Global Journal of
Otolaryngology
ISSN 2474-7556

Juniper PUBLISHERS
key to the Researchers

Research Article
Volume 28 Issue 4- March 2026
DOI: 10.19080/GJO.2026.28.556241

Glob J Otolaryngol

Can AI Match Expert Consensus? A Comparison with 2025 ATA Guidelines in Differentiated Thyroid Cancer

Emily Kwon*, Amina Shaikh, Alexandra Filipkowski, Annie Xu, Sudeepti Vedula, Brandon Gold, Rachel Kaye, Wayne D Hsueh

Rutgers New Jersey Medical School, Newark, New Jersey, USA

Submission: February 19, 2026; **Published**: March 03, 2026

***Corresponding author:** Emily Kwon, Rutgers New Jersey Medical School, Newark, New Jersey, USA

Abstract

Objective: To evaluate and compare artificial intelligence-generated clinical recommendations against the 2025 American Thyroid Association (ATA) Management Guidelines for Adult Patients with Differentiated Thyroid Cancer (DTC).

Study Design: Cross-sectional analysis using ChatGPT, Google Gemini, OpenEvidence (OE), and DTC guidelines.

Methods: 39 clinical queries were derived from the ATA guidelines and entered into ChatGPT, Gemini, and OE. The ATA guidelines served as the gold standard for comparison. Two physician raters, blinded to the response source, evaluated responses for accuracy, comprehensiveness, patient safety, and appropriateness using a 5-point Likert scale (1=strongly disagree, 5=strongly agree). Mixed-effects modeling compared AI-generated outputs. Readability scores are presented as Tukey-adjusted pairwise comparisons and word count. Analyses were performed using RStudio version 4.5.0.

Results: OE demonstrated greater concordance with the ATA guidelines than both ChatGPT and Gemini. For accuracy, comprehensiveness, and safety, OE responses had significantly higher odds of receiving higher ratings compared with Chat and Gemini (OR = 5-50, all p<0.01), while no significant difference was observed between Chat and Gemini (p>0.05). In contrast, for patient-appropriateness, Chat received significantly higher ratings than OE (OR=2.7, p=0.006), while Gemini did not differ significantly from either platform (both p>0.05). All platforms demonstrated readability scores that exceeded recommended levels.

Conclusion: There is strong alignment between AI-generated responses and ATA guidelines for DTC management. OE provided outputs most concordant with the ATA guidelines throughout accuracy, comprehensiveness, and patient safety for clinicians. AI platforms offer promising adjunctive tools to assist clinicians in the management of adult patients with DTC.

Keywords: Consensus; Thyroid Cancer; Artificial Intelligence

Abbreviations: ATA: American Thyroid Association; DTC: Differentiated Thyroid Cancer; AI: Artificial Intelligence; LLMs: Large Language Models; OE: Open Evidence, FRES: Flesch Reading Ease Score; FKGL: Flesch Kincaid Grade Level; GFI: Gunning Fog Index

Introduction

Differentiated thyroid cancer (DTC) is the most common endocrine malignancy, and its management continues to evolve as new evidence refines indications for DTC surgery and surveillance. The 2025 American Thyroid Association (ATA) Management Guidelines for Adult Patients with DTC remain the gold-standard reference for clinicians, providing evidence-based recommendations that inform risk assessments, planning of treatments and follow-up [1]. However, the expanding complexity nuances, and addendums to these guidelines, presents challenges for physicians to maintain care that is concordant with ATA guidelines [1]. Artificial intelligence (AI), particularly large language models (LLMs), has emerged as a promising tool in both medical education and clinical decision-making. These models, such as GPT 5, Claude, and Med-PaLM have the ability to synthesize large volumes of information and generate concise, evidence-based recommendations. Other AI-powered clinical platforms, including OpenEvidence (OE), aim to provide real-time, research-based decision support for clinicians at the point of care in real time. These systems can rapidly aggregate information and generate clinical recommendations, making them advantageous for evidence-based practice reinforcement. However, the reliability, accuracy, and safety of LLMs due to its novelty and under exploration in specialized domains of medicine, like DTC, remain incompletely understood [2].

