versatile GPT-4V. The model's accuracy was compared with physicians' diagnoses to validate its effectiveness in image-based deep learning. The potential future development of the multimodal AI approach for classifying middle ear diseases is also discussed.

Methods

GPT-4V has been available as an image recognition model since September 25, 2023. The study design was divided into two phases: (1) establishing a model with appropriate prompts and (2) validating the ability of the optimal prompt model to classify images (Figure 1).

Correct Otoscopic Images and Patient Information

This study included 305 otoscopic images of middle ear disease obtained from patients who visited Shinshu University or Jichi Medical University between April 2010 and December 2023. The endoscope used was an Olympus ENF-VH and ENF-V3 (Olympus, Tokyo, Japan), and the video system was an Olympus VISERA ELITE OTV-S190. One image was obtained from each patient. We excluded images with poor quality and those in which multiple diseases were suspected. The remaining images were classified into four disease categories: acute otitis media (AOM), middle ear cholesteatoma (Chole), chronic otitis media (COM), and otitis media with effusion (OME). The final diagnoses were based on the judgment of the otolaryngologists who treated the patients. These images were accompanied by patient-specific information, such as age, sex, and chief complaint (e.g., fever, otalgia, otorrhea, ear fullness, deafness, facial palsy, dizziness, and tinnitus). We excluded images taken after otologic surgery. Of note, only one image was obtained from each patient.

GPT-4V Settings and Prompt Tuning

The GPT-4V settings were established using prompts reported in previous studies [17,18]. Briefly, conditions and prompts for providing answers were verified using 10 images for each disease.

According to a report on prompts [19], image data and/or patient information were manually input into GPT-4V, and the generated results were evaluated by the physicians (N.M. and H.Y.).

Accuracy verification of GPT-4V using the optimal prompt model

The model with the optimal prompt created was used to verify the diagnostic accuracy of GPT-4V on 190 images (37 in AOM; 53 in Chole [Congenital, 6; Acquired, 47]; 51 in COM; and 49 in OME), which were different from those for tuning prompts. To account for the variability in responses, each administration was performed three times, and responses that were answered two or more times were considered to be the actual response.

Comparison of AI Accuracy with Physician Accuracy

To compare the diagnostic accuracy of GPT-4V with that of physicians, 30 clinicians completed a web-based questionnaire consisting of 190 images.

The web-based survey included tympanic membrane images and patient information (age, sex, and chief complaint) in a four-choice question format. The respondents included eight certificated pediatricians, eight otolaryngology residents, eight certificated otolaryngologists, and six experts in otolaryngology (more than 15 years of experience).

To show the trend in the percentage of correct responses according to the difficulty of the questions, the questions were divided into three levels (easy, normal, and hard) according to the overall percentage of correct responses by physicians, and the percentage of correct responses for each level and also each question was compared between the GPT-4V and all doctors, otolaryngologists, and pediatricians.

Ethical Statement

Patient information was anonymized to protect privacy and used only with the approval of the Ethics Committee of the Shinshu University School of Medicine (approval number: 6088).

Statistical Analysis

Groups were compared by one-way analysis of variance (ANOVA). Subsequently, multiple comparison tests (the Bonferroni method) were used to compare groups. Statistical significance was set at $p < 0.05$. A one-sample proportion test was used to compare the performance of physician with that of GPT-4V in terms of the correct response rate.

Results

Establishment of optimal prompts

In the initial stage, we sought an optimal input method using 10 images for each disease (AOM, Chole, COM, or OME; 40 images total). First, we inputted only images or options; GPT mostly requires clinical information, such as patient history and symptoms, although no response regarding the disease was generated (Figure 1 and Multimedia Appendix 1). Second, the names of the four diseases were added as candidate answers, but again, no response regarding the disease was generated. When detailed patient information, such as age, sex, and main symptoms, was inputted, GPT-4V provided answers, indicating that input images with patient data were the optimal prompt for testing the accuracy of GPT-4V.

Accuracy validation of the multimodal AI approach

The performance of the multimodal AI approach in this study for classifying middle ear diseases was validated, with an overall diagnostic accuracy of 82.1% for the GPT-4V-based analysis. Disease-specific accuracy rates were 89.19% for AOM (true positives [TP] = 33, false positives [FP] = 1, and false negatives [FN] = 4, precision = 0.97, recall = 0.89, F-score = 0.93), 76.5% for COM (TP = 39, FP = 7, FN = 12, precision = 0.85, recall = 0.76, F-score = 0.8), 79.3% for cholesteatoma (TP = 42, FP = 13, FN = 11, precision = 0.76, recall = 0.79, F-score = 0.78), and 85.7% for OME (TP = 42, FP = 10, FN = 7, precision = 0.81, recall = 0.86, F-score = 0.83) (Figure 2).

These results indicate high discrimination among various disease types; however, there were also some incorrect responses. Representative images of correct and incorrect GPT-4V classifications for each disease are shown in Figure 3.

Comparison of diagnostic accuracy by physicians and GPT-4V

The same images with patients' information used by GPT-4V were evaluated by pediatricians (n = 8), otolaryngology residents (n = 8), certificated otolaryngologists (n=8), and experts in otolaryngology (n = 6), and the diagnostic accuracy of each group was compared. The mean diagnostic accuracy was 70.6% (standard error: 4.2%) for pediatricians, 95.5% (standard error: 1.0%) for otolaryngology residents, 97.3% (standard error: 0.8%) for certificated otolaryngologists, and 98.2% (standard error: 0.4%) for experts in otolaryngology. ANOVA revealed significant differences among the four groups ($F = 13.43$, $p < 0.001$). In the post hoc comparison, a significant difference was observed between pediatricians and the other three groups ($p < 0.001$). The GPT-4V correct response rate was 82.1%, surpassing that of pediatricians by 11.5% and trailing behind otolaryngologists by an average of just over 10% (Figure 4).

