


Domain-Specific Customization for Language Models in Otolaryngology: The ENT GPT Assistant

Brenton T. Bicknell, BS¹ , Nicholas J. Rivers, MD², Adam Skelton, BS¹, Delaney Sheehan, MD², Charis Hodges, BS¹, Stevan C. Fairburn, BS¹, Benjamin J. Greene, MD², and Bharat Panuganti, MD³

Abstract

Objective. To develop and evaluate the effectiveness of domain-specific customization in large language models (LLMs) by assessing the performance of the ENT GPT Assistant (E-GPT-A), a model specifically tailored for otolaryngology.

Study Design. Comparative analysis using multiple-choice questions (MCQs) from established otolaryngology resources.

Setting. Tertiary care academic hospital.

Methods. Two hundred forty clinical-vignette style MCQs were sourced from BoardVitals Otolaryngology and OTOQuest, covering a range of otolaryngology subspecialties (n = 40 for each). The E-GPT-A was developed using targeted instructions and customized to otolaryngology. The performance of E-GPT-A was compared against top-performing and widely used artificial intelligence (AI) LLMs, including GPT-3.5, GPT-4, Claude 2.0, and Claude 2.1. Accuracy was assessed across subspecialties, varying question difficulty tiers, and in diagnostics and management.

Results. E-GPT-A achieved an overall accuracy of 74.6%, outperforming GPT-3.5 (60.4%), Claude 2.0 (61.7%), Claude 2.1 (60.8%), and GPT-4 (68.3%). The model performed best in allergy and rhinology (85.0%) and laryngology (82.5%), whereas showing lower accuracy in pediatrics (62.5%) and facial plastics/reconstructive surgery (67.5%). Accuracy also declined as question difficulty increased. The average correct response percentage among otolaryngologists and otolaryngology trainees was 71.1% in the question set.

Conclusion. This pilot study using the E-GPT-A demonstrates the potential benefits of domain-specific customizations of language models for otolaryngology. However, further development, continuous updates, and continued real-world validation are needed to fully assess the capabilities of LLMs in otolaryngology.

Keywords

artificial intelligence, comprehensive otolaryngology, machine learning, natural language processing

Received March 14, 2025; accepted April 18, 2025.

The field of otolaryngology (ORL) continues to experience rapid technological advancements, offering the potential to enhance patient outcomes and assist otolaryngologists in both clinical practice and research. One such emerging technology is large language models (LLMs). LLMs are a type of artificial intelligence (AI) designed to process and generate human-like text based on vast amounts of training data. Early investigations suggest that these models are experiencing rapid advancements^{1,2} and have shown preliminary capabilities in clinical applications, including the preparation of patient education materials,³ assistance in cancer staging of oral squamous cell carcinoma,⁴ and the enhancement of patient care efficiency through real-time charting, among other uses.⁵

Despite the promise of LLMs, there is a notable gap in research specifically evaluating custom-developed LLMs for otolaryngology. Several approaches exist to create LLMs tailored to the needs of otolaryngologists, such as building models from the ground up by fine-tuning parameters and algorithms, or by integrating existing LLMs with specialized application programming interfaces (APIs), custom instructions, or additional reference materials. For instance, one study in rhinology developed

¹UAB Heersink School of Medicine, University of Alabama at Birmingham, Birmingham, Alabama, USA

²Department of Otolaryngology-Head and Neck Surgery, University of Alabama at Birmingham, Birmingham, Alabama, USA

³Department of Otolaryngology-Head and Neck Surgery, Washington University in St. Louis, St. Louis, Missouri, USA

This article was presented at the AAO-HNSF 2024 Annual Meeting & OTO EXPO; September 27-October 1, 2024; Miami Beach, Florida.

Corresponding Author:

Brenton T. Bicknell, BS, UAB Heersink School of Medicine, University of Alabama at Birmingham, 1670 University Blvd, Birmingham, AL 35233, USA. Email: brentonb@uab.edu

a chatbot using the LangChain/OpenAI API, tailored with the *International Consensus Statement on Allergy and Rhinology: Rhinosinusitis* guidelines.⁶ This chatbot provided direct and actionable recommendations, demonstrating how consensus statements can be effectively leveraged in AI applications within health care.

