

Large Language Models in Hearing Care

Shangqiguo Wang

`sqgw@connect.hku.hk`

The University of Hong Kong <https://orcid.org/0009-0008-6356-9604>

Changgeng Mo

Education University of Hong Kong

Jibin Wu

Hong Kong Polytechnic University

Alice K. Y. Siu

Chinese University of Hong Kong Shenzhen <https://orcid.org/0000-0002-2318-6050>

Fei Zhao

Cardiff Metropolitan University

Jan-Willem A. Wasmann

Radboud University Medical Center

De Wet Swanepoel

University of Pretoria

Vinaya Manchaiah

University of Colorado Anschutz Medical Campus

Systematic Review

Keywords: hearing loss, hearing care, audiology, large language models, generative artificial intelligence

Posted Date: October 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11384026/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Large language models (LLMs) may support hearing care, but evidence of clinical benefit remains unclear. We conducted a scoping review across eight databases, with searches updated in September 2026. Among 110 included reports, 95 empirical analysis units comprised 62 model-centered benchmarks, 31 application-oriented investigations, and two adoption surveys. Applications addressed assessment scoring, clinical documentation, rehabilitation content, biomedical evidence processing, psychosocial support, and research. Fourteen empirical units evaluated prototypes or target users; none tested prospective service deployment or comparative clinical impact. Privacy evaluation concerned attitudes or perceived confidentiality; no included study tested technical privacy safeguards. Current evidence supports selected task-specific capabilities, not established improvements in care. Future research should prioritize applications addressing defined hearing-care needs and compare them with usual care or simpler alternatives. Evaluation should measure patient and professional benefit, safety, privacy, accessibility, and workload rather than response accuracy alone.

Introduction

Large language models (LLMs) are artificial intelligence models trained on large-scale text data to learn language patterns and process or generate text.¹ Depending on their architecture and system configuration, they can also process multimodal inputs, retrieve external evidence, and use tools.^{2,3} A recent review identified 4,609 clinical LLM studies published from January 2022 to September 2025.⁴ In an earlier review of 519 studies published through February 19, 2024, 84.2% evaluated responses to medical questions, including knowledge and clinical-reasoning tasks, 5% used real patient-care data, and 4.6% considered deployment.⁵ Evidence of patient, clinician, or service benefit remains scarce. Benchmark performance does not equate to clinical performance. Examination and question-answering studies measure responses to predefined questions or cases; they do not establish competence in information elicitation, longitudinal reasoning, evidence use, communication, or workflow integration.^{1,6} Newer frameworks therefore include documentation, patient communication, decision support, research, administration, and multistep interaction.^{3,7,8} Evaluation must address the configured system within a defined workflow, not a model name alone.

The World Health Organization estimated that more than 1.5 billion people had hearing loss and documented substantial shortages and unequal distribution of hearing-care professionals.⁹ These service gaps motivate evaluation of LLMs for accessible information, triage, and continuing rehabilitation support, without assuming that automation improves access or outcomes.^{10,11} In hearing care, evaluation must reflect the range of tasks, settings, and users. Integrated people-centered ear and hearing care spans prevention, screening, assessment, medical and surgical management, hearing technologies, rehabilitation, and continuing support across the life course.⁹ These services may be delivered in clinics, through teleaudiology, or through direct-to-consumer channels, each with different opportunities for professional oversight.¹² The underlying data are heterogeneous: conversation, clinical notes, audiograms, speech-in-noise responses, otoscopy and imaging, device records, and genetic or neurophysiologic measures may all contribute to one episode of care.^{13,14} Many intended users are older adults or people facing communication, literacy, linguistic, or technology barriers. Consequently, factual accuracy alone is insufficient. A useful system must also be understandable, accessible, responsive to individual goals, and safe when information is incomplete.^{9,15} Evaluation findings may become outdated when the tested system changes. Hosted systems may change checkpoint, interface, web access, retrieval, tools, or

safety policies without changing product name; access date, instructions, conversation history, and prompts can alter results but are often underreported.^{16,17} Retesting, current-literature instructions, and guideline-grounded retrieval have changed hearing-care and ear, nose, and throat (ENT) findings.^{18,19,20} Reported performance therefore applies to the tested version and configuration at the time of evaluation and may not generalize to later versions or different settings.

Existing audiology and otology reviews have generally considered artificial intelligence broadly, combining LLMs with machine learning, deep learning, imaging, signal processing, and prediction.^{21,22,23} Three LLM-focused ENT reviews overlap with part of the present topic but answer narrower or different questions. Filali Ansary and Lechien²⁴ synthesized clinical decision support in 17 studies and found wide variation in diagnostic, investigation, and treatment accuracy. Ahn et al.²⁵ surveyed 25 studies from top-quartile ENT journals and organized future work around evaluation, grounding, multimodality, agents, and explainability. Alter et al.²⁶ examined methodology reporting in 117 ENT studies and found that fewer than half quoted all prompts. These reviews are ENT-wide; they do not map the full hearing-care continuum or separate technical role from proximity to service implementation.

We conducted a scoping review of LLM research across hearing assessment, medical management, hearing-device support, rehabilitation, and related professional and research tasks to answer three questions: (1) which hearing-care tasks and LLM functions have been studied; (2) in which settings systems have been evaluated, from static benchmarks to routine service delivery; and (3) how completely studies report the system tested and outcomes related to safety, communication, privacy, accessibility, and equity. We use this evidence map to identify evaluation priorities and propose evidence requirements for future hearing-care LLM applications.

Results

Evidence Selection, Publication Profile, and Service Scope

The original and updated database searches retrieved 2,248 records. After removing 1,187 duplicate or previously identified record instances and 194 records published before 2022, 867 unique records were screened. Of these, 706 were excluded at title and abstract screening. Full texts were sought for 161 database reports; two could not be retrieved. Of the 159 reports assessed, 52 were excluded and 107 were included. Citation and author-publication searching contributed three additional reports, giving 110 included reports (Figure 1). A conference abstract and journal article reported the same coding investigation;^{27,28} collapsing this pair yielded 95 empirical analysis units from 96 empirical reports.

The remaining 14 reports comprised six reviews and eight commentaries or methodological correspondence. At report level, 54 concerned core hearing, audiology, or rehabilitation; 20 hearing-specific otology; 27 mixed or general ENT samples; and nine broader-scope sources with an identifiable hearing component (Table 1). The hearing-specific sensitivity set contained 66 empirical analysis units. Report counts by publication year were two in 2023, 26 in 2024, 34 in 2025, and 48 in 2026 through the update. Application-oriented research accounted for 6 of 22 empirical analysis units (27.3%) in 2024, 9 of 29 (31.0%) in 2025, and 16 of 43 (37.2%) in 2026. The sole empirical unit from 2023 was a benchmark evaluation. The increasing annual proportion persisted when

the linked coding investigation was assigned to its earliest report year and when adoption surveys were excluded. Supplementary Data 1 lists report-level characteristics and analysis summaries.

Table 1. Research Role, Service Scope, and Implementation Context of Included Sources

Percentages use 110 reports. Analysis units collapse the known conference/journal report pair; contextual reports remain separate sources. The 96 empirical reports yield 95 empirical analysis units. Units do not establish participant independence; possible formative-sample overlap remains across two psychosocial reports. Context categories describe evaluation setting, not quality. Percentages may not sum to 100 because of rounding. ENT, ear, nose and throat; LLM, large language model.

Dimension	Category	Reports, n (%)	Analysis units, n
Primary research role	Patient or public information benchmark	26 (23.6)	26
Primary research role	Professional knowledge or examination benchmark	9 (8.2)	9
Primary research role	Clinical reasoning, guideline, or decision benchmark	24 (21.8)	24
Primary research role	Multimodal clinical interpretation benchmark	3 (2.7)	3
Primary research role	Task-specific output evaluation	15 (13.6)	14
Primary research role	Domain model or system prototype	10 (9.1)	10
Primary research role	Patient/recipient evaluation	4 (3.6)	4
Primary research role	LLM-enabled research workflow	3 (2.7)	3
Primary research role	Professional use or adoption survey	2 (1.8)	2
Primary research role	Review	6 (5.5)	6
Primary research role	Commentary or methodological correspondence	8 (7.3)	8
Service-scope stratum	Core hearing, audiology, or rehabilitation	54 (49.1)	54
Service-scope stratum	Hearing-specific medical ear or otology	20 (18.2)	20
Service-scope stratum	Mixed or general ENT with nonseparable hearing results	27 (24.5)	26
Service-scope stratum	Broader-scope source with an identifiable hearing component	9 (8.2)	9
Implementation context	M0: observed use or adoption	2 (1.8)	2
Implementation context	M1: static benchmark	49 (44.5)	49
Implementation context	M2: task-specific or real-data evaluation outside routine care	31 (28.2)	30
Implementation context	M3: prototype or target-user evaluation	14 (12.7)	14
Implementation context	M4: prospective service pilot	0 (0.0)	0
Implementation context	M5: comparative clinical or service impact	0 (0.0)	0
Implementation context	Not applicable: contextual report	14 (12.7)	14

Reviews and correspondence were contextual rather than additional empirical experiments. They identified external validation, representative populations, data standards, interpretability, governance, explicit reference

standards, reproducible prompts, model drift, and lay understanding as recurring priorities.^{17,21,22,23,29,30,31,32,33} Hearing-health commentaries also proposed screening, triage, rehabilitation, device support, and research while emphasizing privacy, language representation, cost, and locally governed models.^{10,11,34} Methodological correspondence further showed why dependent expert ratings require appropriate analysis: a reply concerning cochlear-implant education retained descriptive model rankings, but no pairwise difference remained significant after Holm correction.^{35,36,37} The reply was linked to the original study rather than counted as a new empirical unit.

Implementation Context, Reporting Completeness, and Sensitivity Analyses

Among 95 empirical analysis units, two examined adoption, 49 were static benchmarks (51.6%), 30 evaluated tasks or real data outside routine care, and 14 evaluated prototypes or target users. None prospectively introduced an LLM into a hearing service or compared an LLM-enabled workflow with usual care.

Implementation context and primary research role were distinct: a benchmark could use real clinical data without becoming a deployed application.

OpenAI-family models or platforms appeared in 77 of 95 empirical analysis units (81.1%) and Google-family models or platforms in 33 (34.7%). These nonexclusive counts describe reported models and platforms, not current capability or market share. Model family or platform was clearly identified in 94 units (98.9%), and a specific variant or checkpoint in 68 (71.6%). An access or run date was reported in 44 of 88 applicable units (50.0%). Complete or supplemental prompts appeared in 52 of 88 applicable units (59.1%), clear reference standards in 74 of 93 (79.6%), and clear evaluator backgrounds in 69 of 79 (87.3%).

