ARTIFICIAL INTELLIGENCE AND TEACHER MEDIATION IN ENGLISH LISTENING INSTRUCTION: AN INTEGRATIVE REVIEW

INTELIGENCIA ARTIFICIAL Y MEDIACIÓN DOCENTE EN LA ENSEÑANZA DE LA COMPRENSIÓN AUDITIVA EN INGLÉS: UNA REVISIÓN INTEGRATIVA

Melissa Dayana Loor Velez[1*]

[1] Docente de la Carrera Pedagogía de los Idiomas Nacionales y Extranjeros, Facultad de Ciencias Sociales, Humanísticas y de la Educación; Universidad Estatal del Sur de Manabí; Jipijapa, Ecuador. ORCID: https://orcid.org/0009-0008-5041-9324. Correo: melissa.loor@unesum.edu.ec

Katherine Elizabeth Intriago Palacios[2]

[2] Docente de la Carrera Pedagogía de los Idiomas Nacionales y Extranjeros, Facultad de Ciencias Sociales, Humanísticas y de la Educación; Universidad Estatal del Sur de Manabí; Jipijapa, Ecuador. ORCID: https://orcid.org/0009-0009-2505-7840. Correo: katherine.intriago@unesum.edu.ec

Cristina Beatriz Macías Mendoza[3]

[3] Docente de la Carrera Pedagogía de los Idiomas Nacionales y Extranjeros, Facultad de Ciencias Sociales, Humanísticas y de la Educación; Universidad Estatal del Sur de Manabí; Jipijapa, Ecuador. ORCID: https://orcid.org/0009-0004-2093-4027. Correo: cristina.macias@unesum.edu.ec

Nelly María Catagua González[4]

[4] Estudiante de la Carrera Pedagogía de los Idiomas Nacionales y Extranjeros, Facultad de Ciencias Sociales, Humanísticas y de la Educación; Universidad Estatal del Sur de Manabí; Jipijapa. Ecuador. ORCID: https://orcid.org/0009-0001-1288-8719. Correo: catagua-nelly7929@unesum.edu.ec

*** Autor para correspondencia: melissa.loor@unesum.edu.ec**

Abstract

The incorporation of artificial intelligence into English listening instruction created a need to examine not only the tools but also the pedagogical intervention that organized their use. This integrative review synthesized evidence on listening outcomes and teacher mediation. Twenty-four full-text files retrieved through structured searches and citation chaining through July 28, 2026, were examined. After two duplicate

files, one from each of two pairs, were removed, twenty-two unique reports were assessed: eleven listening-outcome studies formed the core corpus, four provided contextual evidence on mediation, five reviews supported interpretation, and two reports were excluded. Core-study quality was evaluated with design-specific criteria, and findings were integrated through thematic comparison. Several structured programs reported improvements in listening comprehension, aural decoding, or affective variables; however, two comparisons found no significant between-group listening advantage, and heterogeneity prevented attribution of the effects solely to technology. Five teacher functions were interpreted: orchestration, scaffolding, distribution of responsibilities, affective-relational support, and verification. Mediation appeared as a recurrent pedagogical condition, although its independent causal effect has not yet been established.

Keywords: *artificial intelligence; teacher mediation; English listening; integrative review; instructional scaffolding*

Resumen

La incorporación de la inteligencia artificial a la enseñanza de la comprensión auditiva en inglés planteó la necesidad de examinar no solo las herramientas, sino también la intervención pedagógica que organizó su uso. Esta revisión integrativa sintetizó evidencia sobre resultados auditivos y mediación docente. Se examinaron veinticuatro archivos de texto completo recuperados mediante búsquedas estructuradas y rastreo de citas hasta el 28 de julio de 2026. Tras retirar dos archivos duplicados, uno de cada uno de dos pares, se evaluaron veintidós informes únicos: once estudios de resultados auditivos conformaron el corpus principal, cuatro aportaron evidencia contextual sobre mediación, cinco revisiones sustentaron el marco interpretativo y dos informes fueron excluidos. La calidad de los estudios principales se valoró con criterios específicos para cada diseño y los hallazgos se integraron mediante comparación temática. Varios programas estructurados informaron mejoras en comprensión auditiva, decodificación o variables afectivas; sin embargo, dos comparaciones no mostraron una ventaja auditiva significativa entre grupos y la heterogeneidad impidió atribuir los efectos exclusivamente a la tecnología. Se interpretaron cinco funciones docentes: orquestación, andamiaje, distribución de responsabilidades, apoyo afectivo-relacional y verificación. La mediación apareció como una condición pedagógica recurrente, aunque su efecto causal todavía no fue probado de manera independiente.