Another approach avoided the high costs associated with retraining LLMs by investing time to develop a curated knowledge base from open resources.⁷ Although this was time-intensive, it yielded promising outcomes, including a significant reduction in error rates (up to 58%), improved safety with fewer hallucinations, and enhanced performance on multiple-choice and open-ended questions compared to general-purpose models. Additionally, this method allowed for seamless integration of up-to-date, evidence-based information, ensuring greater reliability and relevance for clinical and educational applications. However, such developments require considerable time and dedicated resources for effective implementation.

Custom-developed LLMs offer the advantage of being meticulously tuned to the specialized language, terminology, and nuances of medicine. This precision could result in greater accuracy, relevance, and utility in clinical settings compared to general-purpose models. When fine-tuned, these custom models have the potential to enhance clinical decision-making, improve efficiency, and ultimately improve patient care. However, the development process remains resource-intensive, as demonstrated by previous studies.^{6,7} There is little research investigating whether less development-intensive approaches, such as inputting targeted instructions, could achieve comparable improvements in model accuracy.

This study aimed to develop and evaluate a domain-specific LLM, the ENT GPT Assistant (E-GPT-A), to assess the benefits of LLM customization for otolaryngology through given instructions. The study focused on comparing the performance of E-GPT-A against leading general-purpose AI models to benchmark its effectiveness and identify areas where further refinement is needed for improved performance in specialized medical contexts.

Methods

Otolaryngology Multiple-Choice Questions

Two hundred forty clinical-vignette style multiple-choice questions (MCQs) assessing otolaryngology content were sourced from two established otolaryngology resources: OTOQuest and BoardVitals Otolaryngology. OTOQuest is a comprehensive self-assessment tool for otolaryngology, offering high-quality, case-based questions developed by experts in the field. It is available through OTO Logic and designed for both Continuing Medical Education (CME) credit and resident education. BoardVitals Otolaryngology is a board review tool offering more than 1200 practice questions with detailed explanations, specifically designed to help candidates prepare for the American Board of Otolaryngology (ABOto) Qualifying Exam, the

Otolaryngology Training Examination (OTE), and the American Osteopathic Board of Otolaryngology (AOBOO) Written Qualifying Exam.

Development of the E-GPT-A

The E-GPT-A was developed as a tailored application of OpenAI's GPT-4 model for otolaryngology-specific inquiries. Rather than building an entirely new language model, the E-GPT-A customization focused on refining the capabilities of GPT-4 by providing targeted instructions and integrating domain-specific enhancements through OpenAI's GPT builder platform. This streamlined approach balanced efficiency with technical sophistication and cost-effectiveness, tailoring the model's applicability to clinical otolaryngology scenarios.

Technical Framework of the Custom AI Software

The foundation of the E-GPT-A is based on GPT-4, a transformer-based language model that uses multihead self-attention mechanisms to process textual inputs. This architecture is the basis for the customization. Customization of the E-GPT-A focused on instruction optimization, which embeds specific operational directives into the model's framework to guide processing and response to inquiries. These instructions were integrated without altering the model's underlying parameters or retraining its algorithms, ensuring that its original computational framework remained intact. This approach allowed for efficient and targeted adaptation of GPT-4 to otolaryngology.

Instruction Optimization

The GPT builder platform facilitated the alignment of GPT-4's capabilities with otolaryngology-specific requirements by embedding a structured set of instructions. These instructions served as a roadmap for the model by leveraging prompt engineering and instruction embedding within the model's architecture rather than altering its underlying parameters or algorithms. Through the GPT builder platform, these instructions are integrated as predefined operational directives that dynamically guide the model's response generation during runtime. The model uses its existing parameters to interpret and execute these instructions, bypassing the need for direct API calls or retraining its algorithms. The instructions included prioritization of otolaryngology-related knowledge and guidance to reference authoritative and evidence-based sources (such as clinical guidelines, consensus statements, and peer-reviewed literature). These instructions, shown in Supplemental Figure S1, available online, were embedded into the model's operational framework, shaping its interpretive and generative processes.

Top AI LLMs for Comparison

For comparison, this study included other top-performing AI LLMs, including ChatGPT 3.5 (GPT-3.5), ChatGPT 4

(GPT-4), Claude 2.0 (Cld-2.0), and Claude 2.1 (Cld-2.1), to evaluate and compare their capabilities against the E-GPT-A.