Safety or explicit-error assessment was reported in 43 of 95 empirical analysis units (45.3%), readability or communication assessment in 50 (52.6%), and accessibility or health-literacy assessment in 16 (16.8%). These broad categories included technical and qualitative outcomes, not necessarily validated patient-safety or accessibility endpoints. Three units (3.2%) assessed privacy-related attitudes or perceived confidentiality: two professional surveys and one qualitative investigation. None tested technical privacy protection. No unit prespecified a formal equity or fairness analysis (0/95; 0%); 16 (16.8%) reported group comparisons without framing them as equity analyses. Language-specific evidence was inconsistently described across reports.

Sensitivity analyses supported the same interpretation. Among 66 hearing-specific empirical analysis units, 42 were model-centered benchmarks, 34 were static benchmarks, and none reached prospective service or comparative impact evaluation. Excluding ten conference-proceedings-only units left 85 empirical units: 61 (71.8%) were model-centered benchmarks, 48 (56.5%) were static benchmarks, and seven involved prototypes or target users. The known conference/journal coding pair remained represented by its journal report. These are descriptive checks within the included literature, not inferential evidence of a field-wide trend.

Model-Centered Benchmark Studies

Sixty-two of 95 empirical analysis units (65.3%) were model-centered benchmarks: 26 patient-information, nine professional-knowledge, 24 clinical-reasoning, and three multimodal benchmarks involving inputs such as

audiograms or magnetic resonance imaging (MRI) videos (Table 1; Figure 2). Heterogeneous prompts, cases, instruments, and thresholds precluded pooling. Patient-information studies addressed hearing loss, tinnitus, otosclerosis, auditory implants, hearing-aid management, postoperative care, and related topics. They commonly rated accuracy, completeness, guideline concordance, readability, actionability, or safety. Sudden-sensorineural-hearing-loss responses were broadly guideline-concordant in English and Spanish, but readability was suboptimal; ChatGPT's English responses had a mean PEMAT-P actionability score of 60%.³⁸ Repeated tinnitus questions produced changing expert ratings and potentially misleading responses.¹⁸ Updated studies likewise distinguished expert-rated quality from accessibility: tinnitus information with higher quality scores could require a higher reading level, while microtia counseling omitted clinical nuance and often exceeded recommended reading levels.^{39,40} A 44-question hearing-aid study assessed expert-rated answers and textual repeatability, not actual troubleshooting by users.⁴¹ No patient-information benchmark measured comprehension, care-seeking behavior, decisional quality, or health outcomes.

Professional-knowledge studies used examinations, certification questions, or specialty banks. ChatGPT-4 answered 75% of 299 questions from the 2023 Taiwan audiologist examination, above its 60% passing threshold;⁴² later studies compared newer systems and trainees.^{43,44} The Examinations for Medical PErsoneel in Chinese (EMPEC) benchmark expanded model and year coverage using 5,600 questions from the same Taiwan examination source, rather than an independent clinical task.⁴⁵ These studies measured examination performance, not clinical competence.

Clinical benchmarks tested diagnosis, investigation, referral, treatment, triage, guidelines, or decision quality. Evaluations using clinical records remained outside routine care; no output was delivered prospectively to clinicians or patients. Across 1,000 simulated longitudinal hearing trajectories, majority responses from five ChatGPT queries achieved 0.87 accuracy and Cohen's kappa of 0.69 against expert majority decisions, versus 0.80-0.84 accuracy for single queries.⁴⁶ GPT-4o remained inferior to physician decisions in real sudden-sensorineural-hearing-loss cases.⁴⁷ In 100 auditory-implant candidates, ChatGPT recommendations showed substantial agreement with multidisciplinary decisions, with lower agreement for adult and complex cases; this did not test whether assistance improved decisions.⁴⁸ Numerical audiogram data were also used to approximate clinical diagnoses without testing changes in care.⁴⁹ Explicit requests for current literature and guideline-grounded retrieval changed tympanostomy and emergency ENT ratings.^{19,20}

Three studies tested multimodal clinical interpretation outside routine care. Two used pure-tone audiograms; the best model in a multicenter study achieved 67% accuracy for hearing-loss severity and 54% for type,⁵⁰ whereas another incompletely documented its model checkpoint.⁵¹ Gemini 2.5 interpreted temporal-bone MRI videos but had poor sensitivity for cholesteatomas smaller than 5 mm.⁵² None tested effects on clinical decisions or outcomes. AUDIOLOGYBENCH extended evaluation to 3,139 educational items, 3,175 literature-derived items, and 67 multimodal clinical cases.⁵³ Across 804 case-task evaluations, recommendations scored higher than numerical audiogram interpretation or differential diagnosis. Automated adjudication flagged critical errors in 35.4% of the 268 numerical-interpretation responses, but item-level agreement with an image-informed human audit was poor; human calibration preserved the broad task ranking.

Application-Oriented Studies

Thirty-one application-oriented analysis units were described in 32 reports: 14 task-specific output evaluations, ten domain prototypes, four patient/recipient evaluations, and three research workflows (Table 2; Figure 2). LLM functions included document transformation, assessment scoring, evidence extraction, and generation of speech or rehabilitation content. Evaluation ranged from isolated output testing to prototype use. Table 2 provides study-level findings and limitations; Supplementary Data 1 retains configurations, samples, and outcomes.

Table 2. Application-Oriented LLM Research in Hearing Care

Thirty-two reports contribute 31 application-oriented analysis units. The linked coding reports are counted once. M2, task-specific or real-data evaluation outside routine care; M3, prototype or target-user evaluation. These categories do not establish clinical effectiveness. Possible shared formative samples across the Choi reports preclude summing their participants as independent people.

Abbreviations: AUROC, area under the receiver operating characteristic curve; CBT, cognitive-behavioral therapy; DNS, Deep Noise Suppression; EEG, electroencephalography; ENT, ear, nose and throat; fMRI, functional magnetic resonance imaging; GPT, generative pretrained transformer; LLM, large language model; NLP, natural-language processing; OTC, over-the-counter; PMID, PubMed identifier; RAG, retrieval-augmented generation; SNR, signal-to-noise ratio; SSNR, segmental SNR; THI, Tinnitus Handicap Inventory.

Report	Primary role	LLM function	Context	Main finding	Evidence boundary
Herrmann ⁵⁸	Task-specific output evaluation	Automated speech-intelligibility scoring	M2	Embedding scores closely matched human scoring (r approximately 0.95) in typed speech-in-noise transcripts from 144 adults.	Experimental English data; no prospective workflow evaluation or validation of diagnostic cutoffs for clinical use.
Tu et al. ⁶⁰	Domain model or system prototype	Grounded genetic diagnosis and discovery	M3	A grounded pipeline recovered known factors and proposed a hearing-loss gene supported by experiments; experts agreed with most predictions in 20 hearing-loss cases.	Small heterogeneous case series; prompt-sensitive; no blinded diagnostic-accuracy or prospective utility study.
Shaari et al. ⁷⁹	Task-specific output evaluation	Research-topic generation	M2	Experts judged many of 100 proposed review topics feasible and relevant, but none was unique.	Topic ideation only; no completed research workflow or comparison with researcher-generated ideas.
Zhang et al. ⁶⁷	Patient/recipient evaluation	LLM-generated Mandarin speech	M3	In 25 cochlear-implant users, three speaking-rate-matched comparisons detected no statistically significant recognition or response-time differences; slowing the synthetic female voice reduced recognition.	Exact Seed-TTS variant unreported; small single-language laboratory sample; no rehabilitation, adherence, or service outcome.
Herrmann ⁵⁹	Task-specific output evaluation	Multilingual speech-recall scoring	M2	Automated scores captured noise and serial-position effects across 11 languages in 55 adults reporting normal hearing.	Small language-specific samples; no clinical audiologist comparison, diagnostic cutoff, test-retest reliability, or service study.

Report	Primary role	LLM function	Context	Main finding	Evidence boundary
Lotto et al. ⁸⁰	Task-specific output evaluation	Assessment-item generation	M2	Twenty generated ENT questions were generally grammatical but had average quality and required expert revision when tested with 64 learners and clinicians.	One model and small item set; item-generation efficiency and consequential validity were not established.
Du et al. ⁶²	Domain model or system prototype	Multimodal tinnitus diagnostic model	M3	The EEG-fMRI and LLM framework outperformed reported cross-subject technical baselines.	Target dataset comprised 20 people; no independent cohort, prospective testing, or clinical utility assessment.
Müller et al. ⁷⁷	LLM-enabled research workflow	LLM-assisted statistical exploration	M2	An LLM helped explore 1,365 regression specifications for speech outcomes after Vibrant Soundbridge implantation.	Twenty-patient single-center dataset; final predictor was conventional regression without independent testing.
Habib et al. ⁸¹	Task-specific output evaluation	Qualitative corpus classification	M2	Three LLMs classified 996 cochlear-implant Reddit posts rapidly, but agreement with human coding was only fair and positive predictive value was low.	One social-media community; subjective codebook and nonrepresentative self-reported data.
Kılıçtaş et al. ⁵⁵	Task-specific output evaluation	Operative-note simplification	M2	Six LLMs simplified 39 fictional ENT notes; expert-rated accuracy and clarity varied by model and task.	No real patient notes, patient comprehension, editing burden, downstream safety, or clinical use.
Agarwal and Bitragunta ⁷²	Domain model or system prototype	Speech-denoising model	M3	The report combined DNS-corpus denoising experiments with a Whisper-Large, noise-	Quantitative comparisons did not isolate the LLM contribution; no hearing-device-user or

Report	Primary role	LLM function	Context	Main finding	Evidence boundary
				encoder, and language-model architecture.	clinical-workflow evaluation.
Jeong et al. ⁶³	Domain model or system prototype	Tinnitus text mining and outcome prediction	M3	GPT-2 clustered CBT text and T5 generated post-treatment THI estimates from clinical and synthetic data.	Only 42 original records, extensive synthetic augmentation, overfitting risk, and no external validation.
Andrijašević and Vukelić ⁷¹	Task-specific output evaluation	Phoneme-constrained training text	M2	ChatGPT generated Croatian words and sentences meeting phonetic-saturation constraints, but many compliant sentences lacked natural meaning.	Preliminary linguistic evaluation; no therapist, hearing-aid user, or cochlear-implant user testing.
Fernando et al. ⁶⁸	Domain model or system prototype	Personalized auditory-verbal therapy platform	M3	A RAG-enabled prototype linked clinician input, personalized activities, audio, and mobile and web interfaces.	No reported evaluation with children, parents, or therapists; bespoke technical metrics were not clinically validated.
Liu and Zheng ⁶⁴	Domain model or system prototype	Edge otologic diagnostic system	M3	The paper reported 94% accuracy and favorable latency and power but did not isolate the LLM contribution.	Unclear data provenance, reference standard, ethics, training and test separation, and system description; no external or clinician validation.
Warrier et al. ⁷⁰	Task-specific output evaluation	Patient-education material generation	M2	Materials from 144 hospitals and professional societies were compared with ChatGPT texts, which averaged a 6.8-grade reading level.	Readability and expert content measures did not test patient comprehension, behavior, or clinical outcomes.