The accuracy rates for specific diseases were as follows: 92.3% for AOM (pediatricians, 80.4%; otolaryngology residents, 94.9%; certificated otolaryngologists, 97.0%; and experts in otolaryngology, 98.2%), 95.9% for COM (pediatricians, 89.5%; otolaryngology residents, 96.6%; certificated otolaryngologists, 99.8%; and experts in otolaryngology, 98.4%), 81.8% for Chole (pediatricians, 46.0%; otolaryngology residents, 93.2%; certificated otolaryngologists, 93.6%; and experts in otolaryngology, 98.4%), and 93.7% for OME (pediatricians, 81.6%; otolaryngology residents, 97.2%; certificated otolaryngologists, 99.0%; and experts in otolaryngology, 98.0%).
In the confusion matrix of all doctors, there was a notable tendency to misclassify Chole as OME and AOM as OME. Among pediatricians, there were more errors in classifying Chole as AOM or COM (Figure 5).

Regarding difference in the trend of the percentage of correct answers between GPT-4V and physicians according to the difficulty of the questions, even the percentage of correct answers for GPT-4V tended to decrease gradually from 85.7% for easy, 84.0% for normal, and 71.1% for hard questions (Table 1).

Table 1. Comparison of the scores by GPT-4V and Human validation with physicians across various difficulty levels (N = 190).

Difficulty level	Questions, n (%)	GPT-4V, %	All doctors, % (95% CI)	Differences	P value	Otolaryngologists, % (95% CI)	Differences	P value	Pediatricians, % (95% CI)	Differences	P value
Easy (< 95%)	77 (40.5)	85.7	97.8 (97.4–98.2)	12.1	<.001[a]	99.7 (99.5–99.9)	14.0	<.001[a]	96.6 (95.3–97.9)	10.9	.006[a]
Normal (< 85%, < 95%)	75 (39.5)	84.0	90.4 (89.7–91.0)	6.4	.13	97.1 (96.2–98.0)	13.1	<.001[a]	76.3 (73.6–79.0)	-7.7	.07
Hard (< 85%)	38 (20.0)	71.1	76.8 (73.7–79.8)	5.7	.44	90.8 (87.2–94.3)	19.7	<.001[a]	45.4 (39.5–51.3)	-25.7	<.001[a]

[a] Statistically significant.

Furthermore, compared with otolaryngologists, GPT-4V had a significantly lower percentage of correct answers for all questions (99.7% for easy, 97.1% for normal, and 90.8% for hard questions; all $P < 0.001$). In contrast, the results of the "hard" and "normal" groups were similar. Compared with pediatricians, the GPT-4V outperformed the pediatricians in easy questions with 96.6%, although no statistically significant difference was observed ($P = 0.006$). However, the GPT-4V had a predominantly higher percentage of correct answers for normal (76.3%, $P = 0.07$) and hard questions (45.4%, $P < 0.001$).

Discussion

Principal Results

In this study, we assessed the accuracy of the GPT-4V multimodal AI approach in classifying middle ear disorders, yielding the following three key findings: (1) GPT-4V, a general-purpose model focusing on large-scale language models, achieved approximately 80% accuracy in classifying

middle ear disease. The model's performance, evaluated using images and patient data, was superior to that of non-otolaryngologists, although it was lower than the average accuracy of otolaryngologists. (2) The GPT-4V was able to classify diseases when patient information and disease options were input. Further improvements in accuracy could be achieved with more detailed patient information. (3) Accuracy varied by disease, suggesting the potential for optimizing AI usage and improving accuracy by understanding the specificity of GPT-4V in classifying particular diseases.

Comparison with Prior Work

The GPT-4V model has undergone training and utilizes zero-shot learning, which recognizes image features based on natural language to classify diseases based on image information and previously learned disease features [20]. GPT-4V can yield effective results with fewer resources than previous deep learning models, which typically require a large amount of image data, computational resources, time, and parameter adjustments for training. By inputting new information rather than simply classifying image data, it becomes possible to tailor diagnoses and diagnostic aids for each individual. Furthermore, GPT-4V and other large-scale language processing models feature prompt development that is appropriate for its usage purposes, since the accuracy of such models varies depending on the prompt adjustments.

Compared with physicians' accuracy, the model's performance in this study was higher than that of a pediatrician but lower than that of an otolaryngologist. In a previous comparison between deep learning and humans, Crowson et al. [21] classified 22 tympanic membrane images and found that the deep learning model achieved an accuracy of 95.5%, compared with an accuracy of 65% for 39 clinicians. Suresh et al. [22] also reported that a machine-learning model created from 1,000 images was more effective than pediatricians, with an accuracy rate of 90.6%, surpassing the clinicians' accuracy of 59.4%. Our results indicated that the model did not reach the proficiency level of otolaryngologists; however, it could be valuable for utilizing tympanic membrane images in medical practice outside of otolaryngology. In particular, GPT-4V judgments predominantly exceeded pediatricians' correct response rates for questions with normal to hard difficulty, suggesting that the present model may be useful for non-otolaryngologists who have difficulty in making such judgments.

Moreover, previous reports on deep learning classification models have determined the presence or absence of inflammation and exudates based on photographs alone. Further studies are needed to identify the optimal stage in the examination for implementing the image classification model and the subsequent policy decisions that should follow.

GPT-4V allows for the classification of diseases using patient information. While comments about medical or harmful content (with restrictions on medical advice) may result in a lower correct response rate, informative or educational responses are still possible if they are well-informed. Efforts have been made to use large language models (LLMs) to improve the accuracy of prompts. Therefore, it is possible to develop appropriate prompts for medical imaging and middle ear disorders. The accuracy of the LLM is expected to further improve with the development of prompts that are specifically tailored for medical imaging and middle ear disease [23,24].