Protocol Utilized for Assessment

The study employed a rigorous assessment framework developed following a review of prior methodologies. The AI LLMs were assessed using a standardized prompt: “Answer the following question and provide an explanation for your answer choice: [input question and multiple choice responses].” The selected answer and its explanation were recorded for each model. Since the study focused on text-based assessments, clinical vignettes that required interpretation of diagnostic imaging or involved histologic or gross examination findings were excluded.

Measures of Analysis

For each MCQ, we recorded the language models' chosen responses and their explanations, labeling the response based on its accuracy (categorized as “accurate” or “inaccurate”) in relation to the correct answer noted by the resource and verified by three reviewers. The following variables and measures were included for analysis.

Comparisons Between E-GPT-A, the LLMs, and Human Respondents

The study compared the performance of various models to determine whether the customized E-GPT-A demonstrated improved accuracy over the others. Additionally, we evaluated for differences between the overall performance of the models and the average performance of human respondents, presumably otolaryngology residents and practicing otolaryngologists.

Subspecialty

We categorized and compared the percentage of correct responses by subspecialty, including facial plastics/reconstructive, head and neck, pediatrics, allergy and rhinology, laryngology, and neurotology.

Influence of Question Difficulty

Based on the average percentage of correct responses by human respondents, each question was assigned a difficulty tier ranging from 1 (easiest, 75.0%-100% correctness by human respondents) to 4 (most difficult, 0%-24.9% correctness by human respondents).

Measuring Consistency and Reliability Over Time

The first wave of analysis of the language models was conducted in January 2024. To assess consistency and reliability over time, we repeated the assessments using the same protocol in June 2024. This comparison aimed to identify any variations in performance, including whether the models showed improvement, made the same choices, or exhibited any changes.

Domains of Diagnostics and Management

Of the 240 MCQs, 198 were subcategorized based on their question stems into two clinical domains: diagnostics (82 vignettes) and management (116 vignettes). The diagnostic MCQs assessed the ability to identify the most likely diagnosis, recognize signs and symptoms of specific diagnoses, and determine the best diagnostic tests. The management MCQs evaluated the choice of the next treatment step, risk prevention, ongoing care, surgical indications, and patient counseling.

Statistical Analysis

IBM SPSS Statistics version 29.0.1.1 was utilized for all statistical analyses. The threshold for statistical significance was set at $P < .05$. Pearson's chi-square tests were employed to assess differences across the AI language models. For the AI models, the confidence intervals reflect the margin of error for the predicted overall proportion with 95% confidence. For human respondents, the confidence intervals represent the margin of error for the estimated overall average percent correct within 95% confidence.

Ethical Considerations

The study was deemed not human subjects research by the University of Alabama at Birmingham Institutional Review Board (IRB-300011939).

Results

The E-GPT-A demonstrated the highest overall accuracy among the AI LLMs, achieving 74.6% (179/240; 95% CI, 69.1%-80.1%), as shown in **Table 1**. This accuracy was significantly higher than that of GPT-3.5 ($P < .001$), which achieved 60.4% (95% CI, 54.2%-66.6%), Claude 2.0 ($P < .001$) at 61.7% (95% CI, 58.5%-66.2%), and Claude 2.1 ($P < .001$) at 60.8% (95% CI, 57.5%-66.2%) (all $P < .001$). Although the average score of E-GPT-A was higher than that of GPT-4, which had an accuracy of 68.3% (95% CI, 62.4%-74.2%), this difference was not statistically significant ($P = .13$). A comparison of the performance of E-GPT-A, the other language, and human respondents is shown in **Figure 1**.

Performance Across Otolaryngology Subspecialties and Resources

Comparison of Performance Across Subspecialties

E-GPT-A consistently performed well across subspecialties, achieving its highest accuracies in allergy and rhinology (85.0%) and laryngology (82.5%). The lowest accuracies for E-GPT-A were observed in pediatrics (62.5%) and facial plastics/reconstructive surgery (67.5%). Across all models, the top overall accuracies were achieved in allergy and rhinology (69.0%) and laryngology (68.5%), whereas the lowest performances

Table 1. Performance Overall and by Subspecialty of ENT GPT Assistant (E-GPT-A) and Top Artificial Intelligence Language Models