Studies in lung cancer low-dose CT, orthopedic trauma management, pediatric acute-care decision-support, and dermatology skin-lesion triage have demonstrated that AI tools can often mirror expert guidelines, yet may also diverge in subtle but clinically meaningful ways [3–6]. Collectively, these investigations highlight the potential of AI to align with, and occasionally deviate from, expert consensus, underscoring the need for specialty and AI domain specific validation. To date, no study has evaluated AI alignment with the updated 2025 ATA DTC guidelines and evaluated the use of OE in thyroid cancer management. Given the significant increase in usage of AI resources by clinicians and patients, understanding AI reliability, accuracy, and safety in regards to DTC is timely and important.

Materials and Methods

Data Collection

The 2025 ATA guidelines were chosen as the gold standard for this study as it was developed by a taskforce selected for their expertise on the topic and reviewed by the ATA® Board of Directors [1]. ATA was accessed on September 10th, 2025, and information for initial DTC management was accessed [1]. Under the heading “Initial Management of Thyroid Cancer,” statements convey recommendations for initial treatment decisions; assessment of treatment responses; monitoring approaches; diagnostic testing; and subsequent therapies based on the strength of evidence for response and consideration of side-effects and outcomes. Each subsection was titled as a question e.g. “Does surgical experience influence complication rates for thyroidectomy?”. DTC statements and their associated questions were downloaded onto a Microsoft Word (V.16.86, Microsoft Corporation) document for a total of 39 questions.

AI Response Generation

On October 2nd, 2025, all 39 questions were individually queried to the ChatGPT-5 (released August 2025) with a “new chat” to avoid bias and influence on subsequent questions within the same chat. Responses were copied and pasted into the Word document under their respective questions, labeled as “Response A.” All 39 questions were then separately queried via their own “Incognito Window” in Google Gemini 2.5 Flash (released March 2025) into a new Word document under their respective questions, labeled as “Response B.” All 39 questions were then separately queried in OE 2.0 (released December 2024) into a new Word document under their respective questions, labeled as “Response C.” All text formatting was removed from these sources so as to not distinguish responses based on appearance. All questions were separately queried in Microsoft Copilot (released April 2025) and initially given unblinded to physician raters to discuss and come to consensus on grading for the three blinded LLM responses.

Physician Assessment

For each prompt response on the Word document (A, B, or C), we graded based on the following categories: (1) This response was accurate, (2) This response was comprehensive, (3) This response was appropriate for patient-level information, and (4) This response provided safe, non-dangerous information to the patient. The Likert scale was established as 1 (strongly disagree), 2 (disagree), 3 (neutral), 4 (agree), and 5 (strongly agree). Responses were graded by 2 otolaryngology resident physicians (BG and SV). The ATA guidelines served as the gold standard for comparison. Each resident physician was blinded to the other reviewer’s responses and had access to only their own version of the original Word document so as to not be influenced by each other’s responses. All physicians were blinded to what response A, B, or C correlated with in regard to AI models. After each physician completed their gradings, results were input into a Microsoft Excel (Version 16.78.3, Microsoft Corporation).

Readability Assessment

The tool WebFX Readability was used to generate Flesch Reading Ease Score (FRES), Flesch Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), and average word count [7,8]. FRES is a score measured on a scale of 0-100 that measures the reading difficulty of a text based on word count and syllables, with higher scores corresponding to easier readability. An FRES of above 80 correlates with a 6th grade reading level. FKGL is also measured based on word count and syllables but is scored on a 0-18 scale. Each FKGL score correlates with a specific grade level, with numbers 0-12 correlated with their respective grade levels and numbers above 12 with college and graduate levels. GFI is another linguistic measure that focuses on word complexity and sentence length and is measured on a scale from 6-17 where, similar to FKGL, each number correlates with grade level [9].