Report	Primary role	LLM function	Context	Main finding	Evidence boundary
Alabdalhussein et al. ⁶⁹	Task-specific output evaluation	Patient-information generation	M2	For six ENT conditions, overall between-source comparisons detected no statistically significant differences in expert-rated DISCERN or PEMAT-P scores, whereas LLM texts required higher reading levels than ENT UK materials.	No patient involvement and no formal clinical-accuracy assessment; reviewer variability limited reliability.
Gaeta ⁵⁶	Task-specific output evaluation	Audiology communication outputs	M2	Twenty-one audiologists or AuD students completed at least part of the evaluation of a generated report, note, and counseling script; ratings varied by output.	Single examples; no patient recipient, time-saving measure, routine workflow, or downstream safety evaluation.
Wu ⁷⁸	LLM-enabled research workflow	Child-language corpus annotation	M2	GPT-4o-mini supported exploratory structural judgments on 200 utterances within a larger hearing-loss language study.	Peripheral to service delivery; substantive findings came mainly from conventional corpus and NLP methods.
Bennett et al. ⁷³	Patient/recipient evaluation	Patient-facing hearing-health chatbot	M3	Ten adults used the prototype for two weeks (39 interactions); it was easy to use for basic support but engagement was low and users requested memory and personalization.	Small iPhone-only sample; binary nonblinded accuracy ratings; no usual-care comparator or clinical, service, or cost outcome.
Derbarsegian and colleagues ^{27,28}	Task-specific output evaluation	Otologic coding and billing	M2	Across 191 operative notes, exact code agreement was low and both	Single retrospective task; templated cochlear-implant notes inflated

Report	Primary role	LLM function	Context	Main finding	Evidence boundary
				models omitted component procedures and tended to underbill.	performance relative to variable procedures.
Choi et al. ⁷⁴	Patient/recipient evaluation	AI-assisted songwriting and reflection	M3	All 23 cochlear-implant users completed songwriting; 20 completed reflection. Satisfaction averaged 5.86/7, but some users felt pressured or rushed.	One supervised session; selected users and possible formative-sample overlap with the other Choi report. No treatment efficacy or sustained benefit established.
Choi et al. ⁷⁵	Patient/recipient evaluation	Songwriting diary for psychosocial support	M3	Twelve diary participants completed 12 sessions each over three weeks. A separate 10-person component comparison favored the full system descriptively.	No usual-care comparison or hearing benefit; uncertain handling of within-person comparisons and possible shared formative samples. Superiority to every component was not established.
Younan et al. ⁵⁴	Task-specific output evaluation	Operative-note coding	M2	The LLM correctly coded 107/124 notes (86.3%) versus 61/124 (49.2%) for human coding against the study reference.	Single institution; financial gains were estimated, not observed. Source inconsistencies affect error and revenue summaries. Distinct from the earlier 191-note study.
Lee et al. ⁶⁵	Domain model or system prototype	Genotype feature selection and engineering	M3	LLM-selected or engineered features supported conventional classifiers; hearing-loss AUROCs were approximately 0.50-0.54 in the reported 16-shot comparison.	Near-chance hearing prediction; stronger ancestry results are not hearing outcomes. No clinical validation or demonstrated diagnostic benefit.

Report	Primary role	LLM function	Context	Main finding	Evidence boundary
dos Santos et al. ⁸³	Task-specific output evaluation	Literature information extraction	M2	Across 117 speech-enhancement papers, explicit country/channel information was easier to extract than contextual architecture or application information.	Missing or implicit information produced erroneous or invented answers; manual checking remained necessary. No improvement in speech enhancement or listening was tested.
Nguyen et al. ⁶⁶	Domain model or system prototype	Rare-disease genetic-test selection	M3	A fine-tuned LLM achieved 86% internal accuracy for test-choice prediction but was not significantly superior to the random-forest comparator; external accuracy was 61%.	Only a synthetic hearing-loss illustration; cohort-wide accuracy is not hearing-specific. Historical test choice is not improved diagnosis or prospective clinical benefit.
Washington et al. ⁵⁷	Task-specific output evaluation	Cochlear-implant insurance appeal letters	M2	Of 29 generated letters, 15 contained hallucinations. The reported reference table classified only 6/94 citations as appropriate.	Abstract and table denominators conflict; one older model and no comparator. Letter quality does not establish better insurance approval or reimbursement.
Wickramasinghe et al. ⁷⁶	Domain model or system prototype	Personalized hearing self-management advice	M3	For 10 synthetic profiles, 67/83 advice sentences were rated accurate, 71 relevant and 74 understandable.	Synthetic-profile prototype evaluation, not target-user testing or clinical deployment; no behavior, hearing-preservation or patient benefit was measured.
Ye et al. ⁶¹	Domain model or system prototype	Genetic-variant literature evidence extraction	M3	The mechanism-aware evidence pipeline supported hearing-loss variant	Whole-system AUROC is not the LLM contribution. Resolvable PMIDs do not prove that citations support

Report	Primary role	LLM function	Context	Main finding	Evidence boundary
				interpretation. In a separate rules-off ablation, the LLM module added 0.011 AUROC.	claims; no prospective diagnostic or patient benefit was evaluated.
Mo et al. ⁸²	LLM-enabled research workflow	Abstract-level research synthesis	M2	LLM-assisted analysis of 106 OTC hearing-aid abstracts characterized changes in themes and evidence composition across publication periods.	One database and abstracts only; cross-model agreement is not an expert reference standard. The pre/post comparison cannot establish a causal regulatory effect.

Clinical and administrative applications transformed notes, codes, and correspondence. In 191 deidentified otology notes, ChatGPT-3.5 and GPT-4 exactly matched certified coding in 22% and 14%, respectively, often missing consequential modifiers and graft codes.^{27,28} A separate retrospective evaluation of 124 neurotology notes reported 86.3% coding accuracy for its LLM versus 49.2% for centralized human coders against a surgeon-adjudicated reference; financial gains were projected rather than observed.⁵⁴ These findings arose from different systems and reference procedures and were not a direct model comparison. Other studies simplified operative notes, generated audiology reports and counseling scripts, or drafted cochlear-implant insurance appeals.^{55,56,57} They assessed output accuracy or professional ratings rather than observed reimbursement or patient comprehension.

Assessment automation linked LLM-derived scores to listening tasks. Embedding models scored typed speech-in-noise transcripts from 144 adults with approximately 0.95 correlation to human scores.⁵⁸ An 11-language pipeline generated and synthesized stories and scored free recall from 55 listeners reporting normal hearing, recovering expected noise, order, and serial-position effects.⁵⁹ Neither established test-retest reliability, clinical diagnostic cutoffs, validity in hearing-loss populations, or workflow benefit. Commentary identified additional concerns about proprietary embeddings, model retirement, data transfer, and language representation.³⁴

Diagnostic and analytic systems used LLMs to generate hypotheses, process evidence, or combine data modalities. Med-PaLM 2 generated a hearing-loss gene hypothesis tested in knock-in mice, and a Gemini 2.5 Pro pipeline ranked candidate variants from 20 people with hearing loss using retrieved literature and structured evidence. Two predictions were rejected by experts, and other variants were later ruled out through family evaluation.⁶⁰ In OtoVCE, an LLM extracted literature evidence within a hearing-loss variant-interpretation pipeline; a rules-off ablation attributed an area under the receiver operating characteristic curve (AUROC) increment of 0.011 to the LLM module.⁶¹ Other systems addressed tinnitus, otoscopic assessment, genotype features, or genetic-test selection, with limitations in sample size, external validation, hearing-specific evidence, or isolation of the LLM contribution (Table 2).^{62,63,64,65,66}

Speech and rehabilitation applications generated training content, synthesized audio, or supported personalized therapy. Twenty-five cochlear-implant users and 25 normal-hearing adults heard Mandarin sentences generated by Seed-TTS, described as LLM-based but without an exact variant or checkpoint. Implant users showed no statistically significant recognition or response-time differences in three speaking-rate-matched synthetic-versus-natural comparisons, although slowing the synthetic female voice reduced recognition.⁶⁷ A personalized auditory-verbal-therapy prototype combined patient profiles, clinician feedback, retrieval, generated stories and questions, synthesized audio, and a clinician interface; use or benefit with children, parents, or therapists and the number of clinician users were not reported.⁶⁸ Other work evaluated speech denoising or generated patient and auditory-training materials, without demonstrating improved listening, comprehension, or rehabilitation outcomes.^{69,70,71,72}

Patient-facing evaluations examined usability and engagement outside established clinical services. Ten adults with hearing loss used a custom-GPT hearing-health app for two weeks. It prioritized clinically reviewed troubleshooting content and vetted web sources; across 39 interactions, users valued basic support but requested greater depth, memory, multiquestion handling, and voice input. Audiologists rated all responses accurate, but the binary rubric was unvalidated, assessment was unblinded, and there was no usual-care comparator.⁷³ Two related studies evaluated AI-assisted songwriting and reflection with cochlear-implant users through supervised sessions and a three-week diary study. Participants reported engagement and emotional reflection, alongside pressure or rushing in the supervised study; auditory rehabilitation efficacy and sustained clinical benefit were not evaluated. Possible formative-sample overlap precludes summing participants as independent people.^{74,75} By contrast, a hearing self-management agent was evaluated using ten synthetic profiles rather than actual users.⁷⁶

Research applications supported analysis rather than replacing the final scientific method. GPT-4o's Python environment explored speech-intelligibility prediction after Vibrant Soundbridge implantation, but the final model was conventional regression without independent testing.⁷⁷ In a corpus of 64,814 child utterances, GPT-4o-mini was tested on a balanced subset of 200 utterances as an exploratory adjunct to conventional syntax analysis and performed inconsistently.⁷⁸ Other studies generated research topics or examination items, classified patient narratives, or extracted information from research abstracts and articles. Their evaluations identified unoriginal or incorrectly keyed outputs, limited coding agreement, and a continuing need for human verification.^{79,80,81,82,83}

Two adoption surveys examined professional use and attitudes rather than system performance. One surveyed 106 academics across 16 countries and found specific AI training in 8.5%;⁸⁴ another surveyed 95 audiologists and speech-language pathologists and identified ChatGPT as the most-used platform without quantifying its uptake.⁸⁵ Both addressed AI more broadly than LLMs alone.