For the clinical application of the GPT-4V model, collecting clinical data and adjusting parameters are needed to further improve its diagnostic accuracy for each middle ear disease. Upon reviewing the incorrect responses of GPT-4V for each disease, we found that Chole might demonstrate a retraction pocket, which may be mistaken for a perforation. However, images with keratin debris accumulation in the retraction pocket were less prone to misclassification. In cases of COM with calcification, a white lesion was considered to be Chole calcification, emphasizing the importance of distinguishing between these two diseases. AOM cases without the chief complaint of acute inflammation (fever, ear pain, or ear discharge) were occasionally misclassified, even with characteristic findings such as a bulging tympanic membrane, suggesting that GPT-4V was likely to

prioritize patients' information over images. In OME cases, a white lesion was sometimes considered to be a pearly tumor (Chole) or tympanic membrane perforation (COM), particularly when it involved a small amount of effusion and/or air. For physicians, Chole and AOM were often misidentified as other diseases and OME, respectively. When comparing the GPT-4V model with the entire group of physicians, the percentage of correct responses was generally higher among the physicians. However, the GPT-4V diagnostic accuracy for Chole was higher than that of pediatricians, indicating that GPT-4V could help non-otolaryngologists diagnose Chole. In a previous report, a dedicated AI model had a diagnostic accuracy of approximately 90% for Chole [25]; therefore, the combination of such a system and GPT-4V would be useful to improve the accuracy of Chole detection.

As demonstrated in this study, the application of AI, including LLM, is believed to offer advantages in terms of improving efficiency and providing assistance in clinical work, enabling the delivery of high-quality medical care, and overcoming language barriers in medicine. The use of GPT-4V has already been reported to diagnose complicated cases [26], and its application can be expanded by integrating it with imaging information. In the field of orthopedics, trials are underway to determine treatment methods based on MRI reports [27], showcasing the effectiveness of GPT-4V as an aid in image interpretation. GPT has been shown to return answers and provide details about the disease, including risk factors and treatment methods. This allows for the evaluation of images alone and assists in medical treatment. Such insights are valuable for understanding the practical use and challenges of AI in real-world applications. Unlike the simplistic deep learning models of the past, the LLM can enhance accuracy by presenting evidence for judgments and asking a series of questions. When used by physicians with a certain level of specialized knowledge, the LLM effectively aids judgment, leading to increased efficiency in medical care. GPT-4V provides answers in just a few seconds, which is significantly shorter than the time it takes a physician to provide a diagnosis, thereby confirming its efficiency. GPT-4V can be used on smartphones, potentially making medical treatment more location-independent. However, there are associated risks, including the reliance on AI for medical care, misdiagnoses due to system malfunctions, and patient information leakage. ChatGPT is trained based on information up to a certain period but may respond differently at different times or provide answers using outdated criteria. Furthermore, legal and personal literacy measures must be developed to protect personal information and address ethical concerns. Foreign countries and the United Nations are actively promoting laws and regulations governing the use of AI [28,29].

Limitations

One limitation of this study is the use of a limited number of images (n=190). Further analysis is required to assess the impact of using a larger dataset that encompasses various diseases. Additionally, as there are large variations in the quality of otoscopic images, accurate diagnosis might be challenging in some cases.

The recognition and content of the answers may change depending on the doctor, clinics, and designed prompt; the accuracy may also change due to changes in the image quality used or the method used to capture the image. While this is common to deep learning, the advantage of GPT, which does not require prior training, is that it is not affected by the data to be trained; thus, the possibility of such changes is considered to be small.

For these reasons, further exploration is needed on strategies for handling challenging images and facilitating open-ended responses without giving predefined options. Furthermore, because of the rapid pace of technological evolution, it is essential to regularly fine-tune and make a standalone model that ensures reliability and consistency over time.

Conclusions

A multimodal AI approach using GPT-4V has revealed a potential new diagnostic approach for classifying middle-ear diseases. This confirms the ability of AI to assist in clinical diagnosis and identify disease-specific features. The significant improvement in accuracy compared with conventional deep learning models indicates that even general-purpose AI technology can assist in medical treatment with a certain level of accuracy. It can be applied to highly specialized diagnoses, depending on the method. Further improvements in diagnostic accuracy are expected in future studies by integrating more diverse data types.

Acknowledgements

Author contributions
Noda Masao: conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, validation, visualization, writing - original draft. Hidekane Yoshimura: conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, validation, writing - review & editing. Takuya Okubo, Ryota Koshu, Yuki Uchiyama: investigation, project administration, writing - review & editing. Akihiro Nomura, Makoto Ito, Yutaka Takumi: supervision, writing - review & editing.
Other Acknowledgements
We are thankful to our colleagues at Shinshu University: Dr. Sota Ichimura, Dr. Kota Hirose, Dr. Mariko Kasuga, Dr. Shu Yokota, Dr. Kentaro Hori, Dr. Arisa Oguchi, Dr. Kenjiro Sugiyama, Dr. Jun Shinagawa, Dr. Yoichiro Iwasa, Dr. Keita Tsukada, Dr. Tomohiro Oguchi and Dr. Nobuyoshi Suzuki of the Department of Otolaryngology – Head and Neck, and those at Jichi Medical University: Dr. Kota Matsuyama, Dr. Akiko Uchida, Dr. Yuki Miura of the Department of Otolaryngology and Dr. Shinya Fukuda, Dr. Kazuki Okumura and Dr. Keizo Wakae of the Department of Pediatrics, and Dr. Hitoshi Irabu, Dr. Kazuhiro Noguchi, Dr. Ryo Nakagawa, Dr. Narutoshi Yamazaki, Dr. Yuji Takaso, Dr. Keisuke Koyama, Dr. Yukari Nakamura, Dr. Chia Sasaki and Dr. Keigo Nishida for their invaluable cooperation in this study. We thank Editage (www.editage.com) for English language editing.
Funding
The authors declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflicts of Interest

None declared.