Statistical Analysis

Descriptive statistics are presented as mean Likert scores for ease of interpretation with one-way ANOVA for comparison of means. Because Likert ratings are ordinal and observations were clustered by grader and query, platform comparisons were conducted using cumulative logit mixed-effects models, with results reported as odds ratios. Readability scores are presented as Tukey-adjusted pairwise comparison for FRES, FKGL, GFI, and word count across AI platforms and ATA guidelines. Analysis was performed using R Studio version 4.5.0. Statistical significance was set as $p<0.05$.

Table 1: Average Likert Ratings by AI Platform.

Platform	Accuracy		Comprehensiveness		Patient Appropriate		Patient Safety	
	Mean (SD)	p-value	Mean (SD)	p-value	Mean (SD)	p-value	Mean (SD)	p-value
Chat	4.53 (0.96)	**<0.001**	4.55 (0.88)	**0.001**	4.03 (1.10)	**0.052**	4.60 (0.87)	**0.025**
Gemini	4.64 (0.85)		4.62 (0.86)		3.76 (1.11)		4.64 (0.82)	
OE	4.99 (0.11)		4.92 (0.42)		3.56 (1.20)		4.88 (0.46)	

Chat: ChatGPT, OE: OpenEvidence.

Table 2: Odds of Receiving Higher Likert Ratings between AI Platforms.

Metric	Chat vs Gemini		Chat vs OE		Gemini vs OE	
	OR [95% CI]	p-value	OR [95% CI]	p-value	OR [95% CI]	p-value
Accuracy	0.61 [0.27, 1.37]	0.457	0.02 [0.00, 0.17]	<0.001	0.04 [0.00, 0.28]	**0.005**
Comprehensiveness	0.77 [0.34, 1.72]	0.795	0.08 [0.02, 0.29]	**<0.001**	0.10 [0.03, 0.39]	**0.002**
Patient Appropriate	1.84 [0.99, 3.41]	0.128	2.70 [1.44, 5.07]	**0.005**	1.47 [0.80, 2.71]	0.435
Patient Safety	0.84 [0.37, 1.90]	0.906	0.20 [0.07, 0.58]	**0.008**	0.24 [0.08, 0.70]	**0.023**

OR: Odds Ratio, CI: Confidence interval, Chat: ChatGPT, OE: OpenEvidence.

Table 3: Average Readability Scales of ChatGPT vs Gemini vs OE vs ATA.

Metric	Chat	Gemini	OE	ATA	Pairwise Comparisons
FKGL	16.75 ± 9.42	15.44 ± 2.04	19.42 ± 2.77	14.51 ± 3.60	ATA vs Chat: p=0.2967 ATA vs Gemini: p=0.8848 ATA vs OE: **p=0.0011** Chat vs Gemini: p=0.7307 Chat vs OE: p=0.1601
					Gemini vs OE: **p=0.0117**
FRES	15.42 ± 2.69	21.85 ± 9.74	2.75 ± 10.77	26.29 ± 16.45	ATA vs Chat: **p<0.001** ATA vs Gemini: p=0.1956 ATA vs OE: **p<0.001** Chat vs Gemini: **p=0.0232** Chat vs OE: **p<0.001**
					Gemini vs OE: **p<0.001**
GFI	16.83 ± 1.74	17.67 ± 2.25	22.76 ± 2.71	17.93 ± 3.70	ATA vs Chat: p=0.1314 ATA vs Gemini: p=0.9541 ATA vs OE: p<.001 Chat vs Gemini: p=0.3428 Chat vs OE: **p<0.001**
					Gemini vs OE: **p<0.001**
Word Count	350 ± 176	432 ± 151	321 ± 159	715 ± 796	ATA vs Chat: **p=0.0010** ATA vs Gemini: **p=0.0171** ATA vs OE: **p<0.001** Chat vs Gemini: p=0.8148 Chat vs OE: p=0.9899
					Gemini vs OE: p=0.6357

Chat: ChatGPT, OE: OpenEvidence, ATA: American Thyroid Association, FKGL: Flesch-Kincaid Grade Level, FRES: Flesch Reading Ease Score, GFI: Gunning Fog Index

