Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.264082

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Diagnostic performance and short-interval reproducibility of multimodal large language models in differentiating cholesteatoma from chronic otitis media using key-image temporal bone high-resolution computed tomography

Buket Yağcı[1]
Sergen Palaz[1]
Erdem Atalay Çetinkaya[2]
Utku Şenol[3]

[1]University of Health Sciences Türkiye, Antalya Training and Research Hospital, Clinic of Radiology, Antalya, Türkiye

[2]University of Health Sciences Türkiye, Antalya Training and Research Hospital, Clinic of Otorhinolaryngology, Antalya, Türkiye

[3]Akdeniz University Faculty of Medicine, Department of Radiology, Antalya, Türkiye

PURPOSE

This study aimed to assess whether multimodal large language models (LLMs) can distinguish cholesteatoma from non-cholesteatomatous chronic otitis media (COM) on representative key-image temporal bone high-resolution computed tomography (HRCT) and to evaluate the short-interval reproducibility of their outputs.

METHODS

This retrospective, single-center study (2019–2024) included 101 patients (48 with cholesteatoma, 53 with non-cholesteatomatous COM) who underwent surgical treatment. The reference standard was intraoperative diagnosis with histopathological confirmation for cholesteatoma and surgical documentation for COM. For each case, six anonymized representative HRCT images reflecting standard diagnostic criteria were selected by consensus between a 4th-year radiology resident and a board-certified head and neck radiologist, both unaware of the diagnosis. Subsequently, the same images were analyzed by the GPT-5 and Gemini 2.5 Pro LLMs through their official web interfaces, utilizing structured prompts and a zero-shot approach. These evaluations were conducted in two distinct sessions (S1 and S2) with a 1-week interval. The primary endpoint was accurate binary classification. Accuracy, sensitivity, specificity, positive predictive value, and negative predictive value were calculated with 95% confidence intervals [(CIs); Wilson method] agreement with the reference standard and between sessions and models was assessed with the Cohen kappa (κ) coefficient; and differences in classification were assessed with the McNemar test.

RESULTS

The radiologist achieved an accuracy of 96.0% (95% CI: 90.3–98.4) with almost perfect agreement with the reference standard (κ: 0.921). In S1, GPT-5 and Gemini 2.5 Pro achieved accuracies of 43.6% and 49.5%, and in S2, 46.5% and 47.5%, respectively. Both models combined high sensitivity (83.3%–97.9%) with low specificity (1.9%–13.2%), and balanced accuracy ranged from 0.45 to 0.52. Between-session reproducibility was fair for GPT-5 (κ: 0.360) and moderate for Gemini 2.5 Pro (κ: 0.485), and inter-model agreement was slight at both sessions (κ: 0.035 at S1 and κ: 0.086 at S2). Accuracy did not differ significantly between the two models ($P = 0.211$).

CONCLUSION

In this single-center study, GPT-5 and Gemini 2.5 Pro, in the versions evaluated, combined high sensitivity with low specificity and showed only fair-to-moderate between-session reproducibility and exhibited slight inter-model agreement on temporal bone key-image HRCT. These findings do not support their use as independent second readers, and broader generalization to other multimodal LLMs would require the evaluation of additional models.

CLINICAL SIGNIFICANCE

The evaluated models lacked the spatial precision and consistency required for the accurate assessment of complex middle ear structures. This finding underscores the necessity for verification by a radiologist and continuous monitoring.

KEYWORDS

Large language models, artificial intelligence, computed tomography, cholesteatoma, chronic otitis media

Handling editor: Tuğba Akıncı D'Antonoli

Corresponding author: Buket Yağcı

E-mail: buketyagci@hotmail.com

Received 04 May 2026; revision requested 11 June 2026; last revision requested 12 July 2026; accepted 16 July 2026.

Epub: 11.08.2026

DOI: 10.4274/dir.2026.264082

You may cite this article as: Yağcı B, Palaz S, Çetinkaya EA, Şenol AU. Diagnostic performance and short-interval reproducibility of multimodal large language models in differentiating cholesteatoma from chronic otitis media using key-image temporal bone high-resolution computed tomography. *Diagn Interv Radiol*. XXX August 2026 DOI: 10.4274/dir.2026.264082 [Epub Ahead of Print].

Chronic otitis media (COM) is a common inflammatory disease of the middle ear and mastoid cavity.[1] Distinguishing non-cholesteatomatous COM from cholesteatoma is clinically important because cholesteatoma causes progressive osseous erosion and a higher risk of labyrinthine and intracranial complications, frequently requiring a different surgical approach.[2] Temporal bone high-resolution computed tomography (HRCT) is the standard preoperative modality for assessing disease extent and ossicular and tegmental integrity, particularly when otoscopy alone is insufficient for a confident differential diagnosis. However, the discriminative features, chiefly the pattern of bony erosion and the distribution of middle-ear soft tissue, frequently overlap; therefore, accurate interpretation depends heavily on reader expertise.[3]

Large language models (LLMs) are artificial intelligence (AI) systems trained on large-scale data to process and generate natural language, and they have been shown to encode clinical knowledge.[4] Their most consistent medical performance has been on text-based tasks, such as report summarization and clinical decision support.[5,6] More recent multimodal LLMs can additionally process images, although their ability to interpret radiological images remains comparatively limited and variable and is still being characterized;[6,7] these systems also exhibit characteristic limitations, including stochastic output variability, behavioral drift, and hallucination.[8]

Main points

- On temporal bone key-image high-resolution computed tomography (HRCT), GPT-5 and Gemini 2.5 Pro combined high sensitivity with very low specificity, reflecting a systematic bias toward over-diagnosing cholesteatoma rather than genuine discrimination; their balanced accuracy was near chance, whereas an expert radiologist reading the same images reached 96.0% accuracy.
- Model outputs were only fair-to-moderately reproducible between sessions and showed merely slight inter-model agreement, so the choice of model and the timing of access are themselves non-trivial sources of diagnostic variability.
- Because these findings are specific to the model versions tested, current general-purpose multimodal large language models should not be used as independent second readers for temporal bone key-image HRCT until task-specific calibration, version stability, and external validation are demonstrated.