Palabras clave: inteligencia artificial; mediación docente; comprensión auditiva en inglés; revisión integrativa; andamiaje pedagógico

Fecha de recibido: 12/05/2026
Fecha de aceptado: 20/07/2026
Fecha de publicado: 31/07/2026

Introducción

Listening is central to second and foreign language communication because learners must segment a continuous speech stream, recognize lexical forms, integrate contextual knowledge, and interpret a speaker's meaning in real time. These demands make listening especially difficult in English as a foreign language (EFL) setting, where exposure to varied accents, speech rates, and interactional conditions may be limited. Digital technologies have long expanded access to oral input, but contemporary artificial intelligence (AI) adds adaptive feedback, automated speech processing, conversational interaction, personalization, and content generation. Consequently, AI can alter not only what students hear, but also how listening practice is sequenced, monitored, and supported.

Existing reviews establish the relevance of this development while also revealing an important analytical problem. Zhang et al. (2023) reviewed 46 studies of technology-enhanced second-language listening and identified 13 technologies and 19 learning tasks. Their model showed that technology can support cognitive schemata, sustained attention, strategy use, practice, feedback, and learner emotion. However, technology-enhanced listening is broader than AI and does not by itself explain which pedagogical decisions make an AI-supported activity productive. Similarly, Crompton et al. (2024) synthesized 42 studies on AI in English language teaching and reported affordances across language skills, pedagogy, and self-regulation, alongside technical, ethical, and instructional challenges. These reviews demonstrate a rapidly expanding field, but their breadth limits detailed attention to the interaction between listening processes and teacher action.

The need for a focused interpretation is reinforced by the primary evidence. Some interventions reported substantial gains, but Tai and Chen (2024) found no significant post-test difference among multimodal, audio-only, and conventional groups, Yang et al. (2024) found no significant difference in listening gains between guided and less-guided groups, and Kim (2022) reported uneven results across computer and mobile delivery. The literature therefore does not justify a simple claim that artificial intelligence automatically improves listening. Variation in tools, learners, duration, outcomes, and instructional arrangements points to conditions that must be examined rather than averaged away.

Teacher mediation is one such condition. In this review, mediation refers to the pedagogical work through which a teacher selects or designs a tool, aligns it with a listening objective, models its use, structures interaction, supplies prompts and feedback, monitors learning, interprets automated information, supports motivation, and intervenes when the technology or the learner encounters difficulty. This definition is consistent with instrumental orchestration, where the value of a digital tool emerges through deliberately arranged teaching situations. Borges et al. (2026), although working outside language education, demonstrated that unguided AI use could produce a verification deficit and that explicit teacher scaffolding changed students from passive consumers of generated output into evaluators. This relational view is appropriate for language learning because it treats the tool, task, teacher, peers, and assessment as parts of one instructional arrangement rather than treating a chatbot as an isolated cause.

Mediation is also motivational and relational. Chiu et al. (2024) found that the relationship between chatbot use and student motivation depended on teacher support and learner expertise. Teacher support strengthened competence and relatedness, but could also reduce perceived autonomy when it became overly controlling. This tension is especially relevant to listening, where AI offers private repetition and immediate responses while teachers and peers provide explanation, emotional reassurance, shared attention, and contextual

judgment. Accordingly, teacher mediation should not be reduced to technical assistance; it includes instructional, cognitive, affective, social, and evaluative functions.