How to cite this article: Emily Kwon, Amina Shaikh, Alexandra Filipkowski, Annie Xu, Sudeepti Vedula, et al. Can AI Match Expert Consensus? A Comparison with 2025 ATA Guidelines in Differentiated Thyroid Cancer. Glob J Otolaryngol, 2026; 28(4): 556241. DOI:10.19080/GJO.2026.28.556241

Results

Mean Likert ratings differed significantly across LLMs for accuracy (p<0.001), comprehensiveness (p=0.001), and safety (p=0.025), whereas no significant differences were observed for patient-appropriate (p=0.052) (Table 1). Overall, OE demonstrated greater concordance with the ATA guidelines than both Chat and Gemini. For accuracy, comprehensiveness, and safety, OE responses had significantly higher odds of receiving higher ratings compared with Chat and Gemini (OR=5-50, all p<0.01), while no significant difference was observed between Chat and Gemini (p>0.05) (Table 2). In contrast, for patient-appropriateness, Chat received significantly higher ratings than OE (OR=2.7, p=0.006), while Gemini did not differ significantly from either platform (both p>0.05). Regarding readability, ATA guidelines demonstrated significantly higher FRES scores than ChatGPT and OE (both p<0.0001), while no significant difference was observed with Gemini (p>0.05). Gemini responses had higher FRES scores than both Chat and OE (p<0.05). For FKGL, OE demonstrated significantly higher scores than ATA (p=0.001) and Gemini, p=0.012), and no significant differences were observed among ATA, ChatGPT, and Gemini. Similarly, for GFI, OE had significantly higher scores than ATA, Chat, and Gemini (all p<0.0001), while no significant differences were observed among ATA, ChatGPT, and Gemini. Regarding word count, Chat, Gemini, and OE generated significantly longer responses than ATA (p≤ 0.017), although word counts did not differ significantly among the AI platforms. Mean FRES scores were well below 80, indicating difficult readability. FKGL and GFI values across all platforms exceeded a 12th grade reading level. All pairwise platform comparisons are shown in Table 3.

Discussion

Studies across medical disciplines are examining LLMs as means for medical education for providers and patients [10-19]. However, no consensus has been reached on the most suitable LLM tool, as there is variability in the accuracy, safety, and readability of LLM-generated responses. In this comparative evaluation of LLM-generated DTC management responses, we observed differences in guideline concordance, reference practices, and readability across ChatGPT, Gemini, and OE. To date, few studies have examined the application of AI tools in education related to thyroid cancer management. Helvaci et al. evaluated the accuracy and reliability of ChatGPT responses on thyroid cancer in reference to the 2014 guidelines and found that ChatGPT was generally accurate and reliable. Alarifi et al. [20] assessed thyroid nodule cancer risk using ChatGPT, Gemini, and Claude, with responses reviewed by radiologists [21]. They reported no significant differences in appropriateness or reliability among the platforms when compared with established guidelines, although ChatGPT demonstrated the highest accuracy followed by Claude and Gemini. Moise et al. evaluated ChatGPT recommendations on management of thyroid nodules as classified by the Bethesda System for Reporting Thyroid Cytopathology and similarly found recommendations were consistent with 2015 ATA guidelines [22]. The present study aligns with these prior findings that LLMs show high concordance with established ATA guidelines. Responses were primarily graded as agree or strongly agree across domains of accuracy, comprehensiveness, and patient safety, with ChatPT and Gemini performing similarly overall. Expanding beyond thyroid-specific literature, our findings remain consistent with prior otolaryngology studies showing strong concordance, accuracy, and comprehensiveness of ChatGPT and Gemini when evaluated against established otology or pediatric guidelines [13,23]. In contrast, studies assessing other topics such as sinusitis and olfaction have demonstrated mixed accuracy and quality [16,17], further demonstrating that LLM performance varies topic to topic (Supplemental Table 1).