Category	Respondents		All language models		Language models response accuracy, no. (%) correct					
	Avg. correct (95% CI)	No.	% correct (95% CI)		E-GPT-A	GPT-3.5	GPT-4	Cld 2.0	Cld 2.1	
Overall accuracy	71.1% (68.5%-73.7%)	786	65.5% (62.8%-68.2%)		179 (74.6%)	145 (60.4%)	164 (68.3%)	148 (61.7%)	146 (60.8%)	
OTOQuest CME (n = 120)	71.7% (68.1%-75.3%)	396	66.0% (62.2%-69.8%)		85 (70.8%)	77 (64.2%)	78 (65.0%)	76 (63.3%)	75 (62.5%)	
BoardVitals (n = 120)	70.5% (66.7%-74.3%)	391	65.2% (61.4%-69.0%)		94 (78.3%)	68 (56.7%)	86 (71.7%)	72 (60.0%)	71 (59.2%)	
Subspecialties										
Facial plastics/reconstructive (n = 40)	67.0% (60.3%-73.6%)	113	56.5% (49.6%-63.4%)		27 (67.5%)	23 (57.5%)	22 (55.0%)	21 (52.5%)	20 (50.0%)	
Head and neck (n = 40)	69.1% (62.8%-75.3%)	134	67.0% (60.5%-73.5%)		29 (72.5%)	26 (65.0%)	30 (75.0%)	26 (65.0%)	23 (57.5%)	
Pediatrics (n = 40)	73.9% (66.8%-80.9%)	127	63.5% (56.8%-70.2%)		25 (62.5%)	26 (65.0%)	26 (65.0%)	25 (62.5%)	24 (60.0%)	
Allergy and rhinology (n = 40)	76.9% (71.1%-82.6%)	138	69.0% (62.6%-75.4%)		34 (85.0%)	23 (57.5%)	30 (75.0%)	25 (62.5%)	26 (65.0%)	
Laryngology (n = 40)	67.4% (59.8%-74.9%)	137	68.5% (62.1%-74.9%)		33 (82.5%)	27 (67.5%)	26 (65.0%)	24 (60.0%)	27 (67.5%)	
Neurotology (n = 40)	72.5% (66.6%-78.3%)	134	67.0% (60.5%-73.5%)		31 (77.5%)	20 (50.0%)	30 (75.0%)	27 (67.5%)	26 (65.0%)	

were seen in pediatrics (63.5%) and facial plastics/reconstructive (56.5%) (**Table 1**).

Comparison of Performance Between Otolaryngology Resources

The overall response accuracy for human respondents was similar between OTOQuest (71.7%; 95% CI, 68.1%-75.3%) and BoardVitals Otolaryngology (70.5%; 95% CI, 66.7%-74.3%), with the AI language models achieving the same average accuracy of 65.3% (95% CI, 60.6%-70.0%) across both resources, shown in **Table 2**. On OTOQuest, E-GPT-A led with an accuracy of 70.8%, followed by ChatGPT 4 at 65.0%, and ChatGPT 3.5 at 64.2%. Claude 2.0 and Claude 2.1 had lower accuracies, 63.3% and 62.5%, respectively. For BoardVitals Otolaryngology, E-GPT-A again achieved the highest accuracy at 78.3%, with ChatGPT 4 following at 71.7%. ChatGPT 3.5, Claude 2.0, and Claude 2.1 performed worse, with accuracies of 56.7%, 60.0%, and 59.2%, respectively.

Influence of Question Difficulty

The performance of the models varied significantly by question difficulty tier (**Table 3**). The overall response accuracy decreased from 74.0% for tier 1 questions to 40.0% for tier 4 questions ($P < .001$). E-GPT-A exhibited varying response accuracies by difficulty tier ($P < .001$) with a clear trend of decreasing accuracy with increasing question difficulty, achieving 86.8% for tier 1, 67.1% for tier 2, 58.8% for tier 3, and 42.9% for tier 4.

Reliability and Accuracy Over Time

E-GPT-A maintained a stable accuracy between round 1 (74.6%) and round 2 (73.8%), indicating no significant change in overall accuracy between the two rounds (**Table 4**). A similar trend was seen for GPT-3.5, GPT-4, and Claude 2.1. In contrast, Claude 2.0 demonstrated a significant improvement in accuracy from 61.7% to 70.4% ($P = .04$). E-GPT-A exhibits 188 consistent responses and 52 changes, with a P -value of $< .001$, indicating that the model's answers were not likely to be the same every time. A similar pattern of consistency was observed in the other models.