Within radiology, multimodal LLMs are increasingly evaluated for image-related tasks, yet systematic appraisals emphasize recurrent pitfalls, including reduced specificity on image-based interpretation,[6,9] and evidence in head and neck imaging remains limited. In one of the few otologic studies, GPT-4V (with vision) classified middle-ear disease on otoscopic images with an accuracy intermediate between that of non-specialists and otolaryngologists.[10] By contrast, on cross-sectional head and neck CT, multimodal LLMs have demonstrated significant variability in diagnostic performance depending on the model version and provider; for instance, in a recent study evaluating CT-based polyp detection, GPT-4o achieved a remarkably high diagnostic agreement (accuracy: 0.89, kappa: 0.77), significantly outperforming both GPT-5 and Gemini 2.5 Pro ($P < 0.001$ for each) under identical inputs.[11] To our knowledge, however, the diagnostic performance of current multimodal LLMs on temporal bone HRCT and the short-interval reproducibility of their outputs have not been established.

Although LLMs are anticipated to serve near-term clinical functions mostly in text-based tasks, before their routine usage in image-centric diagnostic support, such as identifying suspected cholesteatoma in environments lacking subspecialty expertise, evidence on their diagnostic accuracy and algorithmic consistency must be established. At present, these fundamental performance metrics remain uncharacterized within the specific domain of temporal bone HRCT. Accordingly, this study aims to evaluate and compare the diagnostic performance of GPT-5 and Gemini 2.5 Pro in differentiating cholesteatoma from non-cholesteatomatous COM using representative key-image temporal bone HRCT and to assess the short-interval reproducibility of their outputs across two reading sessions, with expert radiologist interpretation and intraoperative findings as the reference standard.

Methods

Study design and ethics

This retrospective study was approved by University of Health Sciences Türkiye, Antalya Training and Research Hospital's Institutional Review Board (IRB) of the participating institution (approval number: 1/19, date: January 9, 2025). The study was conducted in accordance with the principles of the Declaration of Helsinki, the European Union Artificial Intelligence Act [Regulation (EU) 2024/1689], and the Minimum Reporting Items for Clear Evaluation of Accuracy Reports of LLMs in Healthcare framework.[12,13] Reporting of the study adheres to the Reporting Checklist for Foundation and Large Language Models in Medical Research guidelines (Supplementary Material 1).[14] Given the retrospective design and exclusive use of de-identified data, the IRB waived informed patient consent. This study constitutes a retrospective performance evaluation and involved no clinical deployment of AI systems.

Patients and reference standard

A total of 147 consecutive patients who underwent surgery due to suspected cholesteatoma (2019–2024) were screened using a Picture Archiving and Communication System. The reference standard for cholesteatoma was intraoperative diagnosis with histopathological confirmation. The comparison group consisted of non-cholesteatomatous COM, defined as surgically treated chronic middle ear disease documented in the surgical report as not having cholesteatoma (supported by clinical evaluation consistent with COM). Surgical confirmation was the reference standard for confirming the diagnosis and distinguishing cholesteatoma cases from COM and performing subclassification.

Patients with alternative diagnoses, incomplete clinical data, prior ear surgery, or suboptimal CT scan quality were excluded. The final cohort comprised 48 cholesteatoma and 53 non-cholesteatomatous COM cases; only the affected ears were evaluated. The detailed patient selection flowchart is illustrated in Figure 1.

Computed tomography scan, radiological feature checklist, and selection of representative images

Imaging was performed at Antalya Training and Research Hospital on two scanners: a 64-detector, 128-slice CT 5300 (Philips Healthcare, Suzhou, China) and a 32-detector, 64-slice Somatom Go Up (Siemens Healthcare, Erlangen, Germany). A standardized HRCT protocol with identical parameters was applied to both systems: 0.67 mm slice thickness, 120 kVp, and 250 mAs.

For each case, six representative images (three axial and three coronal) reflecting standard diagnostic criteria were selected by consensus between a board-certified head and neck radiologist and a 4th-year radiology resident, both blinded to the surgical and histopathological diagnosis. The same head and neck radiologist subsequently performed the blinded diagnostic interpreta-

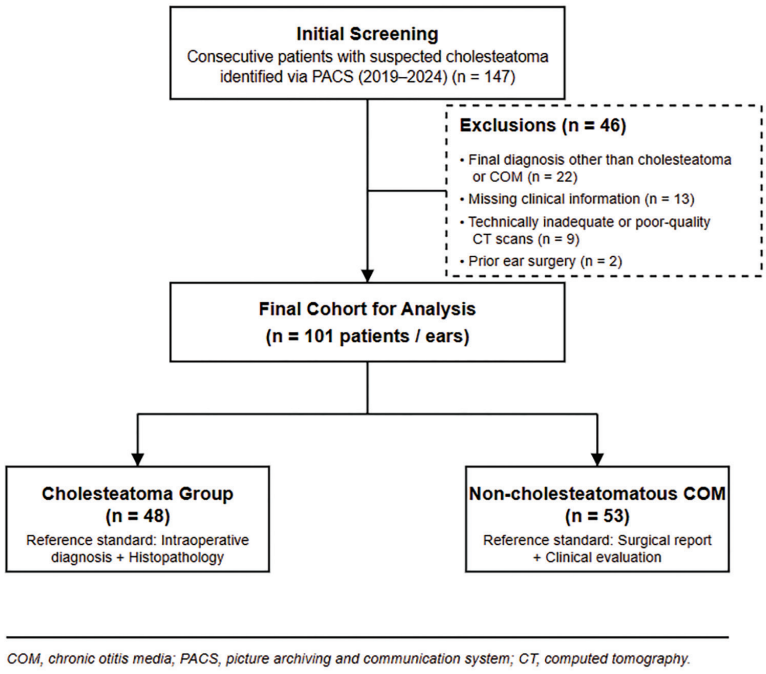

Figure 1. Patient selection flowchart.

tion of these key images, using a predefined checklist to record the distribution of soft-tissue opacity, erosion of bone and the ossicles (malleus, incus, and stapes), tegmen tympani dehiscence, tympanic membrane perforation, and other findings, according to established temporal bone HRCT criteria.[15,16]

The key-image approach was adopted to standardize inputs for the web-based LLM interfaces; because current multimodal LLMs cannot directly process full-volume Digital Imaging and Communications in Medicine (DICOM) datasets, complete interpretation series could not be used. Following patient selection and before model evaluation, all imaging data were fully anonymized by removing DICOM headers, series descriptions, and all free-text fields containing personally identifiable information. The representative slices were exported in Tagged Image File Format (TIFF; bone window; approximately 1.7 MB per image; 1,777 × 1,139 pixels) and then converted to Portable Network Graphics (PNG) for upload, with no resizing, cropping, compression, or windowing change. Three axial and three coronal images from a sample cholesteatoma case are shown in Supplementary Material 2.