Supplemental Table 1: Differentiated thyroid cancer queries input into Artificial Intelligence models.

Topic	Question
Preoperative Process and Risk Factors	1. Does surgical experience influence complication rates for thyroidectomy in differentiated thyroid cancer?
	2. What is the role of preoperative staging with diagnostic imaging and laboratory tests in differentiated thyroid cancer?
	3. When should preoperative cross-sectional or 18F-fluorodeoxyglucose-PET imaging be performed for patients with differentiated thyroid cancer?
	4. Should a serum Tg level be measured prior to surgery in patients with differentiated thyroid cancer?
	5. Should preoperative somatic genomic testing be performed to inform the extent of surgery for patients with differentiated thyroid cancer?
Active Surveillance	1. Are there patients with differentiated thyroid cancer in whom active surveillance and percutaneous ablation are appropriate management options?
	2. What is the optimal approach for patients with differentiated thyroid cancer undergoing active surveillance?
	3. Should serum Tg and TgAb levels be measured during active surveillance in patients with differentiated thyroid cancer?
	4. Are there clear indications for when patients with differentiated thyroid cancer undergoing active surveillance should pursue resection?

Surgical Management	1. What is the optimal operative approach for differentiated thyroid cancer?
	2. When should completion thyroidectomy be performed for patients with differentiated thyroid cancer?
	3. What is the surgical approach to thyroglossal duct carcinoma?
	4. When should completion thyroidectomy following Sistrunk procedure be performed?
Lymph Node Management	1. When should prophylactic central-compartment lymph node resection be performed in patients with differentiated thyroid cancer?
	2. What is the best approach for therapeutic central and lateral compartment node resections in patients with differentiated thyroid cancer?
Perioperative and Intraoperative Management	1. What is the appropriate perioperative approach to voice and parathyroid issues in patients with differentiated thyroid cancer?
	2. Should the patient with differentiated thyroid cancer undergo voice or laryngeal examination prior to surgery?
	3. How should the recurrent laryngeal nerves be assessed intraoperatively in patients with differentiated thyroid cancer?
	4. How should the parathyroid glands be managed intraoperatively and perioperatively in patients with differentiated thyroid cancer?
	5. Should drainage of the thyroidectomy bed be performed in patients with differentiated thyroid cancer?
Postoperative Surveillance and Risk Factors	1. How should the surgeon manage postoperative voice changes and symptoms after surgery if they occur in patients with differentiated thyroid cancer?
	2. What are the basic principles of histopathologic evaluation of thyroidectomy samples in patients with differentiated thyroid cancer?
	3. How should risk of recurrence and initial assessment be performed after surgery in patients with differentiated thyroid cancer?
	4. How should clinical response to surgery be assessed in patients with differentiated thyroid cancer?
	5. When should Tg levels be measured after surgery in patients with differentiated thyroid cancer?
	6. What is the role of ultrasound and other imaging techniques CT, MRI, 18FDG-PET-CT) after primary resection in patients with differentiated thyroid cancer?
RAI Therapy Management and Procedures	1. What is the role of RAI therapy after thyroidectomy in the primary management of differentiated thyroid cancer?
	2. Should radioiodine be administered for oncocytic thyroid carcinoma treatment?
	3. How should patients be prepared for RAI therapy administration in patients with differentiated thyroid cancer?
	4. Should a low-iodine diet be prescribed prior RAI therapy administration in patients with differentiated thyroid cancer?
	5. When and how should diagnostic radioiodine whole-body scan be performed in patients with differentiated thyroid cancer?
	6. Should post-therapy whole-body scan be performed in patients with differentiated thyroid cancer?
	7. Should single photon emission computed tomography with computed tomography be performed with the whole-body scan in patients with differentiated thyroid cancer?
	8. What is the role of radiotherapy, with or without chemotherapy, in patients with differentiated thyroid cancer?
RAI Risk Management	1. How should patients with differentiated thyroid cancer be educated regarding radiation safety?
	2. How do you counsel and minimize risks of RAI therapy side effects to the salivary glands and lacrimal ducts in patients with differentiated thyroid cancer?
	3. How should patients with differentiated thyroid cancer be counseled regarding the risk of second primary malignancy after receiving RAI therapy?
	4. What other testing should patients with differentiated thyroid cancer receiving RAI therapy undergo?
	5. How should patients with differentiated thyroid cancer be counseled about RAI therapy and pregnancy, nursing, and gonadal function?