Diagnostics and Management Domains

In the diagnostics and workup category, the E-GPT-A and ChatGPT 4 demonstrated the highest accuracy, each correctly answering 58 out of 82 questions (70.7%). This performance was higher than that of Claude 2.0, Claude 2.1, and GPT-3.5, which achieved accuracies of 64.6%, 63.4%, and 54.9%, respectively (**Table 5**).

In the management and treatment category, E-GPT-A again exhibited the highest accuracy, correctly answering 89 out of 116 questions (76.7%). This was followed by ChatGPT 4, which achieved an accuracy of 69.0%. Claude 2.1 and Claude 2.0 demonstrated lower accuracies

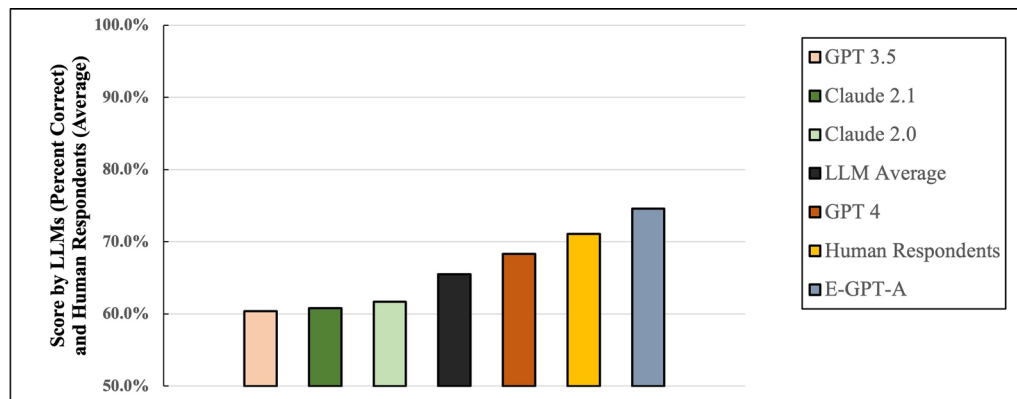


Figure 1. Comparative accuracy of large language models (LLMs) and human respondents on otolaryngology multiple-choice questions. This figure compares the accuracy of LLMs (ENT GPT Assistant [E-GPT-A], GPT-3.5, GPT-4, Claude 2.0, and Claude 2.1) with human respondents, showing percentage correct responses.

of 64.7% and 62.1%, respectively, whereas ChatGPT 3.5 had a slightly higher accuracy of 65.5% (**Table 5**).

Discussion

The findings of this study highlight the potential of domain-specific LLM customization, demonstrated by the performance of the E-GPT-A. Achieving an accuracy of 74.6% on otolaryngology-specific MCQs, the results underscore the potential benefits of customization in improving the relevance and accuracy of LLMs in specialized fields. By leveraging targeted instructions to otolaryngology-specific applications, the E-GPT-A achieved higher accuracy than general-purpose models in answering clinical MCQs. This represents a promising, cost- and time-efficient approach to enhancing the utility of LLMs in specialized medical fields, eliminating the need for resource-intensive retraining or complete model redevelopment.

These results align with emerging literature on the utility of specialized LLMs in health care.⁶ Customization through provided instructions allows for the fine-tuning of language models to the specific language, terminology, and clinical nuances inherent in a particular medical field, thereby improving the model's ability to provide contextually appropriate—and presumably more accurate—responses. For the field of otolaryngology, this improved accuracy can enhance the reliability of AI, making integration into clinical workflows more realistic, as dependable, accurate information is critical for supporting clinical decision-making.

We speculate that the absence of statistical significance in the improved accuracy of the E-GPT-A over GPT-4 might be attributed to the limited sample size of questions utilized in this study. The study's use of only 240 questions was likely not sufficient to detect this significant difference of 6.3% between the overall performance of GPT-4 and E-GPT-A. Additionally, domain-specific models are often designed to prioritize contextual

relevance over absolute accuracy, which can result in differences in performance depending on the nature of the evaluation criteria. Future studies with larger question sets and additional benchmarking techniques are warranted to further assess and optimize the capabilities of domain-specific models like E-GPT-A.