Large language model evaluation and prompts

Anonymized images were evaluated using two contemporary multimodal LLMs: OpenAI's GPT-5, publicly released in August 2025,[17] and Google DeepMind's Gemini 2.5 Pro, announced in March 2025.[18] Access to GPT-5 was through the official ChatGPT web interface (chatgpt.com) and Gemini 2.5 Pro through the official Gemini web application (gemini.google.com)[19,20] on August 17, 2025, for the first evaluation and on August 24, 2025, for the second. Each model was evaluated with a zero-shot approach using the same structured prompt for every case. To minimize transfer effects, a new chat session was initiated for each case, and the conversation history and memory/personalization features were disabled wherever the interface allowed (for GPT-5, by disabling memory in the account settings and using a new temporary chat; for Gemini 2.5 Pro, by turning off Gemini Apps Activity and personalization). The two sessions were conducted from separate user accounts.

All cases were evaluated by both models at the first session (S1), with the order of cases and of model evaluation randomized by a computer-generated list to minimize ordering effects. Each model first received the following prompt: "In this CT image, describe the normal and abnormal findings in the temporal bone, indicating which side they are on, as a radiologist would. Based on any abnormal findings you see, list the three most likely differential diagnoses." A second prompt was then offered, as follows: "Based on the findings in these images, which of the following differential diagnoses is the most likely diagnosis for this patient: cholesteatoma or chronic otitis media? If the diagnosis is cholesteatoma, classify it as acquired cholesteatoma (flaccida type and/or tensa type) or congenital cholesteatoma." The evaluation process is illustrated in Figures 2 and 3, and example image descriptions with model outputs from both sessions are provided in Supplementary Material 3.

Temporal validation (two sessions)

The entire evaluation was repeated at S2, 1 week after S1, using the same images, prompts, and randomization, to assess short-interval reproducibility under real-world conditions in which commercial models may exhibit stochastic variation or background updates.

Outcome evaluation

The first diagnosis listed in each model's response was taken as its final output. The second prompt required a binary choice between cholesteatoma and COM, with cholesteatoma subtyping, to ensure response consistency. The primary endpoint was correct binary classification (cholesteatoma vs. COM) against the reference standard; secondary endpoints were intermodel consistency at each session and intramodel consistency across sessions. Baseline imaging-feature completeness was documented but did not affect the primary classification.

Statistical analysis

All statistical analyses were performed using IBM SPSS Statistics for Windows, version 31.0 (IBM Corp., Armonk, NY, USA), and Python 3.11 (scikit-learn and scipy libraries). Continuous variables are reported as mean ± standard deviation and categorical variables as frequencies and percentages. All tests were two-tailed, exact *P* values are reported, and a *P* value of < 0.05 was considered statistically significant.

For all diagnostic-performance analyses, cholesteatoma was designated the positive class (coded 1) and non-cholesteatomatous COM the negative class (coded 0); accordingly, sensitivity denotes the correct classification of cholesteatoma and specificity the correct classification of non-cholesteatomatous COM. The surgical–pathological diagnosis served as the reference standard.

Between-group differences in continuous variables were assessed using the independent-samples t-test. Categorical variables were compared using the Pearson chi-square test or the Fisher exact test when any expected cell frequency was < 5.

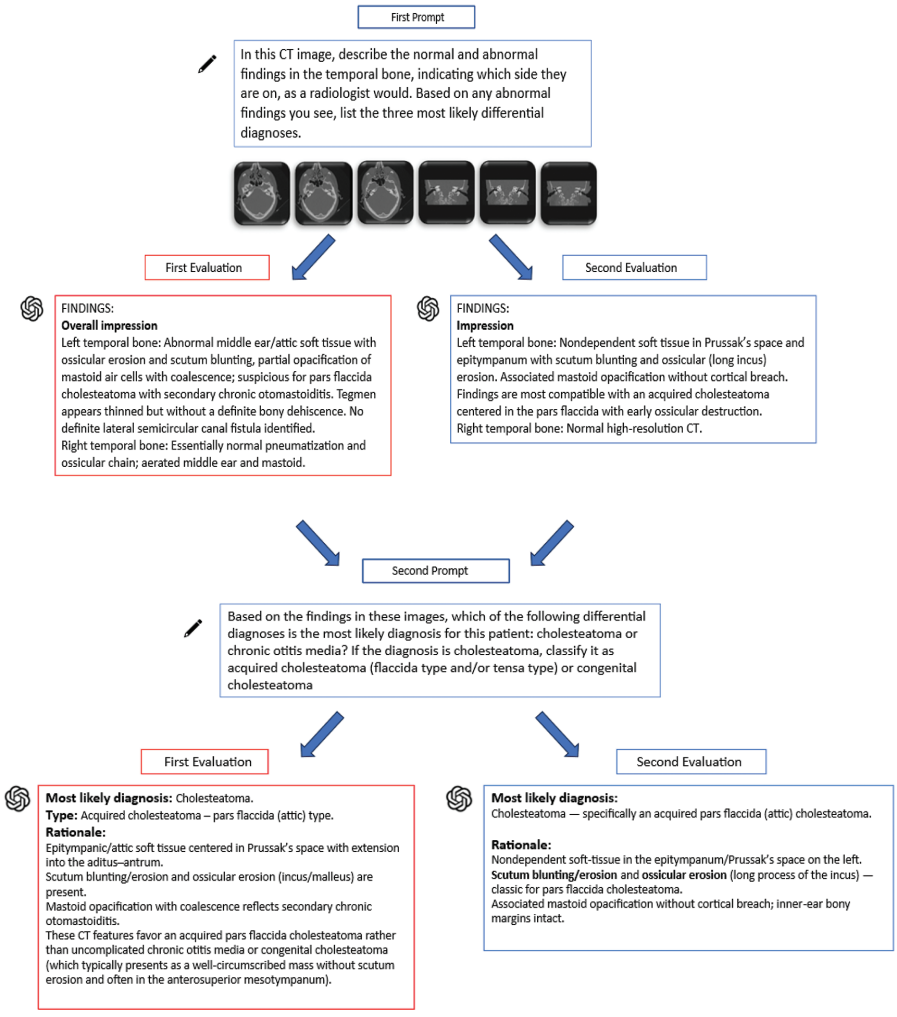

Figure 2. GPT-5 evaluation workflow. It shows the prompts given to GPT-5 and the model's responses after loading six high-resolution computed tomography images of a cholesteatoma case, evaluated at two sessions. CT, computed tomography.