Abbreviations

AOM: acute otitis media
AI: artificial intelligence
COM: chronic otitis media
FP: false positives
FN: false negatives
GPT-4V: Generative Pre-trained Transformer 4 Vision
LLMs: large language models
Chole: middle ear cholesteatoma
ANOVA: one-way analysis of variance
OME: otitis media with effusion
TP: true positives

Multimedia Appendix 1

References

1. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F, Dong J, Prasadha MK, Pei J, Ting MYL, Zhu J, Li C, Hewett S, Dong J, Ziyar I, Shi A, Zhang R, Zheng L, Hou R, Shi W, Fu X, Duan Y, Huu VAN, Wen C, Zhang ED, Zhang CL, Li O, Wang X, Singer MA, Sun X, Xu J, Tafreshi A, Lewis MA, Xia H, Zhang K. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018;172:1122–1131.e9. doi:10.1016/j.cell.2018.02.010.
2. Schaefferkoetter J, Yan J, Moon S, Chan R, Ortega C, Metser U, Berlin A, Veit-Haibach P. Deep learning for whole-body medical image generation. Eur J Nucl Med Mol Imaging 2021;48:3817–3826. doi:10.1007/s00259-021-05413-0.
3. Lee CC, Lin CS, Tsai CS, Tsao TP, Cheng CC, Liou JT, Lin WS, Lee CC, Chen JT, Lin C. A deep learning-based system capable of detecting pneumothorax via electrocardiogram. Eur J Trauma Emerg Surg 2022;48:3317–3326. doi:10.1007/s00068-022-01904-3.
4. Choi Y, Chae J, Park K, Hur J, Kweon J, Ahn JH. Automated multi-class classification for prediction of tympanic membrane changes with deep learning models. PLoS One 2022;17:e0275846. doi:10.1371/journal.pone.0275846.
5. Park YS, Jeon JH, Kong TH, Chung TY, Seo YJ. Deep learning techniques for ear diseases based on segmentation of the normal tympanic membrane. Clin Exp Otorhinolaryngol 2023;16:28–36. doi:10.21053/ceo.2022.00675.
6. Alhudhaif A, Cömert Z, Polat K. Otitis media detection using tympanic membrane images with a novel multi-class machine learning algorithm. Peer J Comput Sci 2021;7:e405. doi:10.7717/peerj-cs.405.
7. Zeng X, Jiang Z, Luo W, Li H, Li H, Li G, Shi J, Wu K, Liu T, Lin X, Wang F, Li Z. Efficient and accurate identification of ear diseases using an ensemble deep learning model. Sci Rep 2021;11:10839. doi:10.1038/s41598-021-90345-w.
8. Chen YC, Chu YC, Huang CY, Lee YT, Lee WY, Hsu CY, Yang AC, Liao WH, Cheng YF. Smartphone-based artificial intelligence using a transfer learning algorithm for the detection and diagnosis of middle ear diseases: a retrospective deep learning study. EclinicalMedicine 2022;51:101543. doi:10.1016/j.eclinm.2022.101543.
9. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, Nadkarni G, Klang E. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023;13:16492. doi:10.1038/s41598-023-43436-9.
10. Jang D, Yun TR, Lee CY, Kwon YK, Kim CE. GPT-4 can pass the Korean national licensing examination for Korean medicine doctors. PLOS Digit Health 2023;2:e0000416. doi:10.1371/journal.pdig.0000416.
11. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ 2023;9:e48002. doi:10.2196/48002.
12. GPT-4V(ision) System Card. https://openai.com/research/gpt-4v-system-card (accessed September 25, 2023).
13. Wu W, Yao H, Zhang M, Song Y, Ouyang W, Wang J. GPT4Vis: what can GPT-4 do for zero-shot visual recognition? arXiv preprint arXiv:2311.15732; 2023. doi:10.48550/arXiv.2311.15732.
14. Chen RJ, Lu MY, Williamson DFK, Chen TY, Lipkova J, Noor Z, Shaban M, Shady M, Williams M, Joo B, Mahmood F. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell 2022;40:865–878.e6.