Discussion

LLMs have demonstrated capabilities in hearing-related question answering, material generation, automated scoring, information transformation, and research support. These findings provide a basis for developing hearing-care applications, but most evaluations tested static responses, isolated outputs, or prototypes rather than service effectiveness. Neither the broader evidence set nor the sensitivity analyses identified prospective

hearing-service deployment or comparative clinical impact. The increasing annual proportion of application-oriented research suggests greater attention to task and system development within this literature, not a demonstrated transition to effective care. The short observation period, incomplete 2026 coverage, and publication and indexing delays limit temporal interpretation. Performance belongs to a configured system comprising model version, prompts, retrieval, tools, interface, and human workflow.^{16,17} Hearing-care research should therefore test whether a system with defined oversight, approved information sources, and human supervision addresses a defined need and improves relevant outcomes over usual care or a simpler alternative. This pattern is consistent with concerns in the broader medical literature. Bedi et al.⁵ found that most health-care LLM evaluations addressed medical questions and few used real patient-care data. ENT reviews similarly identified variable decision-support performance, limited external validation, and incomplete prompt reporting.^{24,25,26} Their eligibility rules and reporting definitions differ from ours, so percentages should not be treated as direct performance comparisons. Across these reviews, the unresolved issue is whether performance on selected inputs transfers to the decisions, communication, and follow-up required in actual care. Strong benchmark scores and readable outputs are useful intermediate measures, but they cannot substitute for evidence that clinicians or patients use the system appropriately and benefit from doing so.^{3,6,7}

Question-answer benchmarks have helped characterize LLM capabilities and identify misleading advice, unsafe omissions, excessive reading level, and temporal instability.^{17,18,38} They remain useful for monitoring consequential risks and testing hypotheses in new populations, languages, modalities, interactions, or decision contexts. Repeated comparisons can track changes in model behavior, but their contribution to clinical understanding is limited when they address no new question, setting, or clinically relevant outcome. An absence of dangerous-response labels under one evaluation rubric should not be equated with established clinical safety.⁵³ Benchmarks should specify the claim and configuration, use independent representative cases, weight errors by clinical consequence, and publish prompts and reference standards.^{6,16,86}

Alongside targeted benchmarking, research should evaluate applications that address defined hearing-care needs. Earlier hearing-health commentaries proposed screening, education, triage, rehabilitation, clinician support, and research.^{10,11} High-value applications should address a defined bottleneck, require capabilities beyond a simpler method, support meaningful outcomes, and permit safe governance. Patient communication and care navigation supported by approved information could explain audiograms, troubleshoot devices, prepare patients for appointments, reinforce postoperative instructions, support tinnitus self-management, and escalate urgent symptoms. Retrieval can link responses to approved guidelines and device documents but depends on corpus and system quality.^{87,88} Evaluation should measure comprehension, action, trust calibration, continuity, and escalation rather than response accuracy alone.^{3,7,89}

Multimodal support could combine history, audiometry, speech testing, otoscopy or imaging, and device data to identify missing information and present evidence and uncertainty. Its role should be case preparation, structured interpretation, or decision support rather than autonomous diagnosis; information elicitation, sequencing, tool use, and recovery from incomplete data require direct testing.^{2,3,6} Assessment and rehabilitation applications include transcript scoring, multilingual test materials, synthetic speech, auditory training, and personalized rehabilitation.^{58,59,67,68} Evaluation should first establish linguistic or acoustic validity and measurement reliability, then test whether the application reduces editing burden, supports engagement, or improves listening and everyday communication, according to its intended purpose.

Professional and research tasks with defined inputs and outputs include documentation, coding, information extraction, cohort identification, and evidence synthesis. Outputs can be checked against source records, coding rules, or a documented analysis. These tasks may save time without delegating final judgment; outcomes should include correction burden, consequential errors, reproducibility, cost, and source traceability.^{88,90,91} Progress also requires interoperable audiology data for external validation.^{13,14} Rules, templates, signal-processing models, or conventional natural-language processing may be safer, more transparent, or computationally more efficient for some tasks; studies should justify LLM complexity and compare it with the best simpler alternative.^{6,92}

Table 3 links four evaluation questions to the gaps observed in this review: incomplete configuration reporting, limited validation on representative data, few prototype or target-user evaluations, and no prospective service or comparative impact evaluations. The questions address system identity, performance on representative data, safe use by intended users, and impact on care, workload, and cost. Aligned with the Developmental and Exploratory Clinical Investigations of Decision support systems driven by Artificial Intelligence (DECIDE-AI) guideline, it is an organizing aid rather than a validated maturity scale or reporting guideline.⁹² Evidence requirements should match the claim and risk.^{86,93}

Table 3. Four Questions for Developing Hearing-Care LLM Applications

These questions link observed evidence gaps to proposed evaluation requirements and are aligned with DECIDE-AI; they are not a replacement reporting guideline or validated maturity scale. The final column reports empirical analysis units, not independent participant samples. Reporting-item denominators include only applicable units; all other denominators are 95. Low-risk research tools may not require every step. DECIDE-AI indicates Developmental and Exploratory Clinical Investigations of Decision support systems driven by Artificial Intelligence; LLM, large language model.

Question	Aim	Minimum evidence	Decision before progression	Hearing-care evidence to date
1. What exactly is the system?	Define the intervention and intended claim	Intended users and setting; inputs and outputs; model version; prompts, retrieval, and tools; human oversight; prohibited uses; comparator; data flow, consent, retention, and deletion	The system is versioned, governed, and linked to a testable patient or service outcome	Specific variant/checkpoint: 68/95 (71.6%). Run/access date: 44/88 (50.0%). Full prompts: 52/88 (59.1%).
2. Does it work on representative data?	Establish validity beyond development examples	External data; reference standard; clinically weighted errors; abstention; relevant subgroups; privacy and security tests matched to the data and system, including prompt injection or leakage where relevant	Performance is adequate for limited target-user exposure and severe failure modes are understood	Clear reference standard: 74/93 (79.6%). Prespecified equity analysis: 0/95 (0%). Technical privacy testing: 0/95 (0%).
3. Can intended users use it safely?	Test comprehension, accessibility, and human–system interaction	Representative patients and professionals; simulation or silent prospective use; editing burden; automation bias; escalation; consent comprehension; privacy acceptability	Users can identify and manage errors, and workflow fit is acceptable	Prototype or target-user evaluation (M3): 14/95 (14.7%). These studies did not evaluate prospective service deployment.
4. What is its impact on care?	Assess clinical and service outcomes, workload, cost, and sustained performance	Prospective comparison with usual care or the best available alternative; patient and service outcomes; safety; workload; time; access; cost; equity; audit and incident processes; drift and rare-harm monitoring	Patient or service benefit is demonstrated with acceptable safety, equity, workload, and cost	Prospective service pilot (M4): 0/95 (0%). Comparative clinical/service impact evaluation (M5): 0/95 (0%).

The intervention includes users, setting, inputs, outputs, model version, prompts, retrieval, tools, oversight, prohibited uses, and data flows; material changes may require revalidation. External testing should emphasize clinically weighted errors, uncertainty, abstention, and diverse patients, professionals, and care settings rather than average accuracy alone. Human-factors testing should precede benefit claims. Simulation can assess comprehension, correction burden, automation bias, handoffs, and escalation; silent prospective testing can expose workflow and safety problems before outputs influence care. Comparative studies should then use function-matched outcomes and usual care or the best available alternative, including simpler or more transparent methods. These include comprehension and action for communication, correction burden and financial consequences for coding, safe clinician behavior for decision support, and engagement and functional

benefit for rehabilitation. Applicable trials should follow the artificial-intelligence extensions to SPIRIT and CONSORT.⁹⁴

Across these applications, accuracy alone cannot make a hearing-care system acceptable. Privacy concerns the full data path for notes, audiograms, device and implant records, genetic or neurophysiologic data, and audio, which may retain speaker identity beyond a transcript.⁹⁵ Governance should specify data provenance and rights, collection, deidentification, vendor access, retrieval stores, logs, retention, secondary use, and deletion. Data-management plans should state what can be deleted from each store, how deletion is verified, and any technical or legal limits, rather than promise unconditional erasure.^{96,97,98} Deployment architecture alters the risk profile but does not eliminate privacy risks. Local models may reduce transfer to external providers^{91,99} but still require access control, encryption, retrieval permissions, logging, retention rules, and maintenance. Hosted systems additionally require review of vendor processing, storage, reuse, and jurisdiction; no deployment label is a proxy for privacy. Evaluation should also consider reliance on external providers, continuity if access or terms change, and the feasibility of maintaining an alternative. Computational requirements and environmental impacts should be weighed against the benefit of the application.^{97,98}

Retrieval-augmented generation and agents enlarge the attack surface because retrieved documents, images, websites, or messages can redirect a model. Prompt injection has been demonstrated through medical images and conversations, and custom systems have leaked instructions and knowledge-base content.^{100,101,102} Security testing should be proportionate to the data, tools, and intended use. Systems handling patient records or external content need targeted tests of unauthorized retrieval, cross-patient disclosure, prompt injection, and unsafe tool use. Reports should describe the risks tested, safeguards, and remaining failures. Applicable data-protection requirements provide a baseline, but do not by themselves establish system security.^{97,98} Safety evaluation should prioritize severe omissions, false reassurance, failed escalation, automation bias, and recovery after intercepted errors.^{93,98} Equity should be prespecified across language, literacy, age, disability, communication mode, culture, device, and technology access. Accessibility may require text and audio options, captions, controllable pace, plain language, multilingual support, and assistive-technology compatibility.^{9,103} Studies should also examine how LLM use changes professional responsibilities, team communication, training needs, and accountability, including who checks outputs and manages errors.¹⁰⁴

Strengths include an LLM-specific search across biomedical, psychological, engineering, and computing databases; broad hearing-care coverage; duplicate human eligibility review and charting; explicit classification; and sensitivity analyses. Separating research role from implementation context prevented promising task outputs from being misread as clinical deployment. Supplementary Data 1 makes source-level classifications inspectable. Limitations include English-only eligibility, omission of standalone preprint repositories, variable conference detail, and possible missed ENT-wide reports. Search strategies were not peer reviewed by an information specialist. The 2022 start date may have omitted earlier work despite an informal preliminary search. The original ACM search fields were not retained, limiting comparability with the reconstructed update search. Eligibility of hybrid systems depended on published descriptions of the LLM contribution, which were sometimes incomplete. Mixed ENT studies were retained when hearing results were inseparable, with sensitivity analysis assessing their influence. Possible shared formative participants in the two psychosocial reports precluded assuming sample independence. No formal risk-of-bias assessment was performed. Differences in tasks, configurations, comparators, and outcomes precluded meta-analysis. Counts, percentages, and reporting

indicators describe the included literature rather than overall study quality or population parameters. The application priorities and development questions are interpretive proposals, not stakeholder-derived or validated frameworks.