For each evaluator (the radiologist, GPT-5, and Gemini 2.5 Pro at each session), accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated from the 2 × 2 contingency table, with 95% confidence intervals (CIs) obtained using the Wilson score method. Agreement with the reference standard was assessed using the Cohen κ coefficient with its 95% CI and a z-test for significance; κ was interpreted according to Landis and Koch[21] as poor (< 0.00), slight (0.00–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), or almost perfect (0.81–1.00). As both models produced binary outputs, discrimination at the single operating point was summarized as balanced accuracy, defined as (sensitivity + specificity)/2, with 95% CIs from the Wald approximation for the mean of two independent binomial proportions.

Pairwise differences in diagnostic accuracy were assessed using the McNemar test with continuity correction for radiologist versus GPT-5 (S1), radiologist versus Gemini 2.5 Pro (S1), and GPT-5 (S1) versus Gemini 2.5 Pro (S1). Within-model between-session reproducibility and inter-model agreement were assessed using the Cohen κ coefficient and the McNemar test. The Bonferroni correction was applied within two predefined comparison families: primary diagnostic-accuracy comparisons (κ: 3; corrected α: 0.0167) and reproducibility and inter-model comparisons (κ: 4; corrected α: 0.0125); both unadjusted *P* values and corrected thresholds are reported.

For each of the six predefined CT features (scutum erosion, ossicular erosion, tegmen dehiscence, epitympanic involvement, mastoid involvement, and tympanic-membrane perforation), the S1 model outputs were compared with the radiologist-defined checklist as the reference standard, and accuracy, sensitivity, and specificity were calculated per feature and per model.

Results

A total of 101 patients (48 with cholesteatoma, 53 with non-cholesteatomatous COM) were analyzed, with no significant differences in age or sex between groups. Scutum erosion, ossicular erosion, tegmen dehiscence, and tympanic membrane perforation were significantly associated with cholesteatoma, whereas epitympanic and mastoid involvement were not. Baseline characteristics and CT findings are detailed in Table 1.

The radiologist's diagnostic accuracy was significantly superior to that of both models across all metrics (Table 2). Against the pathological reference standard, the radiologist achieved 96.0% accuracy (sensitivity: 97.9%, specificity: 94.3%, PPV: 94.0%, NPV: 98.0%) with almost perfect agreement (κ: 0.921, 95% CI: 0.845–0.997; $P < 0.001$; balanced accuracy: 0.961, 95% CI: 0.924–0.998), correctly identifying 47 of 48 cholesteatoma cases (1 false negative) and 50 of 53 COM cases (3 false positives).

At S1, GPT-5 reached 43.6% accuracy (sensitivity: 83.3%, specificity: 7.5%, PPV: 44.9%, NPV: 33.3%; κ: −0.088, 95% CI: −0.274 to 0.099; $P = 0.357$; balanced accuracy: 0.454, 95% CI: 0.391–0.518) and Gemini 2.5 Pro 49.5% (sensitivity: 95.8%, specificity: 7.5%, PPV: 48.4%, NPV: 66.7%; κ: 0.032, 95% CI: −0.155 to 0.219; $P = 0.735$; balanced accuracy = 0.517, 95% CI: 0.471–0.562). Despite their high sensitivity, both models showed very low specificity: GPT-5 labeled 89 of 101 cases and Gemini 2.5 Pro 95 of 101 cases as cholesteatoma (49 false positives each). Both performed significantly worse than the radiologist ($P < 0.001$ for both), with no significant difference between the two models ($P = 0.211$).

Performance at S2 was similarly poor: GPT-5 achieved 46.5% accuracy (sensitivity: 83.3%, specificity: 13.2%; κ: −0.033; $P = 0.728$; balanced accuracy: 0.483, 95% CI: 0.413–0.552), and Gemini 2.5 Pro showed the most extreme positive bias of all evaluations, labeling 99 of 101 cases as cholesteatoma (accuracy: 47.5%, sensitivity: 97.9%, specificity: 1.9%; κ: −0.002; $P = 0.984$; balanced accuracy. 0.499, 95% CI: 0.472–0.526).

Feature-level agreement between the model outputs and the radiologist-defined reference checklist is presented in Table 3.

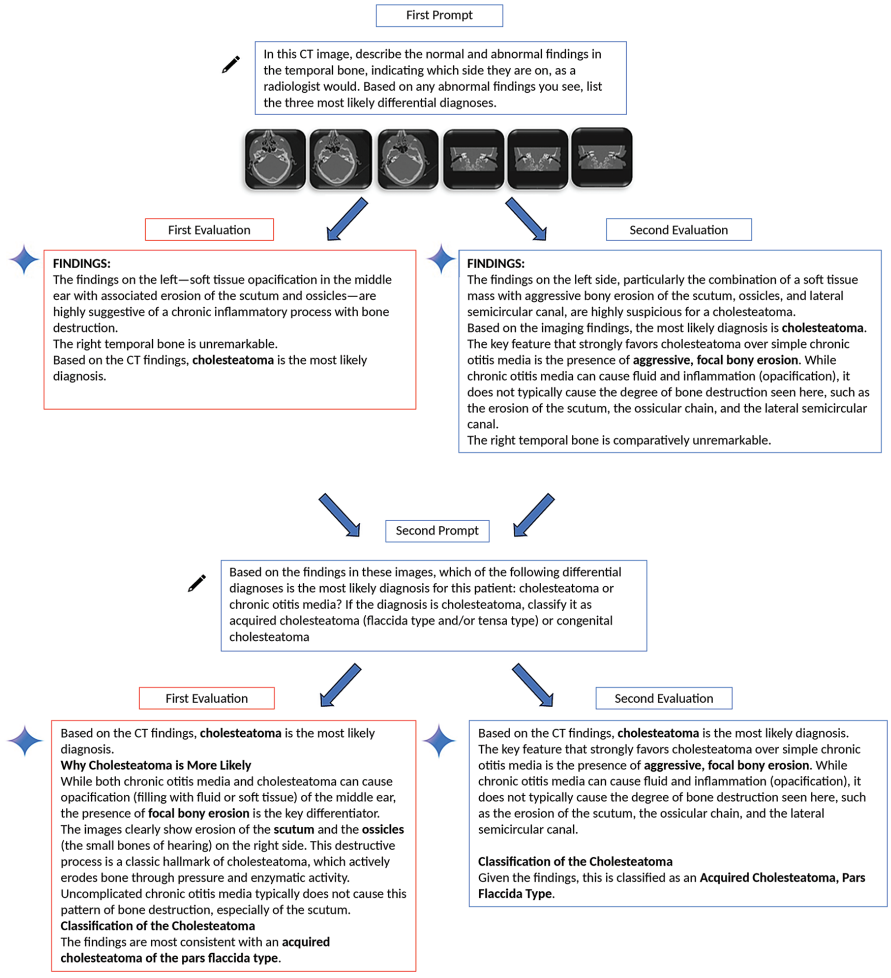

Figure 3. Gemini 2.5 Pro evaluation workflow. It shows the high-resolution computed tomography images of the same case loaded into Gemini 2.5 Pro, the prompts given, and the model's responses, evaluated at two sessions. CT, computed tomography.