Another critical finding of this study is the impact of question difficulty on model performance and inconsistency of responses, which further highlights areas for future development. The accuracy of E-GPT-A declined significantly from 86.8% for tier 1 (easiest) questions to 42.9% for tier 4 (most difficult) questions. This trend was consistent across all tested models and points to a common limitation among current LLMs in handling complex clinical scenarios requiring deeper analytical reasoning and synthesis. Addressing this limitation may involve incorporating additional training data focused on complex cases or integrating multimodal inputs, such as imaging data, to enhance the model's reasoning capabilities. Moreover, E-GPT-A provided a different answer on a second run of questions 5 months after the initial run to 27.7% of inquiries, though no information had changed. However, it is of note that the accuracy on the second run was not significantly different, only 0.8% less.

The implications of these findings are important for the future development of LLMs in otolaryngology. Although E-GPT-A shows promise in handling otolaryngology-specific queries with improved accuracy over general models, developing models that integrate not only context-specific instructions but a combination of APIs, customized instructions, and authoritative resources directly into the model itself could lead to even greater accuracy—and, hence, reliability for clinical applications. The potential utility of text-based language models in the clinic setting may lie in assisting with educational tools, such as training programs or study resources for residents and clinicians pursuing CME, among other clinical applications. Future iterations of

Table 2. Characterization of Response Accuracies by Resource

Category	Respondents		All language models		Language models response accuracy, no. (%) correct					
	Avg. correct (95% CI)	No.	% correct (95% CI)		E-GPT-A	GPT 3.5	GPT 4	Cld 2.0	Claude 2.1	
OTOQuest	71.7% (68.1%-75.3%)	391	65.3% (60.6%-70.0%)		85 (70.8%)	77 (64.2%)	78 (65.0%)	76 (63.3%)	75 (62.5%)	
Facial plastics/reconstructive (n = 20)	65.9% (56.7%-75.0%)	62	51.7% (42.7%-60.6%)		14 (70.0%)	13 (65.0%)	11 (55.0%)	11 (55.0%)	13 (65.0%)	
Head and neck (n = 20)	72.1% (65.0%-79.2%)	70	58.3% (49.5%-67.2%)		14 (70.0%)	14 (70.0%)	14 (70.0%)	15 (75.0%)	13 (65.0%)	
Pediatrics (n = 20)	78.4% (69.2%-87.5%)	56	46.7% (37.7%-55.6%)		11 (55.0%)	11 (55.0%)	12 (60.0%)	11 (55.0%)	11 (55.0%)	
Allergy and rhinology (n = 20)	80.1% (72.1%-88.1%)	63	52.5% (43.6%-61.4%)		15 (75.0%)	13 (65.0%)	14 (70.0%)	11 (55.0%)	10 (50.0%)	
Laryngology (n = 20)	64.3% (52.3%-76.3%)	68	56.7% (47.8%-65.5%)		15 (75.0%)	14 (70.0%)	11 (55.0%)	13 (65.0%)	15 (75.0%)	
Neurotology (n = 20)	69.4% (60.6%-78.3%)	72	60.0% (51.2%-68.8%)		16 (80.0%)	12 (60.0%)	16 (80.0%)	15 (75.0%)	13 (65.0%)	
BoardVitals	70.5% (66.7%-74.3%)	391	65.3% (60.6%-70.0%)		94 (78.3%)	68 (56.7%)	86 (71.7%)	72 (60.0%)	71 (59.2%)	
Facial plastics/reconstructive (n = 20)	68.1% (57.7%-78.4%)	51	42.5% (33.7%-51.3%)		13 (65.0%)	10 (50.0%)	11 (55.0%)	10 (50.0%)	7 (35.0%)	
Head and neck (n = 20)	66.0% (55.1%-76.9%)	64	53.3% (44.4%-62.3%)		15 (75.0%)	12 (60.0%)	16 (80.0%)	11 (55.0%)	10 (50.0%)	
Pediatrics (n = 20)	69.4% (58.2%-80.6%)	70	58.3% (49.5%-67.2%)		14 (70.0%)	15 (75.0%)	14 (70.0%)	14 (70.0%)	13 (65.0%)	
Allergy and rhinology (n = 20)	73.6% (64.9%-82.3%)	75	62.5% (53.8%-71.2%)		19 (95.0%)	10 (50.0%)	16 (80.0%)	14 (70.0%)	16 (80.0%)	
Laryngology (n = 20)	70.4% (60.3%-80.5%)	69	57.5% (48.7%-66.3%)		18 (90.0%)	13 (65.0%)	15 (75.0%)	11 (55.0%)	12 (60.0%)	
Neurotology (n = 20)	75.5% (67.3%-83.8%)	62	51.7% (42.7%-60.6%)		15 (75.0%)	8 (40.0%)	14 (70.0%)	12 (60.0%)	13 (65.0%)	

Abbreviation: E-GPT-A, ENT GPT Assistant.