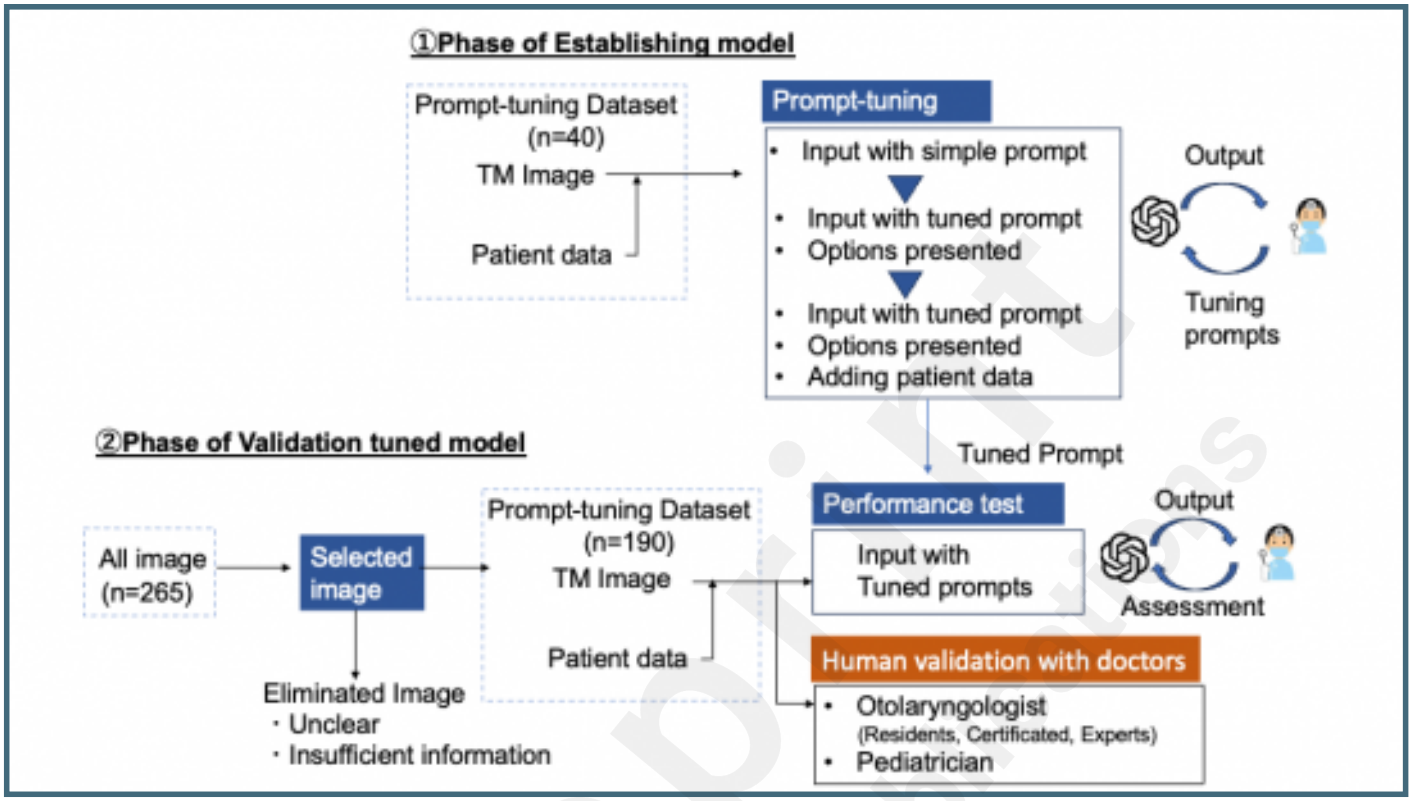
doi:10.1016/j.ccell.2022.07.004.
15. Zhang L, Jiang Y, Jin Z, Jiang W, Zhang B, Wang C, Wu L, Chen L, Chen Q, Liu S, You J, Mo X, Liu J, Xiong Z, Huang T, Yang L, Wan X, Wen G, Han XG, Fan W, Zhang S. Real-time automatic prediction of treatment response to transcatheter arterial chemoembolization in patients with hepatocellular carcinoma using deep learning based on digital subtraction angiography videos. Cancer Imaging 2022;22:23. doi:10.1186/s40644-022-00457-3.
16. Ming Y, Dong X, Zhao J, Chen Z, Wang H, Wu N. Deep learning-based multimodal image analysis for cervical cancer detection. Methods 2022;205:46–52. doi:10.1016/j.ymeth.2022.05.004.
17. Tanaka Y, Nakata T, Aiga K, Etani T, Muramatsu R, Katagiri S, Kawai H, Higashino F, Enomoto M, Noda M, Kometani M, Takamura M, Yoneda T, Kakizaki H, Nomura A. Performance of generative pretrained transformer on the national medical licensing examination in Japan. PLOS Digit Health 2024;3:e0000433. doi:10.1371/journal.pdig.0000433.
18. Masao N, Takayoshi U, Ryota K, Ito M, Yamoto N, Yoshizaki T, Nomura A. A study of the performance of the Generative Pretrained Transformer in the Japanese Otorhinolaryngology Specialty Examination. Nippon Jibiinkoka Tokeibugeka Gakkai Kaiho (Tokyo) 2023;126:1217-1223. doi:10.3950/jibiinkotokeibu.126.11_1217.
19. Bharat SM, Myrzakhan A, Shen Z. Principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4, arXiv:2312.16171 [cs.CL] (or arXiv:2312.16171v2 [cs.CL] for this version). doi:10.48550/arXiv.2312.16171.
20. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. doi:10.48550/arXiv.2103.00020
21. Crowson MG, Bates DW, Suresh K, Cohen MS, Hartnick CJ. "Human vs machine" validation of a deep learning algorithm for pediatric middle ear infection diagnosis. Otolaryngol Head Neck Surg 2023;169:41–46. doi:10.1177/01945998221119156
22. Suresh K, Wu MP, Benboujja F, Christakis B, Newton A, Hartnick CJ, Cohen MS. AI model versus clinician otoscopy in the operative setting for otitis media diagnosis. Otolaryngol Head Neck Surg 2023. Advance online publication. doi:10.1002/ohn.559.
23. Hu Y, Chen Q, Du J, Peng X, Keloth VK, Zuo X, Zhou Y, Li Z, Jiang X, Lu Z, Roberts K, Xu H. Improving large language models for clinical named entity recognition via prompt engineering. J Am Med Inform Assoc 2024 Jan 27:ocad259. Advance online publication. doi:10.1093/jamia/ocad259
24. Meskó B. Prompt engineering as an important emerging skill for medical professionals: Tutorial. J Med Internet Res 2023;25:e50638. doi:10.2196/50638
25. Tseng CC, Lim V, Jyung RW. Use of artificial intelligence for the diagnosis of cholesteatoma. Laryngoscope Investig Otolaryngol 2023;8:201–211. doi:10.1002/lio2.1008.
26. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023;330:78–80. doi:10.1001/jama.2023.8288
27. Truhn D, Weber CD, Braun BJ, Bressem K, Kather JN, Kuhl C, Nebelung S. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. Sci Rep 2023;13:20159. doi:10.1038/s41598-023-47500-2
28. Vlahou A, Hallinan D, Apweiler R, Argiles A, Beige J, Benigni A, Bischoff R, Black PC, Boehm F, Céraline J, Chrousos GP, Delles C, Evenepoel P, Fridolin I, Glorieux G, van Gool AJ, Heidegger I, Ioannidis JPA, Jankowski J, Jankowski V, Jeronimo C, Kamat AM, Masereeuw R, Mayer G, Mischak H, Ortiz A, Remuzzi G, Rossing P, Schanstra JP, Schmitz-Dräger BJ, Spasovski G, Staessen JA, Stamatialis D, Stenvinkel P, Wanner C, Williams SB, Zannad F, Zoccali C, Vanholder R. Data sharing under the general data protection regulation: Time to harmonize law and research ethics? Hypertension 2021;77:1029–1035. doi:10.1161/