Both models showed a consistent pattern of high sensitivity but very low specificity for most features, with the number of false positives exceeding that of true negatives for most features. For scutum erosion (72 of 101 cases), GPT-5 achieved 75.0% sensitivity but only 17.2% specificity (accuracy: 58.4%), and Gemini 2.5 Pro performed similarly (83.3%, 13.8%, 63.4%). For ossicular erosion (51 cases), both again combined high sensitivity (GPT-5: 80.4% and Gemini: 90.2%) with negligible specificity (10.0% and 12.0%, respectively). Tegmen dehiscence was slightly better balanced (GPT-5: 68.6%/28.0%; Gemini: 78.4%/38.0%). Epitympanic involvement, present in 100 of 101 cases, yielded high sensitivity (GPT-5: 88.0%; Gemini: 93.0%), but neither model detected the single negative case (specificity: 0.0%). Mastoid involvement showed the highest agreement, with sensitivities of 98.9% (GPT-5) and 96.6% (Gemini) and a specificity of 50.0% for Gemini (7 of 14 negative cases). In contrast, tympanic membrane perforation (94 of 101 cases) was frequently missed despite its high prevalence (sensitivity: 40.4% for GPT-5 and 54.3% for Gemini).

Facial nerve canal defects (5 cholesteatoma cases) and lateral semicircular canal fistulas (3 cases) were excluded from the formal feature-level analysis for two reasons: their low prevalence precluded statistically robust comparison, and their reliable assessment requires multiplanar review of the full HR stack, which the six representative slices could not reproduce. Consistent with this, neither model identified any of these findings.

With respect to cholesteatoma subtype, the cohort comprised 34 combined-type, 12 pars flaccida, and 2 congenital cholesteatomas. When subtype classification was requested in the second prompt, both models classified all cholesteatoma cases as pars flaccida, consistently across S1 and S2.

Reproducibility and inter-model agreement are presented in Table 4. Between-session agreement was fair for GPT-5 (86 of 101 concordant; κ: 0.360, 95% CI: 0.061–0.659; P = 0.606) and moderate for Gemini 2.5 Pro (97 of 101 concordant; κ: 0.485, 95% CI: –0.010 to 0.980; P = 0.134). Neither between-session comparison demonstrated a statistically significant asymmetric shift across sessions. However, these less-than-perfect kappa values indicate incomplete agreement between sessions, from which output inconsistency can be indirectly inferred. Crucially, because the 95% CI for Gemini 2.5 Pro included zero, its repeatability could not be statistically distinguished from chance.

Because neither model can serve as a reference for the other, inter-model agreement was examined descriptively, as an index of output consistency across interchangeable tools rather than of diagnostic validity. Inter-model agreement was slight at both sessions: GPT-5 and Gemini 2.5 Pro agreed in 85 of 101 cases at S1 (κ: 0.035, 95% CI: –0.399 to 0.469; P = 0.211) and 86 of 101 at S2 (κ: 0.086, 95% CI: –0.341 to 0.513; P = 0.002). Because both kappa intervals included zero, this significant result reflects asymmetric discordance between the models rather than above-chance agreement. After Bonferroni correction within this family of four comparisons (α: 0.0125), the significant S2 inter-model result (P = 0.002) was preserved.

Discussion

This study evaluated GPT-5 and Gemini 2.5 Pro in differentiating cholesteatoma from non-cholesteatomatous COM using key-image temporal bone HRCT. Both models demonstrated a systematic positive classification bias toward cholesteatoma, characterized by high sensitivity but very low specificity, resulting in near-chance balanced accuracy. Furthermore, the analysis revealed limited within-model reproducibility between sessions and slight inter-model agreement.

The dominant finding is clinical over-calling (i.e., a systematic tendency to assign the positive label); because both models labeled most cases as cholesteatoma, their apparently high detection of true disease reflects this bias rather than genuine discrimination. As the models returned categorical labels without localization, a correct label did not imply that the relevant structure was recognized; near-universal findings, such as epitympanic involvement, were correctly labeled "present" almost regardless of discriminative ability. The low specificity for scutum and ossicular erosion is more informative, indicating that both models reported erosion when it was absent and did not reliably separate partial-volume effects or mucosal thickening

Table 1. Baseline characteristics and CT features of the study cohort, stratified by pathological diagnosis

Variable	Cholesteatoma (n = 48)	COM (n = 53)	*P* value
Age, years (mean ± SD)[a]	42.06 ± 15.01	45.75 ± 17.41	0.259
Gender, n (%)			
Male[b]	32 (66.7%)	24 (45.3%)	0.050
Female	16 (33.3%)	29 (54.7%)	—
Cholesteatoma subtype, n (%)			
Combined type (pars flaccida + pars tensa)	34 (70.8%)	—	—
Pars flaccida type	12 (25.0%)	—	—
Congenital type	2 (4.2%)	—	—
CT features			
Scutum erosion, n (%)			
Present[b]	48 (100.0%)	24 (45.3%)	< 0.001
Absent	0 (0.0%)	29 (54.7%)	—
Ossicular erosion, n (%)			
Present[b]	45 (93.8%)	6 (11.3%)	< 0.001
Absent	3 (6.3%)	47 (88.7%)	—
Tegmen dehiscence, n (%)			
Present[b]	32 (66.7%)	19 (35.8%)	0.004
Absent	16 (33.3%)	34 (64.2%)	—
Epitympanic involvement, n (%)			
Present[c]	48 (100.0%)	52 (98.1%)	1.000
Absent	0 (0.0%)	1 (1.9%)	—
Mastoid involvement, n (%)			
Present[b]	40 (83.3%)	47 (88.7%)	0.437
Absent	8 (16.7%)	6 (11.3%)	—
Tympanic membrane perforation, n (%)			
Present[c]	48 (100.0%)	46 (86.8%)	0.013
Absent	0 (0.0%)	7 (13.2%)	—