E-GPT-A and similar clinical applications should integrate multimodal data sources, including diagnostic imaging, histopathology, and real-time clinical parameters. Although incorporating such data presents technical challenges, it has the potential to significantly enhance AI-driven diagnostic accuracy and expand its applicability beyond text-based queries.

However, several limitations must be acknowledged. The study's reliance on text-based MCQs, although useful for standardizing comparisons across models, does not fully capture the complexities of real-world clinical practice, where decision-making often involves the integration of real-time, multimodal data that includes imaging and laboratory results. The study's controlled environment may not reflect the dynamic and variable conditions encountered in clinical practice, where factors such as time pressure and patient variability are critical. The rapid development of AI technology also means that the findings presented here may quickly become outdated, necessitating ongoing evaluation and adaptation of the models.

The rapid advancements in AI technology necessitate regular updates to domain-specific models like E-GPT-A. As general-purpose LLMs evolve, customizing improved and field-specific models would ideally outpace and outperform these developments to remain relevant and effective. Future research should prioritize continuously updating the model's training data as new clinical guidelines and research emerge, while also exploring the integration of more complex data types and real-time clinical scenarios. Expanding the scope of assessment to include multimodal data and testing the model across a broader range of otolaryngology conditions, including rare and complex cases, will further validate the benefits of domain-specific customization in improving LLM performance in specialized fields. Additional abilities that will be of importance are being able to recognize findings related to diagnostic imaging, histopathological data, and real-time clinical data (ie, case scenarios). The ability to customize AI for ENT—through developing models that integrate a combination of APIs, customized instructions, and authoritative resources—could be the key to developing a valuable tool, such as a text-based clinical AI assistant, for otolaryngologists.

Conclusion

This pilot study of the E-GPT-A demonstrates the potential of domain-specific customization for LLMs in otolaryngology. Its improved accuracy over general-purpose models highlights the benefits of tailoring AI to specialized fields through customized instructions. However, to fully realize the capabilities of language models as a text-based clinical assistant for otolaryngology, further development, as well as ongoing refinement, real-world validation, and the integration of multimodal data are essential. With continued advancements, these models could soon be a valuable asset in the toolkit of the otolaryngologist and otolaryngology trainee.

Table 3. Influence of Difficulty on Language Model Response Accuracy

Language model	Response accuracy by difficulty tier, % (95% CI) correct				P
	Tier 1 (n = 114)	Tier 2 (n = 85)	Tier 3 (n = 34)	Tier 4 (n = 7)	
All LLMs	78.6% (448)	56.7% (241)	48.8% (83)	28.6% (10)	<.001
E-GPT-A	86.8% (99)	67.1% (57)	58.8% (20)	42.9% (3)	<.001
ChatGPT 3.5	73.7% (84)	49.4% (42)	50.0% (17)	28.6% (2)	<.001
ChatGPT 4	86.0% (98)	57.6% (49)	44.1% (15)	28.6% (2)	<.001
Claude 2.0	72.8% (83)	58.8% (50)	41.2% (14)	14.3% (1)	<.001
Claude 2.1	73.7% (84)	50.6% (43)	50.0% (17)	28.6% (2)	<.001

Abbreviations: E-GPT-A, ENT GPT Assistant; LLM, large language model.