HYPERTENSIONAHA.120.16340
29. Fournier-Tombs E, McHardy J. A medical ethics framework for conversational artificial intelligence. J Med Internet Res 2023;25:e43068. doi:10.2196/43068

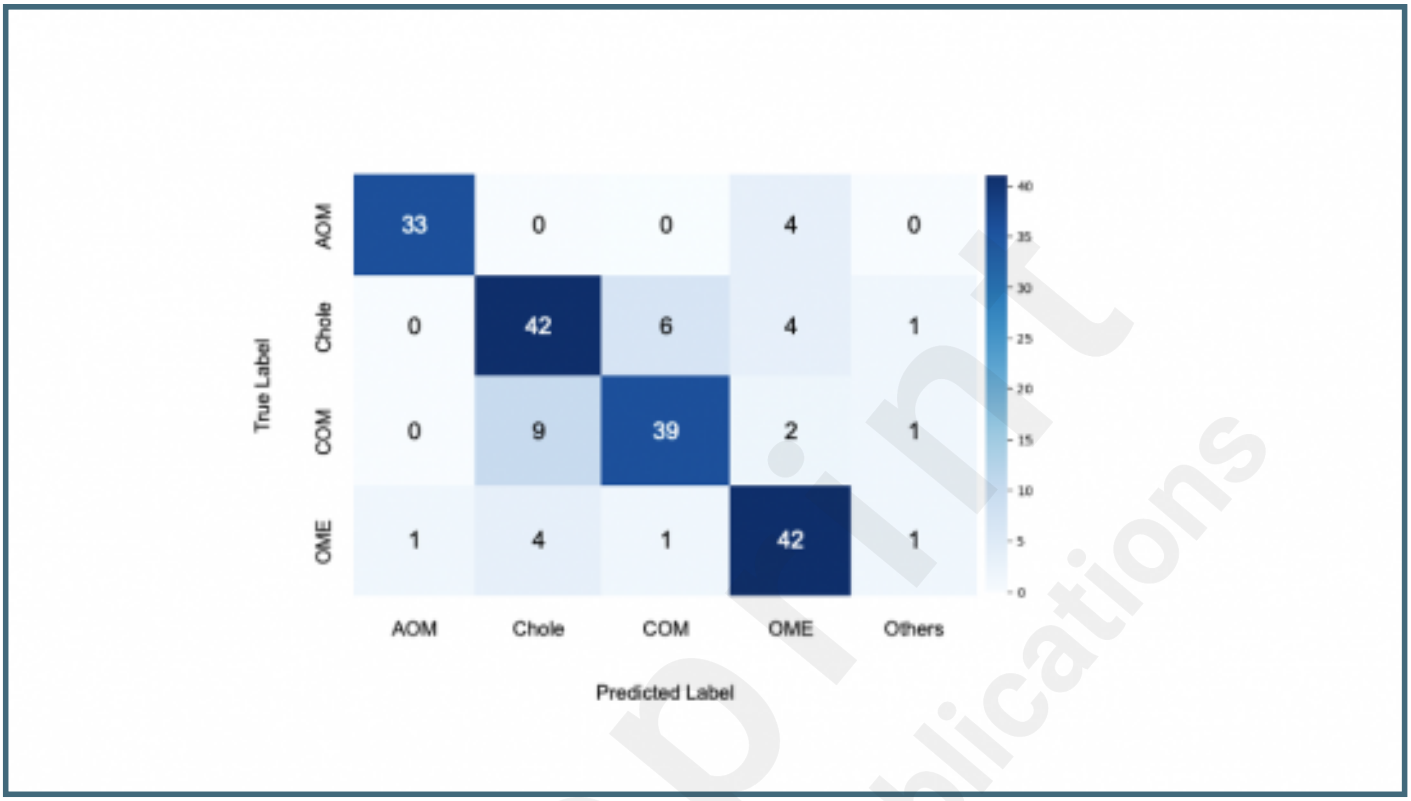
Supplementary Files

Figures

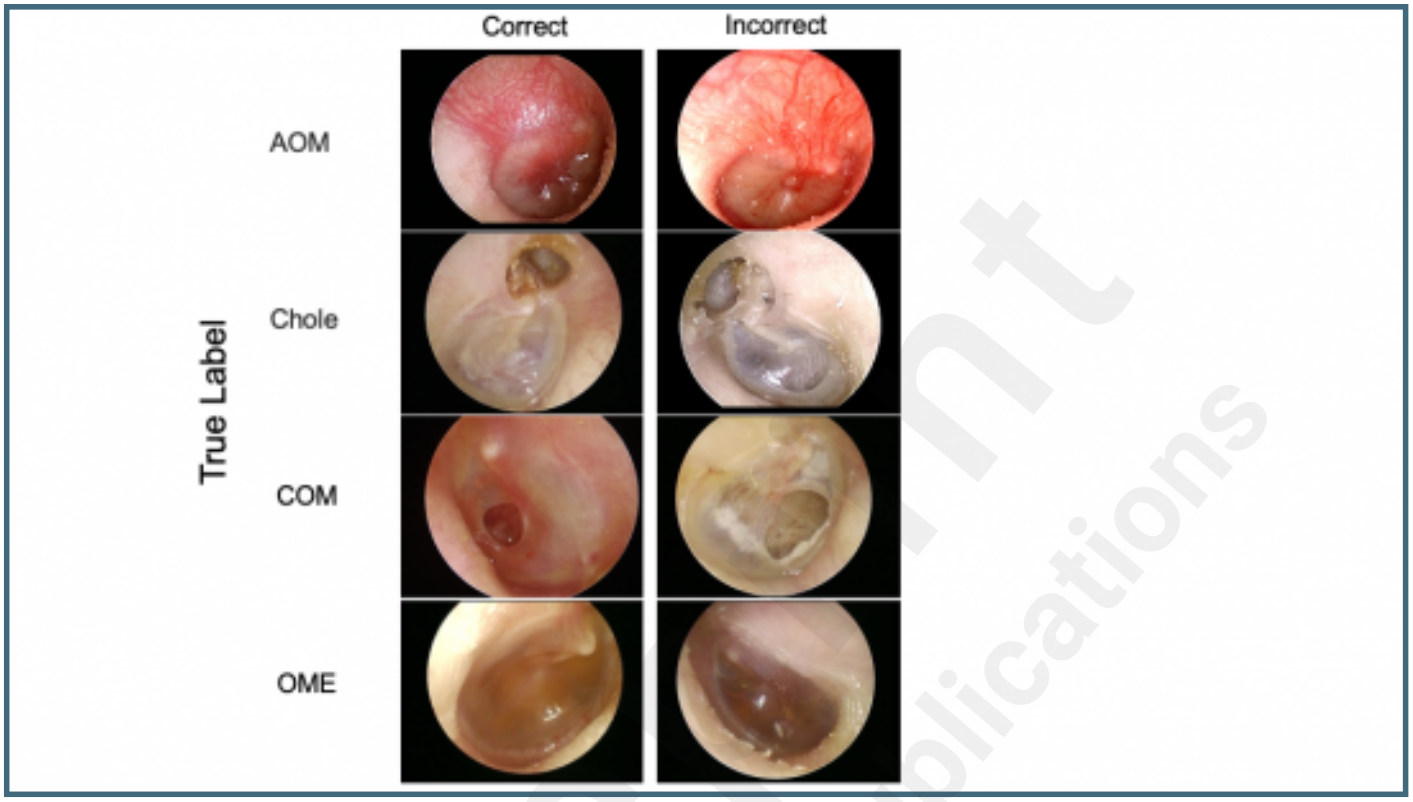
Overview of the study. The model was divided into two phases: (1) establishment and (2) tuned model validation. TM, Tympanic membrane; GPT-4V, Generative Pre-trained Transformer 4 Vision.

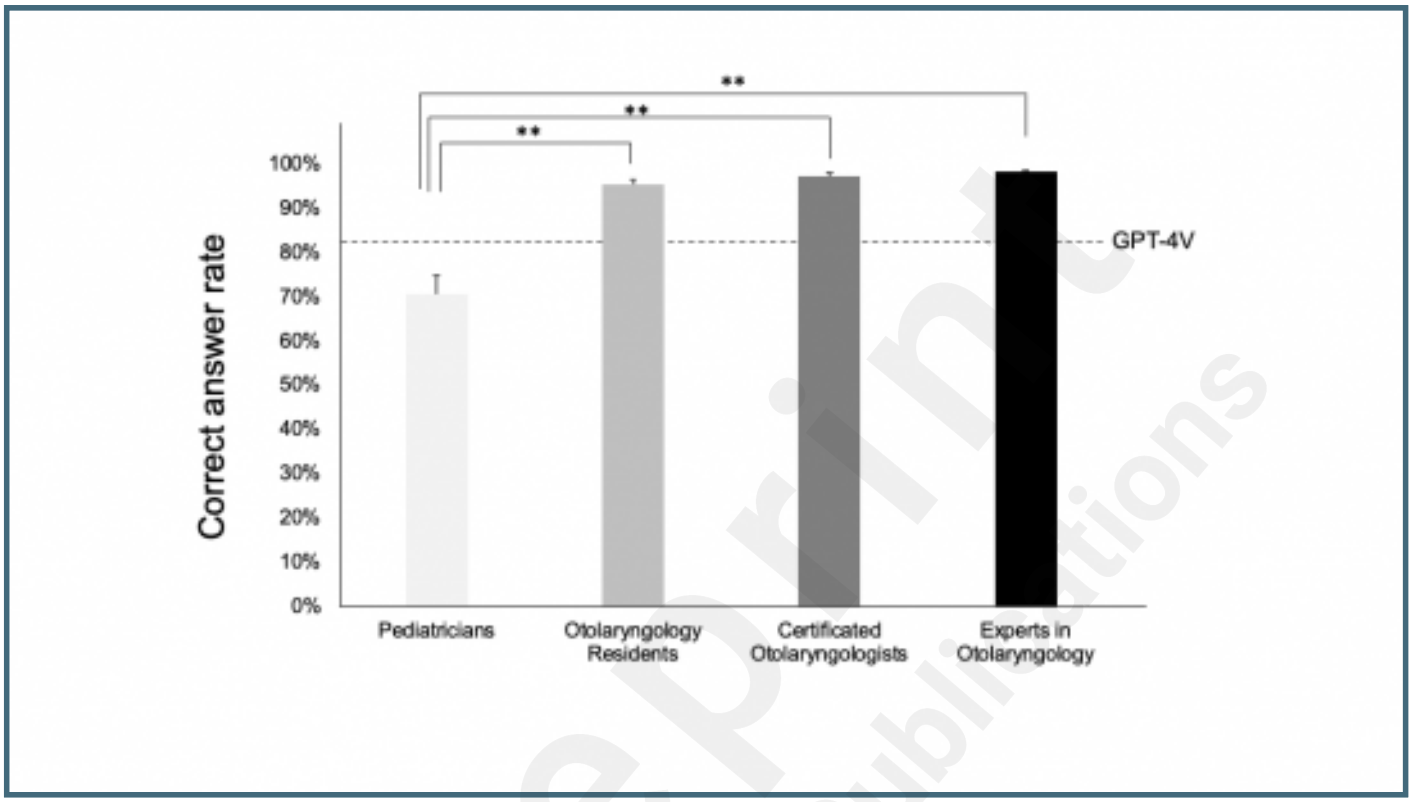


Confusion matrix of GPT-4V for classifying four middle ear diseases. AOM, acute otitis media; Chole, middle ear cholesteatoma; COM, chronic otitis media; OME, otitis media with effusion.