Continuous variables are reported as mean ± SD; categorical variables as n (%). Statistical tests: [a]independent samples t-test (continuous variables); [b]Pearson's chi-square test (categorical variables when all expected cell frequencies were ≥ 5); [c]Fisher's exact test (categorical variables when at least one expected cell frequency was < 5). All tests were two-tailed; $P < 0.05$ was considered statistically significant. COM, chronic otitis media; SD, standard deviation; CT, computed tomography.

from true bone loss. A tool that flags nearly every ear as cholesteatoma offers little decision support because its positive output rarely changes management. This is quantitatively evident in the PPVs (44.9%–48.4% across models and sessions), which were essentially identical to the cholesteatoma prevalence in the cohort (47.5%); a positive model output therefore conveyed almost no information beyond the pre-test probability.

Over-calling appears to be a recurrent but model-dependent failure of current general-purpose LLMs rather than a peculiarity of this task. Büyüktoka et al.[22] evaluated five multimodal LLMs for bone-lesion detection and reported markedly heterogeneous behavior, with some platforms reproducing the high-sensitivity, very-low-specificity profile observed here, whereas others remained conservative. Consistent with this, the same key-image approach applied to paranasal sinus CT showed performance differing markedly across versions and vendors, with GPT-5 tending toward false-negative-prone classification—the exact opposite of our findings.[11] The divergence in error tendencies across these related head and neck CT tasks indicates that the direction of diagnostic error is task- and version-specific; therefore, error profiles cannot be assumed to transfer across clinical applications. A related classification failure was observed for subtype assignment: both models defaulted exclusively to the pars flaccida subtype, failing to identify congenital cases or the combined types that predominated in our cohort. This likely reflects the near-ubiquitous epitympanic involvement of combined-type disease, together with the rarity of congenital cholesteatoma, which typically lacks overt erosion or perforation.

Table 2. Diagnostic performance of the radiologist and the two large language models against the pathological reference standard

Rater	Accuracy, % (95% CI)	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)	Cohen κ (95% CI)	*P* (κ)	Balanced accuracy (95% CI)
Radiologist	96.0 (90.3–98.4)	97.9 (89.1–99.6)	94.3 (84.6–98.1)	94.0 (83.8–97.9)	98.0 (89.7–99.7)	0.921 (0.845–0.997)	< 0.001	0.961 (0.924–0.998)
GPT-5 (S1)	43.6 (34.3–53.3)	83.3 (70.4–91.3)	7.5 (3.0–17.9)	44.9 (35.0–55.3)	33.3 (13.8–60.9)	−0.088 (−0.274 to 0.099)	0.357	0.454 (0.391–0.518)
Gemini 2.5 Pro (S1)	49.5 (40.0–59.1)	95.8 (86.0–98.8)	7.5 (3.0–17.9)	48.4 (38.6–58.3)	66.7 (30.0–90.3)	0.032 (−0.155 to 0.219)	0.735	0.517 (0.471–0.562)
GPT-5 (S2)	46.5 (37.1–56.2)	83.3 (70.4–91.3)	13.2 (6.5–24.8)	46.5 (36.3–57.0)	46.7 (24.8–69.9)	−0.033 (−0.221 to 0.155)	0.728	0.483 (0.413–0.552)
Gemini 2.5 Pro (S2)	47.5 (38.1–57.2)	97.9 (89.1–99.6)	1.9 (0.3–9.9)	47.5 (37.9–57.2)	50.0 (9.5–90.5)	−0.002 (−0.188 to 0.184)	0.984	0.499 (0.472–0.526)

Cholesteatoma was the positive class, and non-cholesteatomatous chronic otitis media was the negative class. The 95% CIs for accuracy, sensitivity, specificity, PPV, and NPV were calculated using the Wilson score method; the 95% CI for κ was derived from the standard error of Cohen's κ, and *P*(κ) denotes its significance by a z-test. Because both models produced binary outputs, a conventional receiver operating characteristic-based area under the curve was not applicable; balanced accuracy [(sensitivity + specificity)/2] is reported as a descriptive single-operating-point measure, with 95% CIs from the Wald approximation for the mean of two independent proportions. McNemar $P < 0.001$ for both radiologist versus GPT-5 (S1) and radiologist versus Gemini 2.5 Pro (S1). CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value; κ, Cohen kappa coefficient; S1, session 1; S2, session 2.

Table 3. Feature-level detection performance of GPT-5 (S1) and Gemini 2.5 Pro (S1) compared with the radiologist-defined reference checklist

CT Feature	Present/ absent (n)	GPT-5 (S1)				Gemini 2.5 Pro (S1)			
		Accuracy (%)	Sensitivity (%)	Specificity (%)	TP/FP/FN	Accuracy (%)	Sensitivity (%)	Specificity (%)	TP/FP/FN
Scutum erosion	72/29	58.4	75.0	17.2	54/24/18	63.4	83.3	13.8	60/25/12
Ossicular erosion	51/50	45.5	80.4	10.0	41/45/10	51.5	90.2	12.0	46/44/5
Tegmen dehiscence	51/50	48.5	68.6	28.0	35/36/16	58.4	78.4	38.0	40/31/11
Epitympanic involvement	100/1	87.1	88.0	0.0	88/1/12	92.1	93.0	0.0	93/1/7
Mastoid involvement	87/14	88.1	98.9	21.4	86/11/1	90.1	96.6	50.0	84/7/3
TM perforation	94/7	43.6	40.4	85.7	38/1/56	56.4	54.3	85.7	51/1/43

Performance metrics were calculated for each CT feature individually by comparing model outputs against the radiologist-defined 6-item checklist as the reference standard. Session 1 data only. Present/Absent: number of cases in which the feature was present or absent in the reference checklist. CT, computed tomography; TP, true positive; FP, false positive; FN, false negative; TM, tympanic membrane.