The review identified 62 model-centered benchmark and 31 application-oriented empirical analysis units, alongside two adoption surveys. Applications extend beyond question answering to assessment, content generation, document transformation, biomedical evidence processing, psychosocial support, and research. However, no included investigation tested prospective hearing-service deployment or comparative clinical impact. Current evidence supports selected capabilities, not established improvements in care. Targeted benchmarking should be complemented by studies of defined applications with intended users, representative data, explicit comparators, and outcomes matched to their purpose. Progress should be judged by patient, professional, and service benefits alongside safety, privacy protection, equitable access, and acceptable workload and cost.

Methods

Design and Reporting

We conducted a scoping review guided by Joanna Briggs Institute (JBI) methodology and reported it with reference to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR).^{105,106} This design supported systematic identification and mapping of heterogeneous prompt-response, record-based, technical, user, survey, review, and commentary evidence. A written protocol, eligibility criteria, search record, and charting plan preceded screening, although the protocol was not registered; subsequent amendments were dated and documented.

Eligibility Criteria

Eligibility followed Population, Concept, and Context criteria. Populations included people receiving or seeking hearing-related care, professionals or trainees, and hearing-care service or research data. The concept was an identifiable LLM, LLM-powered system, or separable LLM component used in hearing care or a directly supporting workflow. Eligible hearing-care activities included patient information and counseling; prevention; screening and self-monitoring; triage and referral; audiological assessment and interpretation; medical or surgical otology and neurotology; tinnitus and hyperacusis; hearing aids and auditory implants; rehabilitation; teleaudiology; communication; documentation and coding; education; and directly related research support.^{9,12} Mixed or general ENT studies were eligible when the task or dataset contained an identifiable ear, auditory, otologic, or neurotologic component, even when item-level hearing results were not separable. Vestibular-only reports and general ENT studies without this connection were excluded. General sign-language translation, captioning, and other accessibility tools were outside the service-focused scope unless linked to hearing assessment, treatment, rehabilitation, or a related care task. Serving Deaf or hard-of-hearing users alone did not establish eligibility; neither sign-language modality nor use of computer vision was itself an exclusion criterion.

An LLM was central when it was the evaluated intervention; generated, transformed, scored, or analyzed a hearing-related input or output; formed an indispensable system component; or produced a separately evaluable

result. Reports were excluded when an LLM appeared only in background, future work, or assistance with writing the report, or when conventional machine learning, deep learning, signal processing, image recognition, or classical natural-language processing drove the study without a chartable LLM contribution. For hybrid systems, eligibility depended on the reported role of the LLM component, not simply the presence of machine learning or a transformer architecture. English-language sources published from 2022 onward were eligible. This start date reflected the limited uptake of LLMs before 2022 and an informal preliminary search that identified no earlier hearing-care LLM studies. Peer-reviewed journal reports and indexed conference sources were included; conference reports had to identify the objective, LLM, sample or input, core methods, at least one interpretable result, and sufficient hearing or otology context. Standalone preprint repositories were not searched, and preprints without an eligible published version were excluded. Reviews and commentaries were included as contextual evidence only when a hearing- or otology-specific LLM contribution could be charted. Broad ENT reviews without separable hearing-specific evidence were used only to contextualize the Introduction and Discussion.

Information Sources and Search

On July 6, 2026, we searched PubMed via NCBI, Embase, Web of Science Core Collection, IEEE Xplore, Scopus, APA PsycInfo, CINAHL, and ACM Digital Library. The searches were updated on September 19, 2026 for PubMed, Embase, IEEE Xplore, Scopus, APA PsycInfo, and CINAHL; September 20 for Web of Science; and September 21 for ACM. Database-specific strategies combined explicit LLM terms and major model families with broad hearing-care terms. Searches covered the eligible period from 2022 onward rather than only newly published records, allowing retrieval of newly indexed or previously missed reports. The Supplementary Information provides complete strategies, limits, and execution details. Citation searching identified one additional report, and targeted author-publication searching identified two hearing-health commentaries.

Source Selection and Data Charting

Database exports were converted to a common record format and deduplicated across databases and search rounds using standardized digital object identifiers, PubMed identifiers, and normalized titles, with manual review of possible matches. Reports retrieved in both search rounds were counted once in the selection flow. Two hearing-care professionals (S.W. and C.M.) separately screened titles and abstracts and assessed full texts using common criteria, resolving differences by discussion. One primary reason was recorded for each full-text exclusion. Two reviewers independently extracted and verified data against the source reports. Discrepancies were resolved through discussion to reach consensus. Charted variables covered publication characteristics; hearing-care scope; design; LLM task, input, output, and configuration; sample or evaluation unit; intended recipient; comparator or reference standard; human evaluators; outcomes; implementation context; and reporting of prompts, safety, privacy, communication, accessibility, language, and equity. Known duplicate reports were linked and represented once in empirical analyses.

Classification and Interpretive Assessment

We classified publication format, primary research role, LLM function, patient/recipient involvement, and implementation context separately. Primary roles distinguished patient information, professional knowledge, clinical reasoning, and multimodal benchmarks from task-specific outputs, domain prototypes, patient/recipient evaluation, research workflows, adoption surveys, reviews, and commentaries. Proposed future use did not convert a static benchmark into an application. Both reviewers checked and reconciled final classifications. Each empirical analysis unit received one implementation-context category: M0, observed use; M1, static benchmark; M2, task-specific or real-data evaluation outside routine care; M3, prototype or target-user evaluation; M4, prospective service pilot; or M5, comparative clinical or service impact evaluation. Reviews and commentaries were not assigned a category. These categories describe implementation context rather than study quality. For the application display, one primary function was assigned from the charted task and intended downstream use: clinical or administrative; research or education; rehabilitation or content; diagnostic or analytic; assessment automation; or speech and audio. Classification followed intended downstream use rather than model architecture. Full operational definitions appear in Supplementary Data 1; the six groups form Figure 2c.

No formal risk-of-bias assessment was performed, and no global quality score was assigned. Instead, we charted directly observable design limitations and source-level reporting features. Prompt, reference-standard, and human-evaluator items were first coded as applicable or not applicable and then as clearly or fully reported, partially reported, or not reported. Safety and error, privacy and governance, readability and communication, and accessibility and health literacy were coded as formally assessed, discussed only, not reported, or not applicable. Formal assessment required an explicit method or measure and a reported result. Privacy-related attitudes and perceived confidentiality were distinguished from technical tests of privacy protection when interpreting the broad governance code. Prespecified equity or fairness analysis was distinguished from descriptive group comparisons.

Analysis and Synthesis

Publication characteristics were summarized at report level. Research roles were summarized by report and, for empirical comparisons, by analysis unit. Implementation context and reporting indicators used empirical analysis units. The known duplicate conference/journal report pair was collapsed using its primary journal report. These units do not imply independent participant samples; possible formative-sample overlap between psychosocial reports remained qualified. Reviews and correspondence were excluded from empirical denominators. Counts and percentages summarized the included literature descriptively; no confidence intervals were calculated for these review-level proportions. Meta-analysis was inappropriate because tasks, configurations, comparators, reference standards, outcomes, and metrics were not comparable. Sensitivity analyses restricted evidence to hearing-specific reports and excluded conference-proceedings-only investigations, retaining journal reports of linked pairs. An additional descriptive analysis tabulated application-oriented units by publication year as counts and percentages of that year's empirical units, including adoption surveys. Separate checks used the earliest report year for the linked pair or excluded surveys. No inferential trend test was performed. Figure 2 shows reports by role, empirical analysis units by implementation context, and application functions by context; its counts were generated from the same source-level dataset as the tables.

Declarations

Data availability

Source-level characteristics and classifications of the included reports, the application-oriented evidence summary, and sensitivity-analysis summaries are provided in Supplementary Data 1. The database search strategies and PRISMA-ScR checklist are provided in the Supplementary Information.

Acknowledgements

This research received no funding.

Author contributions

S.W., C.M., F.Z., J.-W.A.W., and V.M. conceptualized the review. S.W. and C.M. developed the methodology, independently screened titles, abstracts, and full texts, independently extracted and verified data against the source reports, and wrote the original draft. S.W. conducted the literature searches and obtained the reports, curated the data, performed the formal analysis and validation, prepared the visualizations, and administered the project. J.-W.A.W. contributed to the interpretation of the findings. S.W., C.M., J.W., A.K.Y.S., F.Z., J.-W.A.W., D.W.S., and V.M. critically reviewed and edited the manuscript. S.W. and V.M. supervised the study. All authors approved the final manuscript and agreed to be accountable for their contributions and to help resolve questions concerning the accuracy or integrity of the work.

Competing interests

The authors declare no competing interests.

AI disclosure

OpenAI Codex assisted with English-language editing, search-strategy development, evidence summaries, analysis code, and preparation of tables and figures. S.W. reviewed all AI-assisted code and information and takes responsibility for their accuracy and integrity. The authors retain responsibility for the final manuscript.

Human Ethics and Consent to Participate declarations: not applicable.