Table 4. Accuracy and Reliability of the ENT GPT Assistant (E-GPT-A) and Other Language Models Over Runs

Language model	Accuracy, % correct (95% CI)			Consistency, response changes			Characterization of response changes			
	Round 1	Round 2	P	Unchanged	Changed	P	COR twice	INC twice	COR to INC	INC to COR
E-GPT-A	179 (74.6%)	177 (73.8%)	.83	188	52	<.001	152	36	27	25
ChatGPT 3.5	145 (60.4%)	140 (58.3%)	.64	183	57	<.001	114	69	31	26
ChatGPT 4	164 (68.3%)	161 (67.1%)	.77	191	49	<.001	138	53	26	23
Claude 2.0	148 (61.7%)	169 (70.4%)	.04	181	59	<.001	129	52	19	40
Claude 2.1	146 (60.8%)	166 (69.2%)	.06	192	48	<.001	132	60	14	34

Abbreviations: CI, confidence interval; COR, correct; INC, incorrect;

Table 5. Large Language Model Performance in Diagnostics and Management

Language model performance	No. (%) correct
Diagnostics and workup (n = 82)	
E-GPT-A	58 (70.7%)
ChatGPT 3.5	45 (54.9%)
ChatGPT 4	58 (70.7%)
Claude 2.0	53 (64.6%)
Claude 2.1	52 (63.4%)
Management and treatment (n = 116)	
E-GPT-A	89 (76.7%)
ChatGPT 3.5	76 (65.5%)
ChatGPT 4	80 (69.0%)
Claude 2.0	72 (62.1%)
Claude 2.1	75 (64.7%)

Abbreviation: E-GPT-A, ENT GPT Assistant.

Acknowledgments

The authors would like to thank OTOQuest and BoardVitals for providing the questions used to perform the assessment of these AI models.

Author Contributions

Brenton T. Bicknell, contributed to the conceptualization and design of the study, led the data acquisition process, conducted data analysis, and drafted the manuscript, including the

preparation of tables and figures; **Nicholas J. Rivers**, played a key role in the study's conceptualization and design, provided expert interpretation of the data, and was instrumental in the critical revision of the manuscript, ensuring the accuracy and clinical relevance of the content; **Adam Skelton**, played a significant role in data acquisition and drafting of the manuscript; **Delaney Sheehan**, contributed to the study design, the analysis and interpretation of data, and reviewed the manuscript; **Charis Hodges**, involved in data acquisition and drafting of the manuscript; **Stevan C. Fairburn**, involved in the detailed data analysis, contributed to the interpretation of results, and assisted in the preparation and revision of the manuscript; **Benjamin J. Greene**, contributed to the study design, data interpretation, and manuscript revision; **Bharat Panuganti**, provided overall supervision of the study, contributed to its conceptualization and design, guided the interpretation of data, and was heavily involved in the critical revision of the final manuscript.

Disclosures

Competing interests: The authors declare that there are no conflicts of interest.

Funding source: None.


Data Availability Statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request, contingent on access to the original resources.

Supplemental Material

Additional supporting information is available in the online version of the article.

ORCID iD

Brenton T. Bicknell  <http://orcid.org/0009-0000-2101-2543>

References

1. Ye F, Zhang H, Luo X, Wu T, Yang Q, Shi Z. Evaluating ChatGPT's performance in answering questions about allergic rhinitis and chronic rhinosinusitis. *Otolaryngol Head Neck Surg.* 2024;171(2):571-577. doi:10.1002/ohn.832
2. Zalzal HG, Cheng J, Shah RK. Evaluating the current ability of ChatGPT to assist in professional otolaryngology education. *OTO Open.* 2023;7(4):e94. doi:10.1002/oto2.94
3. Hill GS, Fischer JL, Watson NL, Riley CA, Tolisano AM. Assessing the quality of artificial intelligence-generated patient counseling for rhinosinusitis. *Int Forum Allergy Rhinol.* 2024;14:1634-1637. doi:10.1002/alr.23387
4. Baran E, Lee M, Aviv S, et al. Oropharyngeal cancer staging health record extraction using artificial intelligence. *JAMA Otolaryngol Head Neck Surg.* 2024;150:1051. doi:10.1001/jamaoto.2024.1201
5. Tierney AA, Gayre G, Hoberman B, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catal Innov Care Deliv.* 2024;5(3). doi:10.1056/CAT.23.0404
6. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med.* 2023;3(1):141. doi:10.1038/s43856-023-00370-1
7. Long C, Subburam D, Lowe K, et al. ChatENT: augmented large language model for expert knowledge retrieval in otolaryngology-head and neck surgery. *Otolaryngol Head Neck Surg.* 2024;171(4):1042-1051. doi:10.1002/ohn.864