Questions were extracted from the 2025 American Thyroid Association guidelines.

Radioactive Iodine (RAI), Thyroglobulin (Tg), Antibody (Ab)

This is one of the first studies to evaluate OE's performance to other LLMs in applications for head and neck manifestations [24]. Previous studies have examined OE's application in plastic surgery, primary care, and dermatology [24–28]. Rivera et al. suggested that OE generates higher quality and more reliable material than ChatGPT. Hack et al. [25] reported that OE outperformed Gemini but was comparable to ChatGPT-5 in accuracy, comprehensiveness, usefulness, clarity, and relevance in simulated emergency department and primary care otolaryngology cases. Nihal et al. [27] reported that OE generated superior responses than ChatGPT in accuracy when evaluated against the NCCN guidelines for basal cell carcinoma and squamous cell carcinoma.28 In contrast, Kring et al. examined patient queries related to head and neck cancer symptoms and found that Microsoft Copilot produced higher quality health information compared to OE [24]. Our results align with those of Rivera et al., Hack et al., and Nihal et al., demonstrating that OE had a greater likelihood of being rated higher than ChatGPT and Gemini across accuracy, safety, and comprehensiveness [25,27,28]. Although OE received lower scores in patient appropriateness compared with ChatGPT, this finding is consistent with Kring et al. who demonstrated that OE scored lower in expressing clear aims, relevance, and encouraging shared decision making [24]. This may be attributed to the fact that OE is designed to provide clinical advice for health professionals rather than the general public, and sources information primarily from peer-reviewed journals, such as those from the NEJM Group, JAMA Network, and NCCN guidelines. Furthermore, the questions may better reflect physician rather than patient concerns. Overall, OE may represent a reliable adjunct for clinicians in DTC management.

There is established evidence that medical information generated by LLMs often exceeds recommended health literacy standards [10,11,29]. Prior studies by Washington et al. and Haver et al. found response generated by ChatGPT, Google Bard, or Gemini surpassed FLEK and FRES recommendations when addressing topics related to lung cancer screening [11,29]. Similarly, Ho et al. evaluated ChatGPT and GAO in dysphonia education and found that both FLEK and FRES scores exceeded the AMA-recommended 6th grade readability level [10]. In our analysis, both the ATA guidelines and LLM-generated responses similarly exceeded the recommended 6th grade reading level, corroborating the findings from previous studies [10,11,29]. While LLMs can be prompted to adjust the length and readability of their responses, this occurs often at the risk of generalization or omitting nuance, particularly in areas of medical uncertainty or controversy [30]. Although the ATA guidelines are intended for use by clinicians and patients, their content regarding detailed diagnostic testing and management is not primarily designed for patient-initiated inquiries [1]. DTC is a specialized topic, and many of the guideline-based questions may not reflect the language or depth typically posed by patients. Future studies should therefore evaluate patient-specific educational materials, such as patient brochures, when evaluating LLM for this purpose. This study presents updated insights into the performance of LLMs for medical education in DTC management; however, it is not without limitations. As observed in prior studies, minor changes in wording can significantly alter LLM responses [31]. Furthermore, the language used in our queries, developed by clinicians, may not reflect typical patient questions and thus limits generalizability. Third, the AI models evaluated here are rapidly evolving; future versions may display different performance characteristics [32,33].

Conclusion

There is strong alignment between AI-generated responses (ChatGPT, Gemini, OE) and ATA guidelines for DTC management. OE consistently provided outputs most concordant with the ATA guidelines throughout accuracy, comprehensiveness, and patient safety for clinicians. AI platforms, particularly those designed for clinicians, offer promising adjunctive tools to assist clinicians in the management of adult patients with DTC.