References

1. Singhal K et al (2023) Large language models encode clinical knowledge. *Nature* 620:172–180.
<https://doi.org/10.1038/s41586-023-06291-2>
2. Mehandru N et al (2024) Evaluating large language models as agents in the clinic. *npj Digit Med* 7:84.
<https://doi.org/10.1038/s41746-024-01083-y>

3. Tu T et al (2025) Towards conversational diagnostic artificial intelligence. *Nature* 642:442–450. <https://doi.org/10.1038/s41586-025-08866-7>
4. Chen SF et al (2026) LLM-assisted systematic review of large language models in clinical medicine. *Nat Med* 32:1152–1159. <https://doi.org/10.1038/s41591-026-04229-5>
5. Bedi S et al (2025) Testing and evaluation of health care applications of large language models: a systematic review. *JAMA* 333:319–328. <https://doi.org/10.1001/jama.2024.21700>
6. Hager P et al (2024) Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med* 30:2613–2622. <https://doi.org/10.1038/s41591-024-03097-1>
7. Johri S et al (2025) An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat Med* 31:77–86. <https://doi.org/10.1038/s41591-024-03328-5>
8. Bedi S et al (2026) Holistic evaluation of large language models for medical tasks with MedHELM. *Nat Med* 32:943–951. <https://doi.org/10.1038/s41591-025-04151-2>
9. World Health Organization (2021) World report on hearing. World Health Organ <https://www.who.int/publications/i/item/9789240020481>
10. Swanepoel DW, Manchaiah V, Wasmann J-WA (2023) The rise of AI chatbots in hearing health care. *Hear J* 76:32. <https://doi.org/10.1097/01.HJ.0000927336.03567.3e>
11. Swanepoel DW, Manchaiah V (2024) The transformative potential of AI chatbots in hearing health care. *Hear J* 77(1):4. <https://doi.org/10.1097/01.HJ.0001025988.40308.e1>
12. Glista D, Schnittker JA, Brice S (2023) The modern hearing care landscape: toward the provision of personalized, dynamic, and adaptive care. *Semin Hear* 44:261–273. <https://doi.org/10.1055/s-0043-1769621>
13. Wasmann J-WA et al (2021) Computational audiology: new approaches to advance hearing health care in the digital age. *Ear Hear* 42:1499–1507. <https://doi.org/10.1097/AUD.0000000000001041>
14. Vercammen C et al (2026) Data standards in audiology: a mixed-methods exploration of community perspectives and implementation considerations. *Int J Audiol*. <https://doi.org/10.1080/14992027.2026.2619921>
15. Graham MA, Nicholson N, Glade R, Delport S, Mahomed-Asmail F (2026) Person-centered care in adult auditory rehabilitation: a scoping review. *Int J Audiol* 65:11–21. <https://doi.org/10.1080/14992027.2025.2499916>
16. Huo B et al (2025) Large language models for chatbot health advice studies: a systematic review. *JAMA Netw. Open* 8, e2457879 <https://doi.org/10.1001/jamanetworkopen.2024.57879>
17. Wang S, Mo C (2026) Evaluation of AI chatbots in hearing health: model timeliness and prompt design matter. *Int J Audiol* 65:992–995. <https://doi.org/10.1080/14992027.2025.2606193>
18. Jedrzejczak WW et al (2024) ChatGPT for tinnitus information and support: response accuracy and retest after three and six months. *Brain Sci* 14:465. <https://doi.org/10.3390/brainsci14050465>
19. Durgut O, Dikici OU (2026) Does ChatGPT update itself? Accuracy of ChatGPT in tympanostomy tube guidance: a comparative analysis with current literature. *Eur Arch Otorhinolaryngol* 283:117–130. <https://doi.org/10.1007/s00405-025-09630-3>
20. Hack S et al (2026) Guideline-integrated large language models improve decision support for acute ear, nose and throat emergencies. *J Emerg Med* 86:219–230. <https://doi.org/10.1016/j.jemermed.2026.04.018>

21. Frosolini A et al (2024) Artificial intelligence in audiology: a scoping review of current applications and future directions. *Sensors* 24:7126. <https://doi.org/10.3390/s24227126>
22. Koyama H, Kashio A, Yamasoba T (2024) Application of artificial intelligence in otology: past, present, and future. *J Clin Med* 13:7577. <https://doi.org/10.3390/jcm13247577>
23. Adcock K, Arulchelvan E, Shields N, Vanneste S (2026) Towards precision medicine for otology and neurotology: machine learning applications and challenges. *Hear Res* 469:109473. <https://doi.org/10.1016/j.heares.2025.109473>
24. Filali Ansary R, Lechien JR (2025) Clinical decision support using large language models in otolaryngology: a systematic review. *Eur Arch Otorhinolaryngol* 282:4325–4334. <https://doi.org/10.1007/s00405-025-09504-8>
25. Ahn J, Kang BG, Chang M, Yoon S (2025) Applications and future perspectives of large language models in otolaryngology-head and neck surgery: a comprehensive survey. *Clin Exp Otorhinolaryngol* 18:283–295. <https://doi.org/10.21053/ceo.2025-00121>
26. Alter IL, Chan K, Andreadis K, Rameau A (2026) Generative artificial intelligence methodology reporting in otolaryngology: a scoping review. *Laryngoscope* 136:1109–1117. <https://doi.org/10.1002/lary.70165>
27. Derbarsegian A, Gordon SA (2025) ChatGPT vs human clinical coders in determining appropriate CPT billing codes for otologic surgeries [Conference abstract]. *Otolaryngol. Head Neck Surg.* 172, P2 <https://doi.org/10.1002/ohn.983>
28. Derbarsegian A, Vesole AS, Sun DQ, Gordon SA (2026) Can ChatGPT replace human clinical coders? A comparative study in otology billing. *Ann Otol Rhinol Laryngol* 135:773–779. <https://doi.org/10.1177/00034894261452167>
29. Alter IL, Chan K, Lechien J, Rameau A (2024) An introduction to machine learning and generative artificial intelligence for otolaryngologists-head and neck surgeons: a narrative review. *Eur Arch Otorhinolaryngol* 281:2723–2731. <https://doi.org/10.1007/s00405-024-08512-4>
30. Kleebayoon A, Wiwanitkit V (2025) ChatGPT for responding to patient inquiries about otosclerosis: correspondence. *Eur Arch Otorhinolaryngol* 282:2785–2786. <https://doi.org/10.1007/s00405-025-09210-5>
31. Navarathna N, Kanhere A, Gomez C, Isaiah A (2025) Artificial intelligence in pediatric otolaryngology: a state-of-the-art review of opportunities and pitfalls. *Int J Pediatr Otorhinolaryngol* 194:112369. <https://doi.org/10.1016/j.ijporl.2025.112369>
32. Muradova A (2026) Artificial intelligence, real questions: ChatGPT and cochlear implants. *Eur Arch Otorhinolaryngol* 283:1169–1170. <https://doi.org/10.1007/s00405-025-09820-z>
33. Litsou E, Tzotzis I, Malisovas T, Asimakopoulos A, Psychogios G (2026) Artificial intelligence in otolaryngology: current applications, limitations, and future perspectives. *Eur Arch Otorhinolaryngol.* <https://doi.org/10.1007/s00405-026-10560-x>
34. Bosker HR (2025) Making manual scoring of typed transcripts a thing of the past: a commentary on Herrmann Speech Lang. *Hear.* 28, 2514395 (2025). <https://doi.org/10.1080/2050571X.2025.2514395>
35. Birinci M et al (2026) Large language models for cochlear implant education: a comparison of ChatGPT, Gemini, Claude, and DeepSeek. *Otolaryngol Head Neck Surg* 174:1339–1345. <https://doi.org/10.1002/ohn.70192>

36. Sivri I, Babaoglu E, Gul G, Colak T (2026) Comment on Large Language Models for Cochlear Implant Education: A Comparison of ChatGPT, Gemini, Claude, and DeepSeek. *Otolaryngol. Head Neck Surg.* <https://doi.org/10.1002/ohn.70384>
37. Birinci M et al (2026) Reply to Comment on Large Language Models for Cochlear Implant Education: A Comparison of ChatGPT, Gemini, Claude, and DeepSeek. *Otolaryngol. Head Neck Surg.* <https://doi.org/10.1002/ohn.70385>
38. Ajit-Roger E, Moise A, Peralta C, Orishchak O, Daniel SJ (2024) Enhancing multilingual patient education: ChatGPT's accuracy and readability for SSNHL queries in English and Spanish. *OTO Open* 8:e70048. <https://doi.org/10.1002/oto2.70048>
39. Hack S et al (2026) Quality of Tinnitus Information From Generative AI Systems and Web Search: An Expert-Rated Comparative Study. *Otol Neurotol* 47:1196–1204. <https://doi.org/10.1097/MAO.0000000000005015>
40. Ayo-Ajibola O, Park S, Osterbauer B, Kallemeyn K, Gomez G, MicrotiaGPT (2026) Appraisal of Microtia Care Recommendations From an Artificially Intelligent Chatbot. *Laryngoscope Investig Otolaryngol* 11:e70526. <https://doi.org/10.1002/lio2.70526>
41. Batuk IT, Karakuluk-Celebi I (2026) Evaluating the efficacy of artificial intelligence in audiology: a head-to-head comparison of ChatGPT and Gemini on hearing aid management. *Int J Med Inf* 220:106633. <https://doi.org/10.1016/j.ijmedinf.2026.106633>
42. Wang S et al (2024) Exploring the performance of ChatGPT-4 in the Taiwan audiologist qualification examination: preliminary observational study highlighting the potential of AI chatbots in hearing care. *JMIR Med Educ* 10:e55595. <https://doi.org/10.2196/55595>
43. Qi B, Zheng Y, Wang Y, Xu L (2025) Comparison of ChatGPT and DeepSeek on a standardized audiologist qualification examination in Chinese: observational study. *JMIR Form Res* 9:e79534. <https://doi.org/10.2196/79534>
44. Doğan Karataş T, Aksoy A, Bora A, Doğan M (2026) The use of artificial intelligence in medical education: a comparative analysis of theoretical exam performance between ENT residents and ChatGPT-4o. *Lokman Hekim Health Sci* 6:196–202. <https://doi.org/10.14744/lhhs.2024.94858>
45. Luo Z, Yuan C, Xie Q, Ananiadou S (2025) EMPEC: a comprehensive benchmark for evaluating large language models across diverse healthcare professions. *Find Association Comput Linguistics: ACL 2025* 9945–9958. <https://doi.org/10.18653/v1/2025.findings-acl.518>
46. Pastucha MG et al (2025) ChatGPT as a decision-support tool for better self-monitoring of hearing. *Am J Otolaryngol* 46:104711. <https://doi.org/10.1016/j.amjoto.2025.104711>
47. Caragli V et al (2026) Human versus AI in audiological practice: a comparative evaluation of ChatGPT and physician treatment decisions in idiopathic sudden sensorineural hearing loss. *PLoS ONE* 21:e0350549. <https://doi.org/10.1371/journal.pone.0350549>
48. Bilen YE, Eroğlu E, Atay Y, Sennaroğlu L (2026) ChatGPT vs. multidisciplinary team decisions in auditory implantation: a concordance analysis of implant type and side selection. *Eur Arch Otorhinolaryngol.* <https://doi.org/10.1007/s00405-026-10412-8>
49. Arnous S et al (2026) Comparative Evaluation of ChatGPT and Gemini in Approximating Clinically Confirmed Diagnoses From Structured Numerical Pure-Tone Audiogram Data. *Cureus* 18:e111153. <https://doi.org/10.7759/cureus.111153>