Table 4. Short-interval reproducibility and inter-model agreement of GPT-5 and Gemini 2.5 Pro

Comparison	Agreement (n/101)	Cohen's κ	95% CI lower	95% CI upper	McNemar *P*	Interpretation
Within-model agreement (between sessions)						
GPT-5: S1 vs. S2	86/101	0.360	0.061	0.659	0.606	Fair
Gemini 2.5 Pro: S1 vs. S2	97/101	0.485	−0.010	0.980	0.134	Moderate
Inter-model agreement						
GPT-5 vs. Gemini 2.5 Pro (S1)	85/101	0.035	−0.399	0.469	0.211	Slight
GPT-5 vs. Gemini 2.5 Pro (S2)	86/101	0.086	−0.341	0.513	0.002	Slight

CI, confidence interval; κ, Cohen's kappa coefficient; McNemar *P*, *P* value from McNemar's test with continuity correction, which assesses the symmetry of discordant pairs. Agreement: number of cases with concordant binary classification out of 101. Kappa interpretation follows Landis and Koch:[21] < 0.00, poor (worse than chance); 0.00–0.20, slight; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; 0.81–1.00, almost perfect. The McNemar $P = 0.002$ for GPT-5 vs. Gemini 2.5 Pro (S2) is statistically significant after Bonferroni correction, indicating significant asymmetric discordance driven by Gemini 2.5 Pro's near-universal positive classification in S2 (99/101 cases labeled cholesteatoma).

Beyond diagnostic accuracy, reproducibility was limited. Applied twice to identical images and prompts, both models showed only fair-to-moderate between-session agreement, and for Gemini 2.5 Pro, this agreement could not be distinguished from chance. Inter-model agreement was likewise minimal. For a preoperative decision aid, this instability matters as much as accuracy: an output that may change on re-query or with the choice of vendor cannot serve as a dependable second opinion.[23]

Against this background of instability and low specificity, the expert radiologist reading the same six slices achieved near-perfect agreement with the surgical reference standard, far exceeding both models at the case level. This aligns with prior work showing close concordance between expert HRCT interpretation and operative findings[24] and indicates that subspecialty experience remains central to reliable temporal bone assessment.

Most previous imaging studies on cholesteatoma have used task-specific deep learning (DL) models trained directly on CT or magnetic resonance imaging (MRI), reporting accuracies exceeding 90%—comparable with MRI-based reference performance.[25-27] The contrast between the high accuracy reported for fixed-weight, deterministic DL models and the lower, less reproducible performance of the vendor-hosted LLMs—although based on an indirect comparison across different cohorts—indicates that task-specific training remains a meaningful advantage over zero-shot, generalist models at the current state of the technology.

Comparison with prior otologic work is consistent with this interpretation. Noda et al.[10] evaluated GPT-4V on 190 otoscopic images supplemented by clinical data and reported approximately 80% accuracy in classifying middle ear disease, intermediate between non-specialists and otolaryngologists. The substantially lower case-level accuracy in our study likely reflects a more challenging input modality together with our deliberate withholding of clinical history. Our image-only accuracies align with LLM benchmarks in other radiological domains, where image-only performance is likewise modest and improves substantially once descriptive clinical text is added.[28-30] The convergence of this pattern across modalities indicates that, although the direction of error varies across models, the modest image-only performance itself reflects a general limitation of current multimodal LLMs rather than a feature peculiar to temporal bone HRCT.

This study has several limitations. First, this was a retrospective, single-center analysis; larger multicenter datasets including rarer pathologies would enhance generalizability. Second, external validation is lacking, which limits applicability to broader clinical settings. Third, our reliance on selected "key images" constitutes an important limitation, as this image-set approach may not capture all volumetric information present in a complete series. Although these slices were selected by a predefined checklist and consensus to ensure high-yield findings, this manual curation inherently introduces selection bias. Specifically, because the expert radiologist who interpreted the images also participated in their consensus selection, familiarity with the chosen views may have inflated the radiologist's apparent diagnostic performance relative to a real-world workflow in which full DICOM stacks are interpreted without prior filtration. Both the selection and the interpretation were nonetheless performed blinded to the surgical and histopathological reference standard, so the curation could not encode outcome information, and the matched-input design ensured that the models and the radiologist assessed identical images. Importantly, this bias does not affect the study's primary findings, which concern the models: the LLMs were scored against the surgical and histopathological reference standard rather than against the radiologist; therefore, their high sensitivity, low specificity, near-chance balanced accuracy, and limited reproducibility are independent of expert performance. Fourth, current multi-

modal LLMs cannot directly process full-volume DICOM datasets. Converting the source data to a standard display format for upload (TIFF to PNG) entails a departure from the full bit depth and metadata of the original DICOM. Although lossless formats were used at each step, any such conversion may still affect model performance. Fifth, facial nerve canal dehiscence and lateral semicircular canal fistula could not be formally included in the feature-level analysis owing to their low prevalence and the constraints of the key-image approach; their evaluation requires full DICOM series in a larger cohort. Sixth, the zero-shot prompting strategy reflects typical user interaction but may not represent the maximum potential of fine-tuned or retrieval-augmented systems. In addition, only two proprietary models were evaluated; the findings are therefore model-specific rather than a general property of multimodal LLMs, and additional models must be tested before broader conclusions can be drawn. Finally, proprietary "black-box" systems change over time: our findings reflect specific model versions accessed in August 2025 and represent a temporary snapshot, underscoring the necessity of continuous monitoring.

Several directions follow from these findings. As the technical capabilities of these models expand, future studies should evaluate them on complete DICOM series rather than selected key images, which would better reflect real-world interpretation and remove the selection bias inherent in manual image curation. Access through application programming interfaces, rather than public web interfaces, would further allow inference under stable, documented model versions and improve reproducibility. Building on this, the essential next step is prospective, multicenter validation against an independent reference standard, which would test generalizability across scanners and populations. Future work should also examine whether structured or optimized prompting, the integration of relevant clinical information, or task-specific fine-tuning can mitigate the low specificity and over-calling observed here. Given the rapid and often silent iteration of proprietary systems, continuous benchmarking—ideally on shared, expert-annotated temporal bone datasets—will be needed to track whether diagnostic accuracy and stability improve over time.