References

1. Ringel MD, Sosa JA, Baloch Z (2025) American Thyroid Association Management Guidelines for Adult Patients with Differentiated Thyroid Cancer. Thyroid Off J Am Thyroid Assoc 35(8): 841-985.
2. Patel VR, Liu M, Jena AB (2025) Public Interest in an AI-Enabled Clinical Decision Support Tool. JAMA Netw Open 8(11): e2544672.
3. Kassem MA, Hosny KM, Damaševičius R, Eltoukhy MM (2021) Machine Learning and Deep Learning Methods for Skin Lesion Classification and Diagnosis: A Systematic Review. Diagnostics 11(8): 1390.
4. Tian C, Gao Y, Rui C, Qin S, Shi L (2024) Artificial intelligence in orthopedic trauma. Eng Medicine 1(2): 100020.
5. Ramgopal S, Sanchez-Pinto LN, Horvat CM, Carroll MS, Luo Y (2023) Artificial intelligence-based clinical decision support in pediatrics. Pediatr Res 93(2): 334-341.
6. Cheo HM, Ong CYG, Ting Y (2025) A Systematic Review of AI Performance in Lung Cancer Detection on CT Thorax. Healthcare 13(13): 1510.
7. (2026) Readability Test - WebFX.
8. Raja H, Lodhi S (2024) Assessing the readability and quality of online information on anosmia. Ann R Coll Surg Engl 106(2): 178-184.
9. Gunning R (1969) The Fog Index After Twenty Years. J Bus Commun 6(2): 3-13.
10. Ho RA, Shah E, De Armas JS, Yan K, Kaye R (2025) Artificial Intelligence Models for Dysphonia Patient Education. Otolaryngol--Head Neck Surg Off J Am Acad Otolaryngol-Head Neck Surg 173(6): 1455-1462.
11. Haver HL, Lin CT, Sirajuddin A, Yi PH, Jeudy J (2023) Use of ChatGPT, GPT-4, and Bard to Improve Readability of ChatGPT's Answers to Common Questions About Lung Cancer and Lung Cancer Screening. AJR Am J Roentgenol 221(5): 701-704.
12. Aliyeva A, Sari E, Alaskarov E (2024) Enhancing Postoperative Cochlear Implant Care With ChatGPT-4: A Study on Artificial Intelligence (AI)-Assisted Patient Education and Support. Cureus 16.
13. Rossi NA, Corona KK, Yoshiyasu Y, Hajiyev Y, Hughes CA (2025) Comparative analysis of GPT-4 and Google Gemini's consistency with

How to cite this article: Emily Kwon, Amina Shaikh, Alexandra Filipkowski, Annie Xu, Sudeepti Vedula, et al. Can AI Match Expert Consensus? A Comparison with 2025 ATA Guidelines in Differentiated Thyroid Cancer. Glob J Otolaryngol, 2026; 28(4): 556241. DOI:10.19080/GJO.2026.28.556241

pediatric otolaryngology guidelines. Int J Pediatr Otorhinolaryngol 193: 112336.