50. Liang J et al (2026) A multicenter multifunctional assessment of large language models in pure-tone audiogram interpretation for patients. *npj Digit Med* 9:348. <https://doi.org/10.1038/s41746-026-02537-1>
51. Karaçaylı C, Tahir E, Altuntaş EE (2026) Diagnostic interpretation of pure tone audiograms by multimodal LLMs: a comparative study of ChatGPT-5.0 and Gemini 2.5. *Eur Arch Otorhinolaryngol* 283:2227–2236. <https://doi.org/10.1007/s00405-025-09932-6>
52. Yenigun A et al (2026) Large language models for cholesteatoma diagnosis: a pathology-validated study. *J Craniofac Surg*. <https://doi.org/10.1097/SCS.00000000000013019>
53. Li L et al (2026) Evaluating Large Language Models in Clinical Audiology (AUDIOLOGYBENCH): Benchmark Development and Validation Study. *J Med Internet Res* 28:e94755. <https://doi.org/10.2196/94755>
54. Younan S et al (2026) Large Language Models Improve Operative Note Coding Accuracy and Financial Outcomes in Neurotology. *Laryngoscope* <https://doi.org/10.1002/lary.70777>
55. Kılıçtaş AU, Gül OU, Kılıçtaş BA, Kaba E, Erdivanlı B (2026) Can LLMs simplify operative notes? A comparative analysis in otorhinolaryngology. *Eur Arch Otorhinolaryngol* 283:477–489. <https://doi.org/10.1007/s00405-025-09758-2>
56. Gaeta L (2025) ChatAuD: the utility of generative artificial intelligence and large language models in clinical audiology communication. *Hear J* 78:1–3. <https://doi.org/10.1097/01.HJ.0001098600.18834.d1>
57. Washington CJ et al (2026) Leveraging Artificial Intelligence Chatbot to Generate Cochlear Implant Insurance Appeal Letter. *Ear Nose Throat J* 1455613261480757. <https://doi.org/10.1177/01455613261480757>
58. Herrmann B (2025) Leveraging natural language processing models to automate speech-intelligibility scoring. *Speech Lang Hear* 28:2374160. <https://doi.org/10.1080/2050571X.2024.2374160>
59. Herrmann B (2025) Language-agnostic, automated assessment of listeners' speech recall using large language models. *Trends Hear* 29:23312165251347131. <https://doi.org/10.1177/23312165251347131>
60. Tu T et al (2026) Genetic diagnosis and discovery enabled by large language models. *Adv Sci* 13:e18656. <https://doi.org/10.1002/advs.202518656>
61. Ye S, Wang L, Chen P (2026) OtoVCE: a mechanism-aware language-model evidence layer for hereditary hearing-loss variant interpretation. *Front Genet* 17:1880490. <https://doi.org/10.3389/fgene.2026.1880490>
62. Du Y et al (2026) TinnitusLLM: a multimodal large language model framework for tinnitus diagnosis through EEG-fMRI fusion learning. *IEEE J Biomed Health Inf*. <https://doi.org/10.1109/JBHI.2026.3670122>
63. Jeong Y, Song JJ, Yang J, Kang S (2024) Advancing tinnitus therapeutics: GPT-2 driven clustering analysis of cognitive behavioral therapy sessions and Google T5-based predictive modeling for THI score assessment. *IEEE Access* 12:52414–52427. <https://doi.org/10.1109/ACCESS.2024.3383020>
64. Liu D, Zheng Y (2025) Design of an edge computing otological disease diagnosis system based on deep learning and LLM. in *International Conference on Machine Learning, Neural Networks, and Computer Software (MLNCS 2025) Proc. SPIE* 13940, 1394016 <https://doi.org/10.1117/12.3092606>
65. Lee J et al (2025) Knowledge-Driven Feature Selection and Engineering for Genotype Data with Large Language Models. *AMIA Jt. Summits Transl. Sci. Proc.* 250–259 (2025). <https://pubmed.ncbi.nlm.nih.gov/40502214/>
66. Nguyen QM et al (2026) Interpretable fine-tuned large language models facilitate making genetic test decisions for rare diseases. *npj Digit Med* 9:608. <https://doi.org/10.1038/s41746-026-02733-z>

67. Zhang J, Zhang Z, Bai Y, Chen Y (2026) Perception of synthesized Mandarin speech based on a large-scale language model among Deaf adults with cochlear implants. *J Speech Lang Hear Res* 69:3977–3993. https://doi.org/10.1044/2026_JSLHR-25-00295
68. Fernando S et al (2024) Assisting hearing impaired children by personalized gamification of auditory verbal therapy. in. 6th International Conference on Advancements in Computing (ICAC) 414–419 (2024). <https://doi.org/10.1109/ICAC64487.2024.10850970>
69. Alabdulhussein A et al (2025) Artificial intelligence versus professional standards: a cross-sectional comparative study of GPT, Gemini, and ENT UK in delivering patient information on ENT conditions. *Diseases* 13:286. <https://doi.org/10.3390/diseases13090286>
70. Warriar A et al (2025) Readability of hospital online patient education materials across otolaryngology specialties. *Laryngoscope Investig Otolaryngol* 10:e70101. <https://doi.org/10.1002/lio2.70101>
71. Andrijašević A, Vukelić B (2024) Generating speech material for auditory training exercises using ChatGPT chatbot. in 2024 47th MIPRO ICT and Electronics Convention (MIPRO). 126–131. <https://doi.org/10.1109/MIPRO60963.2024.10569423>
72. Agarwal N, Bitragunta S (2025) Deep learning and adaptive large language model for speech denoising and performance enhancement. in IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) 1–6 (2025). <https://doi.org/10.1109/CONECCT65861.2025.11306888>
73. Bennett RJ, Tsiolkas J, Tagudin J (2026) Usability and desirability of a hearing health chatbot: an explorative study. *Int J Audiol* 65:233–243. <https://doi.org/10.1080/14992027.2025.2514586>
74. Choi Y, Moon J, Yoo J, Kim JG, Hong J-H (2026) Designing a generative AI-assisted music psychotherapy tool for Deaf and hard-of-hearing individuals. in Proceedings of the CHI Conference on Human Factors in Computing Systems Article 1412, 1–25 (2026). <https://doi.org/10.1145/3772318.3791385>
75. Choi Y et al (2026) From daily song to daily self: supporting emotional growth of Deaf and hard-of-hearing individuals through generative AI songwriting. in Proceedings of the CHI Conference on Human Factors in Computing Systems Article 1251, 1–26 (2026). <https://doi.org/10.1145/3772318.3791204>
76. Wickramasinghe N, Ulapane N, Wijewickrema S, Wisk D (2026) Feasibility Assessment of an Intelligent Agent to Assist Preserving Healthy Hearing. *Stud Health Technol Inf* 339:82–87. <https://doi.org/10.3233/SHTI260969>
77. Müller C et al (2025) Large language model-based prediction of speech intelligibility after Vibrant Soundbridge implantation using multidimensional outcome data: part 2 of a prospective study. *Sci Rep* 15:39564. <https://doi.org/10.1038/s41598-025-20919-5>
78. Wu T (2026) Delayed wh-question development in children with hearing loss: evidence for morphosyntactic vulnerability from corpus-based NLP and LLM analyses. in Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026) 291–304 <https://doi.org/10.18653/v1/2026.healing-1.26>
79. Shaari AL et al (2024) Exploring the ability of ChatGPT to act as a research aid in otolaryngology. *Indian J Otolaryngol Head Neck Surg* 76:6100–6105. <https://doi.org/10.1007/s12070-024-05101-z>
80. Lotto C et al (2024) ChatGPT generated otorhinolaryngology multiple-choice questions: quality, psychometric properties, and suitability for assessments. *OTO Open* 8:e70018. <https://doi.org/10.1002/oto2.70018>

81. Habib DRS et al (2026) Patient experiences in the cochlear implant Reddit community: comparing human and large language model categorization. *Am J Audiol* 35:487–496. https://doi.org/10.1044/2025_AJA-25-00216
82. Mo C, Manchaiah V, Wasmann JA, Wang S (2026) From policy debate to empirical evidence: a proof-of-concept bibliometric and LLM-assisted abstract-level content analysis of OTC hearing aid research before and after FDA regulation. *Int J Audiol* 1–16. <https://doi.org/10.1080/14992027.2026.2720841>
83. dos Santos A et al (2024) An experiment on an automated literature survey of data-driven speech enhancement methods. *Acta Acust* 8:2. <https://doi.org/10.1051/aacus/2023067>
84. Aggarwal K, Ravi R, Yerraguntla K (2025) Use of artificial intelligence tools by audiologists and speech-language therapists: an international survey of academicians. *J Otol* 20:20–25. <https://doi.org/10.26599/joto.2025.9540004>
85. Aggarwal K, Ravi R, Yerraguntla K (2025) Artificial intelligence and patient care: perspectives of audiologists and speech-language pathologists. *Intell Based Med* 11:100214. <https://doi.org/10.1016/j.ibmed.2025.100214>
86. Mese I et al (2026) Reporting checklist for foundation and large language models in medical research (REFINE): an international consensus guideline. *Diagn Interv Radiol*. <https://doi.org/10.4274/dir.2026.263812>
87. Liu S, McCoy AB, Wright A (2025) Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *J Am Med Inf Assoc* 32:605–615. <https://doi.org/10.1093/jamia/ocaf008>
88. Low YS et al (2025) Answering real-world clinical questions using large language model, retrieval-augmented generation, and agentic systems. *Digit Health* 11:20552076251348850. <https://doi.org/10.1177/20552076251348850>
89. Wan P et al (2024) Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. *Nat Med* 30:2878–2885. <https://doi.org/10.1038/s41591-024-03148-7>
90. Chen D, Alnassar SA, Avison KE, Huang RS, Raman S (2025) Large language model applications for health information extraction in oncology: scoping review. *JMIR Cancer* 11:e65984. <https://doi.org/10.2196/65984>
91. Wiest IC et al (2024) Privacy-preserving large language models for structured medical information retrieval. *npj Digit Med* 7:257. <https://doi.org/10.1038/s41746-024-01233-2>
92. Vasey B et al (2022) Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med* 28:924–933. <https://doi.org/10.1038/s41591-022-01772-9>
93. Lekadir K et al (2025) FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* 388. <https://doi.org/10.1136/bmj-2024-081554>. e081554
94. Ibrahim H et al (2021) Reporting guidelines for clinical trials of artificial intelligence interventions: the SPIRIT-AI and CONSORT-AI guidelines. *Trials* 22:11. <https://doi.org/10.1186/s13063-020-04951-6>
95. Ravi V, Wang J, Flint J, Alwan A (2024) Enhancing accuracy and privacy in speech-based depression detection through speaker disentanglement. *Comput Speech Lang* 86:101605. <https://doi.org/10.1016/j.csl.2023.101605>