In this single-center study, GPT-5 and Gemini 2.5 Pro combined high sensitivity with low specificity, were only fair-to-moderately reproducible between sessions, and showed slight inter-model agreement on temporal bone key-image HRCT. In the versions tested, they are therefore not suitable as independent second readers, and generalization to other multimodal LLMs would require evaluating additional models. Until task-specific calibration, version stability, and external validation are demonstrated, their outputs should be treated as exploratory and reviewed by a qualified radiologist.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Supplementary Material 1: https://d2v96fxpocvxx.cloudfront.net/4458d962-0fb2-48a3-a23a-e35265f70f9b/content-images/96f31dc5-91c2-437c-adbe-1b7a4761c818.pdf

Supplementary Material 2: https://d2v96fxpocvxx.cloudfront.net/1e8a7e-ae-47cf-4a4e-a540-9cee47620b04/content-images/989e4a5a-d486-4845-80fc-843f5c876bee.pdf

Supplementary Material 3: https://d2v96fxpocvxx.cloudfront.net/4458d962-0fb2-48a3-a23a-e35265f70f9b/documents/Supplementary%20Material%203.docx

References

1. Kaspar A, Newton O, Kei J, Driscoll C, Swanepoel W, Goulios H. Prevalence of otitis media and risk-factors for sensorineural hearing loss among infants attending Child Welfare Clinics in the Solomon Islands. *Int J Pediatr Otorhinolaryngol*. 2018;111:21-25. [Crossref]
2. Probst R. The middle ear. In: Probst R, Grevers G, Iro H, editors. Basic otorhinolaryngology: a step-by-step learning guide. 2nd ed. Stuttgart: Georg Thieme Verlag; 2006. p. 227-253. [Crossref]
3. Watts S, Flood LM, Clifford K. A systematic approach to interpretation of computed tomography scans prior to surgery of middle ear cholesteatoma. *J Laryngol Otol*. 2000;114(4):248-253. [Crossref]
4. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. [Crossref] Erratum in: *Nature*. 2023;620(7973):E19. [Crossref]
5. Amin KS, Mayes LC, Khosla P, Doshi RH. Assessing the efficacy of large language models in health literacy: a comprehensive cross-sectional study. *Yale J Biol Med*. 2024;97(1):17-27. [Crossref]
6. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology*. 2024;310(1):e232756. [Crossref]
7. Nam Y, Kim DY, Kyung S, et al. Multimodal large language models in medical imaging: current state and future directions. *Korean J Radiol*. 2025;26(10):900-923. [Crossref]
8. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol*. 2024;30(2):80-90. [Crossref]
9. Keshavarz P, Bagherieh S, Nabipoorashrafi SA, et al. ChatGPT in radiology: a systematic review of performance, pitfalls, and future perspectives. *Diagn Interv Imaging*. 2024;105(7-8):251-265. [Crossref]
10. Noda M, Yoshimura H, Okubo T, et al. Feasibility of multimodal artificial intelligence using GPT-4 vision for the classification of middle ear disease: qualitative study and validation. *JMIR AI*. 2024;3:e58342. [Crossref] Erratum in: *JMIR AI*. 2024;3:e62990. [Crossref]
11. Yağcı B, Palaz S, Senirli RT. Reliability of multimodal LLMs for sinusitis with polyps vs. sinusitis without polyps classification from paranasal sinus CT slices. *J Imaging Inform Med*. 2026. [Crossref]
12. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off J Eur Union*. 2024;L2024/1689:1-144. [Crossref]
13. Park SH, Suh CH, Lee JH, Kahn CE, Moy L. Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM). *Korean J Radiol*. 2024;25(10):865-868. [Crossref]
14. Mese I, Akinci D'Antonoli T, Bluethgen C, et al. Reporting checklist for foundation and large language models in medical research (REFINE): an international consensus guideline. *Diagn Interv Radiol*. 2026. [Crossref]
15. Cavaliere M, Ugga L, Monfregola A, et al. Temporal bone CT-based anatomical parameters associated with the development of cholesteatoma. *Radiol Med*. 2023;128(9):1116-1124. [Crossref]
16. Uz Zaman S, Rangankar V, Muralinath K, Shah V, K G, Pawar R. Temporal bone cholesteatoma: typical findings and evaluation of diagnostic utility on high resolution computed tomography. *Cureus*. 2022;14(3):e22730. [Crossref]

17. OpenAI. Introducing GPT-5 [Internet]. San Francisco (CA): OpenAI; 2025 Aug 7 [cited 2025 Aug 30]. [Crossref]
18. Kavukcuoglu K. Gemini 2.5: our most intelligent AI model [Internet]. Mountain View (CA): Google; 2025 Mar 25 [cited 2025 Aug 30]. [Crossref]
19. OpenAI. ChatGPT [Internet]. San Francisco (CA): OpenAI; 2025 [cited 2025 Aug 30]. [Crossref]
20. Google. Gemini [Internet]. Mountain View (CA): Google; 2025 [cited 2025 Aug 30]. [Crossref]
21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-174. [Crossref]
22. Büyüktoka RE, Salbas A, Cilengir AH, Buyuktoka AD, Sürücü M, Adibelli ZH. Performance of multimodal large language models for the detection and characterization of bone lesions on radiographs. *Diagn Interv Radiol*. 2026. [Crossref]
23. Evstafev E. The paradox of stochasticity: limited creativity and computational decoupling in temperature-varied LLM outputs of structured fictional data. *arXiv [Preprint]*. 2025. [Crossref]
24. Stefanescu EH, Balica NC, Motoi SB, Grigorita L, Georgescu M, Iovanescu G. High-resolution computed tomography in middle ear cholesteatoma: how much do we need it? *Medicina (Kaunas)*. 2023;59(10):1712. [Crossref]
25. Shabi SM, Almutairi LB, Moafa AN, et al. Artificial intelligence in the diagnosis of cholesteatoma: a systematic review of current evidence. *Cureus*. 2025;17(11):e96154. [Crossref]
26. Petsiou DP, Martinos A, Spinos D. Applications of artificial intelligence in temporal bone imaging: advances and future challenges. *Cureus*. 2023;15(9):e44591. [Crossref]
27. Eroğlu O, Eroğlu Y, Yıldırım M, et al. Comparison of computed tomography-based artificial intelligence modeling and magnetic resonance imaging in diagnosis of cholesteatoma. *J Int Adv Otol*. 2023;19(4):342-349. [Crossref]
28. Han T, Jeong WK, Shin J. Diagnostic performance of multimodal large language models in radiological quiz cases: the effects of prompt engineering and input conditions. *Ultrasonography*. 2025;44(3):220-231. [Crossref]
29. Schramm S, Preis S, Metz MC, et al. Impact of multimodal prompt elements on diagnostic performance of GPT-4V in challenging brain MRI cases. *Radiology*. 2025;314(1):e240689. [Crossref]
30. Atakır K, Işın K, Taş A, Önder H. Diagnostic accuracy and consistency of ChatGPT-4o in radiology: influence of image, clinical data, and answer options on performance. *Diagn Interv Radiol*. 2025. [Crossref]