14. Johnson D, Goodman R, Patrinely J (2023) Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model.
15. Li J, Gao X, Dou T, Gao Y, Li X (2024) Quantitative evaluation of GPT-4's performance on US and Chinese osteoarthritis treatment guideline interpretation and orthopedic case consultation. BMJ Open 14(12): e082344.
16. Shaari AL, Saad AM, Patil D (2025) Evaluation of Large Language Models' Concordance With Guidelines on Olfaction. Laryngoscope Investig Otolaryngol 10(2): e70130.
17. Yoshiyasu Y, Wu F, Dhanda AK, Gorelik D, Takashima M (2023) GPT-4 accuracy and completeness against International Consensus Statement on Allergy and Rhinology: Rhinosinusitis. Int Forum Allergy Rhinol 13(12): 2231-2234.
18. Albehairi SA, Alsahli OM, Bander LF, Albilasi TM, Aljardan ES (2025) Postoperative Otoplasty Care With ChatGPT-4: A Study on Artificial Intelligence (AI)-Assisted Patient Concern and Education. J Craniofac Surg 36(1): 296-298.
19. Goodman RS, Patrinely JR, Stone CA Jr (2023) Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Netw Open 6(10): e2336483.
20. Cavnar Helvaci B, Hepsen S, Candemir B (2024) Assessing the accuracy and reliability of ChatGPT's medical responses about thyroid cancer. Int J Med Inf 191: 105593.
21. Alarifi M (2025) Appropriateness of Thyroid Nodule Cancer Risk Assessment and Management Recommendations Provided by Large Language Models. J Imaging Inform Med 38(6): 4324-4335.
22. Moise A, Tatar L, Sela N (2025) Thyroid Nodule Experts Evaluating ChatGPT's Assessment of Thyroid Nodules Classified by the Bethesda System for Reporting Thyroid Cytopathology. J Otolaryngol - Head Neck Surg 54: 19160216251387617.
23. Rossi NA, Corona KK, Yoshiyasu Y, Young DL, McKinnon BJ (2024) Evaluating the Accuracy and Completeness of Artificial Intelligence Responses Against Established Otology Guidelines. Otol Neurotol Open 4(3): e059.
24. Kring T, Prasad S, Dadi S, Sokhn E, Franzmann E (2025) A comparison of quality and readability of Artificial Intelligence chatbots in triage for head and neck cancer. Am J Otolaryngol 46(5): 104710.
25. Perez Rivera LR, Gursky AK, Elmer N, Boyd CJ, Karp NS (2026) Evaluating the Quality and Reliability of Large Language Models for Plastic Surgery Patient Education: A Comparative Analysis of ChatGPT and OpenEvidence. Aesthet Surg J 46(2): 160-167.
26. Hurt RT, Stephenson CR, Gilman EA (2025) The Use of an Artificial Intelligence Platform OpenEvidence to Augment Clinical Decision-Making for Primary Care Physicians. J Prim Care Community Health 16: 21501319251332215.
27. Hack S, Zalzal HG, Attal R (2026) Empowering front-line physicians with AI: Evaluating large language models in everyday ENT care. Am J Emerg Med 102: 90-97.
28. Nihal A, Yang K, Pavlidakey PG (2026) OpenEvidence Is Superior to GPT in Skin Cancer Guideline Queries. Int J Dermatol 65(2): 338-339.
29. Washington C, Divaker J, Gould M (2025) Evaluating the Effectiveness of ChatGPT and Google Gemini in Providing Lung Cancer Screening Recommendations for Vulnerable Communities. CHEST Pulm 3(2): 100167.
30. Kufta AY, Djalilian AR (2025) Enhancing Patient Education with AI: A Readability Analysis of AI-Generated Versus American Academy of Ophthalmology Online Patient Education Materials. J Clin Med 14(19): 6968.
31. Wang L, Chen X, Deng X (2024) Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med 7: 41.
32. Jaleel A, Aziz U, Farid G (2025) Evaluating the Potential and Accuracy of ChatGPT-3.5 and 4.0 in Medical Licensing and In-Training Examinations: Systematic Review and Meta-Analysis. JMIR Med Educ 11: e68070.
33. Hill GS, Fischer JL, Watson NL, Riley CA, Tolisano AM (2024) Assessing the quality of artificial intelligence–generated patient counseling for rhinosinusitis. Int Forum Allergy Rhinol 14(10): 1634-1637.

DOI: 10.19080/GJO.2026.28.556241

How to cite this article: Emily Kwon, Amina Shaikh, Alexandra Filipkowski, Annie Xu, Sudeepti Vedula, et al. Can AI Match Expert Consensus? A Comparison with 2025 ATA Guidelines in Differentiated Thyroid Cancer. Glob J Otolaryngol, 2026; 28(4): 556241. DOI:10.19080/GJO.2026.28.556241