96. Ong JCL et al (2024) Ethical and regulatory challenges of large language models in medicine. *Lancet Digit Health* 6:e428–e432. [https://doi.org/10.1016/S2589-7500\(24\)00061-X](https://doi.org/10.1016/S2589-7500(24)00061-X)
97. Autio C et al (2024) Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). National Institute of Standards and Technology <https://doi.org/10.6028/NIST.AI.600-1>
98. World Health Organization (2024) Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. *World Health Organ* <https://www.who.int/publications/i/item/9789240084759>
99. Vaid A et al (2024) Local large language models for privacy-preserving accelerated review of historic echocardiogram reports. *J Am Med Inf Assoc* 31:2097–2102. <https://doi.org/10.1093/jamia/ocae085>
100. Clusmann J et al (2025) Prompt injection attacks on vision language models in oncology. *Nat Commun* 16:1239. <https://doi.org/10.1038/s41467-024-55631-x>
101. Lee RW et al (2025) Vulnerability of large language models to prompt injection when providing medical advice. *JAMA Netw Open* 8:e2549963. <https://doi.org/10.1001/jamanetworkopen.2025.49963>
102. Truong VQ, Ritchie MD (2026) Prompt injection of OpenAI custom GPTs leaks informatics secrets. *AMIA Jt. Summits Transl. Sci. Proc.* 438–447 (2026)
103. Munson Klyn NA, Shaikh M, Z., Dhar S (2020) Health literacy and self-reported hearing aid use in the Health and Retirement Study. *Ear Hear* 41:386–394. <https://doi.org/10.1097/AUD.0000000000000770>
104. Wang S et al (2026) Perspectives of audiologists in China on artificial intelligence in clinical practice and professional identity: a qualitative study. *BMC Med Educ* 26:916. <https://doi.org/10.1186/s12909-026-09223-6>
105. Tricco AC et al (2018) PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med* 169:467–473. <https://doi.org/10.7326/M18-0850>
106. Peters MDJ et al (2020) Updated methodological guidance for the conduct of scoping reviews. *JB I Evid Synth* 18:2119–2126. <https://doi.org/10.11124/JBIES-20-00167>
107. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA (2020) PRISMA: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. *Campbell Syst. Rev.* 18, e1230 (2022). <https://doi.org/10.1002/cl2.1230>
108. Page MJ et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372:71. <https://doi.org/10.1136/bmj.n71>

Figures

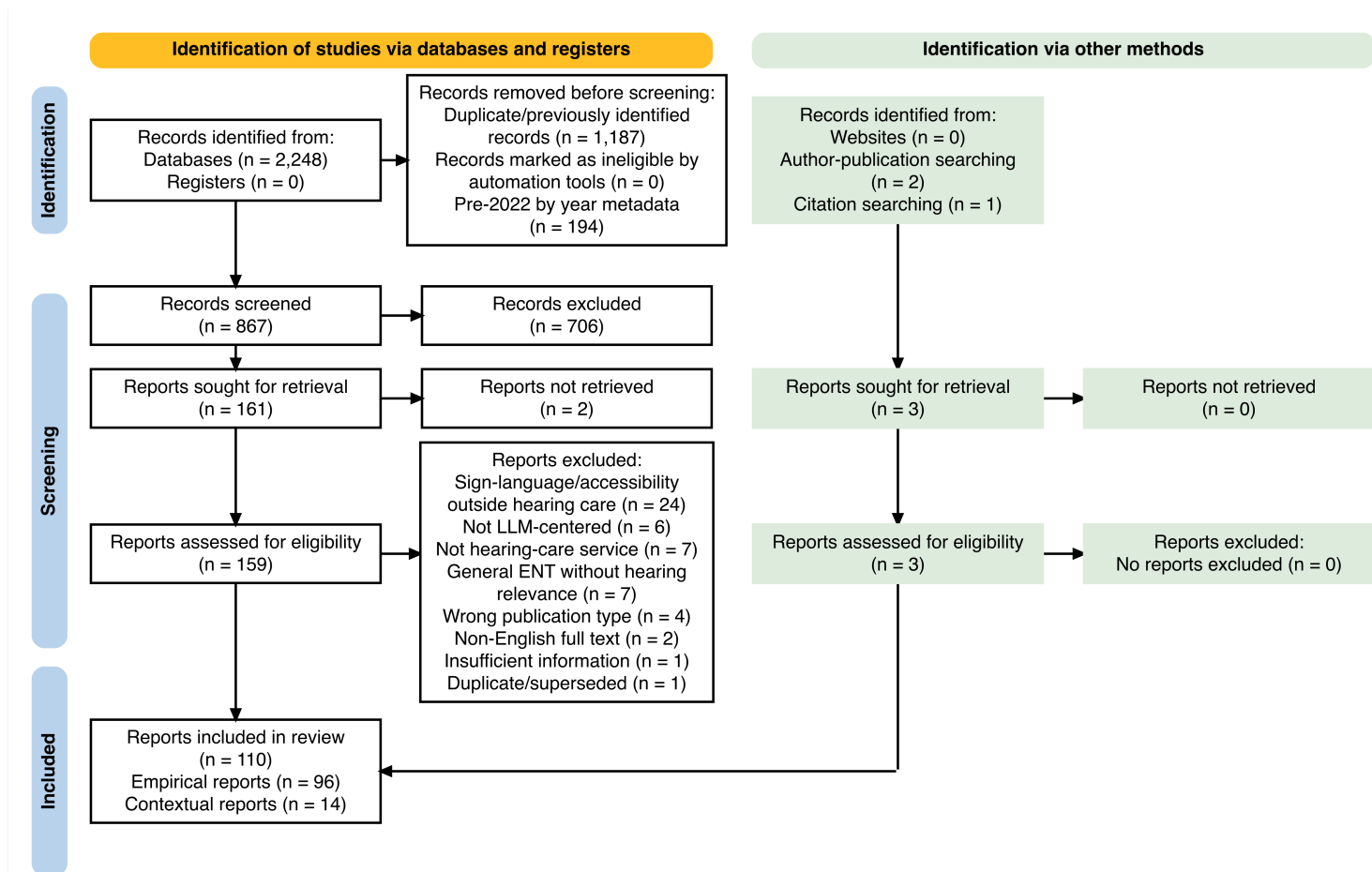


Figure 1

Cumulative PRISMA 2020 selection flow across the original and updated searches. Reports retrieved in both search rounds are counted once. Of 110 included reports, 96 were empirical and 14 were reviews or correspondence. The flow counts reports, not independent participant samples. Generated with the PRISMA2020 package¹⁰⁷ using the official PRISMA 2020 templates,¹⁰⁸ distributed under CC BY 4.0. LLM, large language model; ENT, ear, nose and throat.

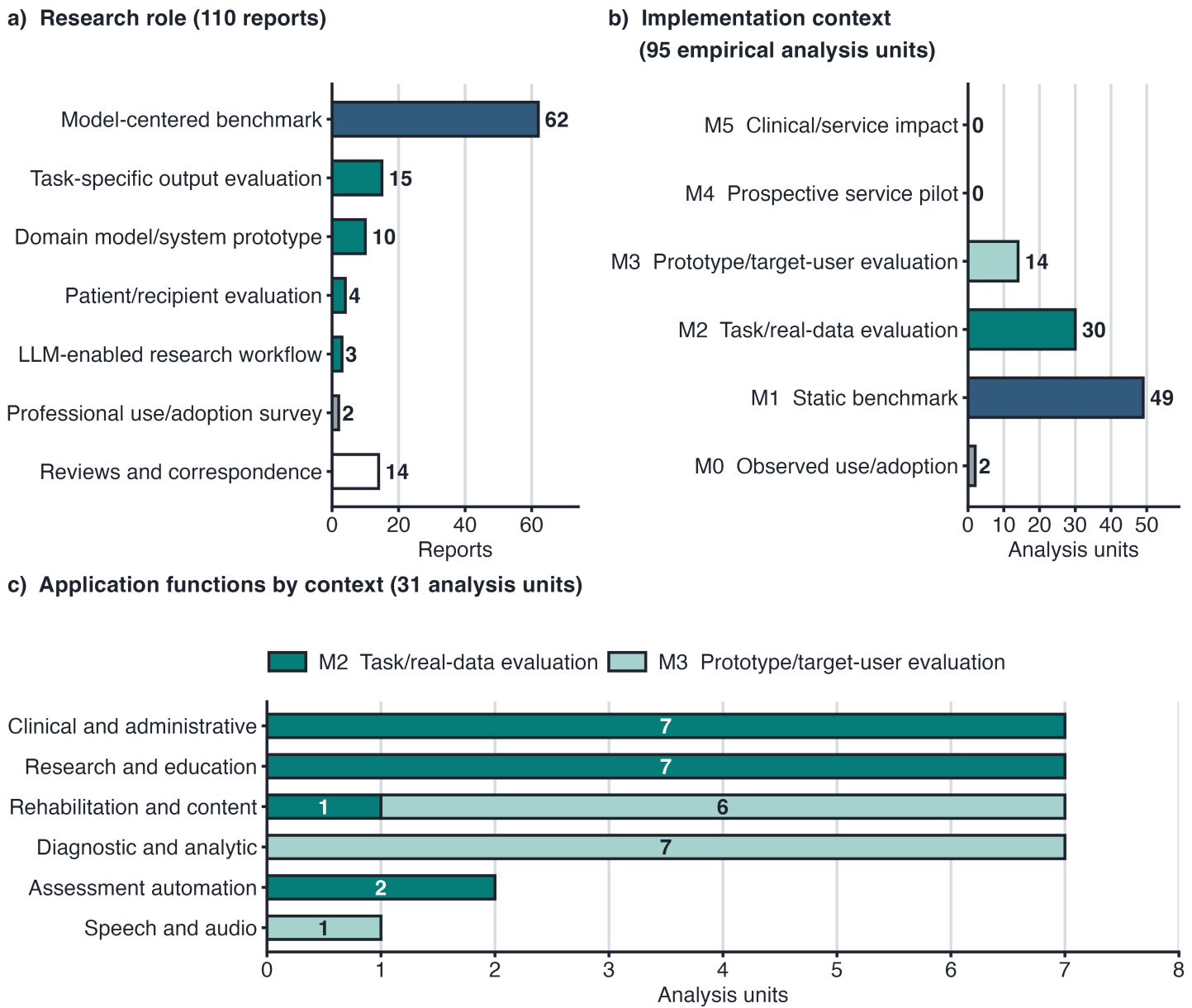


Figure 2

Research role, implementation context and application functions of LLM evidence in hearing care. Panel a classifies 110 reports. Panel b shows implementation context for 95 empirical analysis units: M0, 2; M1, 49; M2, 30; M3, 14; M4 and M5, 0. Panel c maps 31 application-oriented analysis units by primary downstream function and context; stacked segments indicate M2 and M3 counts. Linked duplicate reports are collapsed in panels b and c, but retained in panel a. Possible shared formative samples across psychosocial reports mean these units are not necessarily independent samples. No included study evaluated prospective service deployment or comparative clinical/service impact. LLM, large language model.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation.pdf](#)
- [SupplementaryData1.xlsx](#)