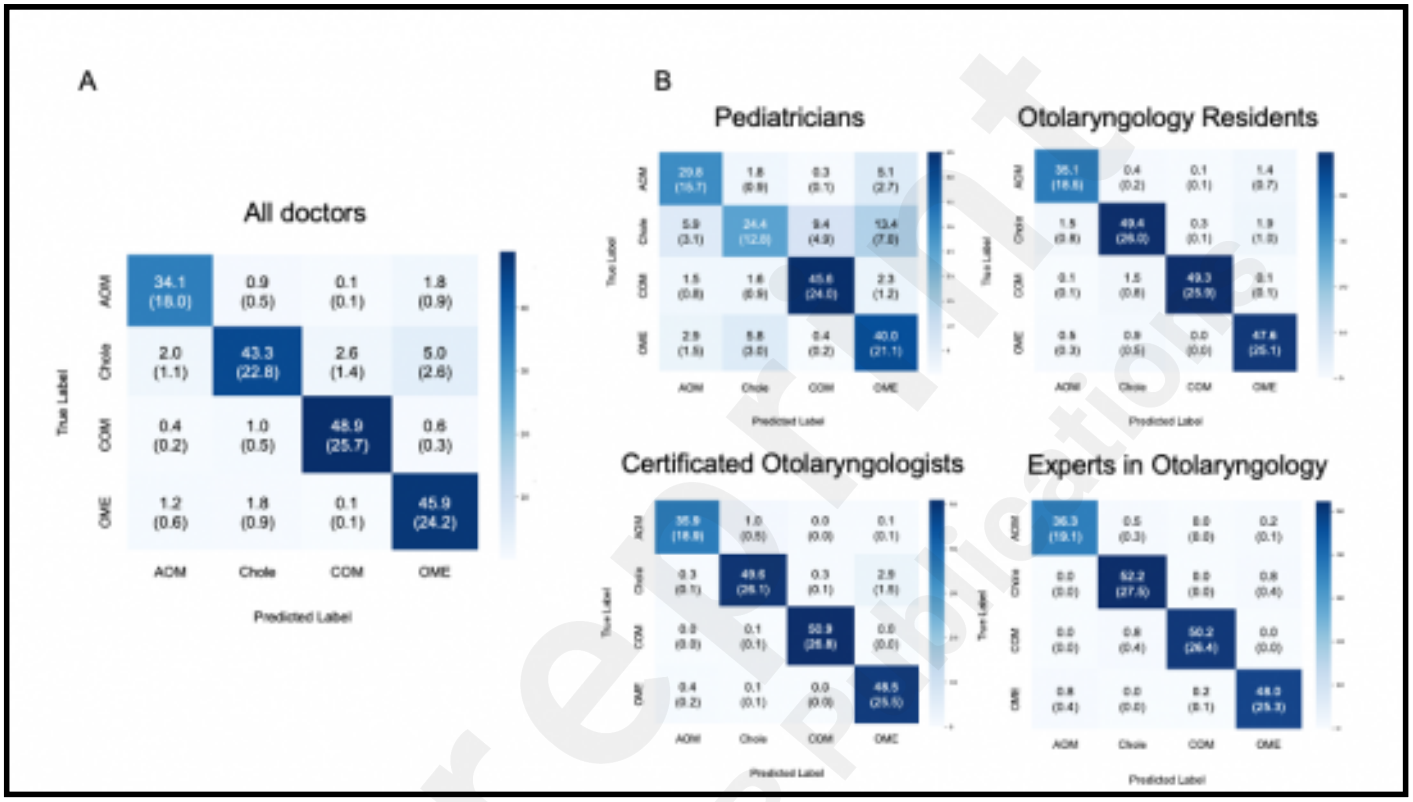


Representative images of correct and incorrect GPT-4V classifications for four middle ear diseases. The left side shows the correct images for GPT-4V classification, and the right side shows the incorrect images for GPT-4V. AOM, acute otitis media; Chole, middle ear cholesteatoma; COM, chronic otitis media; OME, otitis media with effusion.





Confusion matrix of doctors (pediatricians, otolaryngology residents, certificated otolaryngologists, and experts in otolaryngology) for classifying four middle ear diseases. A: Confusion matrix of all doctors (n=30). The average (percentage of total responses) is shown. B: Confusion matrix of doctors in each group: pediatricians (n=8), otolaryngology residents (n=8), certificated otolaryngologists (n=8), and experts in otolaryngology (n=6). The averages of each group (percentage of total responses) are shown. AOM, acute otitis media; Chole, cholesteatoma; COM, chronic otitis media; OME, otitis media with effusion.



Multimedia Appendixes

Representative image and prompt of this study A. Representative image of input and output to GPT-4V. Input can be combined with text and images in input to obtain output B. Example of changing the prompt content and an output that asks for patient information. By presenting a concept as ORDER and adding conditions as Restriction, appropriate prompts were attempted to be developed. In the output, it is required to input patient information such as age, medical history, and chief complaint. C. An example of an answer with an optimized prompt. Present the diagnosis, the rationale for the diagnosis, and treatment and prevention methods.

URL: http://asset.jmir.pub/assets/fc73d5fd02c2b585a34f0243fe2f9496.pdf

Related publication(s) - for reviewers eyes onlies

Graphical abstract of this research.