Despite this conceptual relevance, empirical studies often report an AI tool and a listening score without explicitly examining how teacher-designed sequences, briefings, worksheets, monitoring, or feedback contribute to the outcome. Conversely, studies that richly describe teacher mediation do not always measure listening comprehension directly. This separation makes it difficult for educators to determine whether AI substitutes for teaching, supplements it, or redistributes particular responsibilities.

The purpose of this integrative review was therefore to synthesize recent evidence on AI-supported English listening instruction through the lens of teacher mediation. Three questions guided the review: (1) What AI tools and instructional configurations have been investigated for English listening and aural decoding? (2) How is teacher mediation enacted before, during, and after AI-supported listening activities? and (3) What listening, engagement, and affective outcomes are reported, and which methodological conditions qualify those outcomes? By organizing evidence around mechanisms and instructional conditions rather than around tool novelty, the review seeks to offer a more useful account for classroom practice and future research.

Materials and methods

Review design

An integrative review design was used because the question required the combination of quantitative, qualitative, mixed-methods, and contextual evidence on both learning outcomes and teacher action. The procedure followed the stages proposed by Whittemore and Knafl (2005): problem identification, literature search, data evaluation, data analysis, and presentation. PRISMA 2020 principles informed transparent reporting of identification, screening, eligibility, and inclusion decisions (Page et al., 2021). The integrative designation therefore reflects the heterogeneous evidence and interpretive synthesis, not restricted access to particular databases.

Search strategy and information sources

Structured searches were updated on July 28, 2026, for publications dated from January 2022 through July 2026. ERIC, the Directory of Open Access Journals, SciELO, Crossref-linked metadata, and publisher or journal full-text platforms were searched. Latindex was used only to verify journal information, not as a bibliographic database. Backward and forward citation chaining was applied to the retrieved studies and recent reviews. The evidence workbook preserves the source, query family, retrieval route, duplicate decision, eligibility category, and exclusion reason for every full-text document in the audit trail.

The principal English query combined four concept groups: ("artificial intelligence" OR chatbot* OR "intelligent personal assistant*" OR "voice assistant*" OR "speech recognition" OR "generative AI") AND (EFL OR ESL OR "English as a foreign language" OR "second language") AND (listening OR "listening comprehension" OR "aural decoding") AND (instruction* OR teacher* OR mediation OR scaffold* OR feedback). A sensitivity query omitted the final mediation group so that studies describing teacher action only in the methods would not be missed. Spanish variants combined inteligencia artificial, comprensión auditiva,

enseñanza del inglés, and mediación docente. Flow counts refer to full-text files saved and screened, because unstable web result totals and subscription-only export counts were not treated as reproducible records.

Eligibility and study selection

Core studies were eligible when they: (a) involved learners of English as a foreign or second language in an educational setting; (b) used an identifiable AI-enabled tool or system; (c) assessed listening comprehension, aural decoding, or an explicit listening subskill; and (d) reported enough information about the instructional arrangement to distinguish technology from surrounding pedagogy. Studies were classified as contextual evidence when they directly examined teacher support, mediation, or human–AI division of work but did not measure the target listening outcome. Reviews were used for background and citation chaining, not as primary evidence for the themes. Records were excluded when they fell outside the date window, lacked the target construct, or contained contradictions that prevented a dependable interpretation.

Twenty-four full-text files were entered into screening. Two duplicate files—one copy of Borges et al. (2026) and one version of Alqarni and Alhramelah (2025)—were removed, leaving 22 unique reports. Eleven studies met the core listening-outcome criteria, four were retained as contextual mediation evidence, and five reviews informed background or citation chaining. Two reports were excluded: Dizon (2020) fell outside the prespecified date window, and Hazaymeh et al. (2025) contained incompatible descriptions of sampling, data sources, and intervention procedures. The integrative synthesis therefore used 15 empirical studies, whereas statements about listening outcomes were restricted to the 11 core studies.

Data extraction and quality appraisal

For each empirical record, data was extracted on country, educational level, sample, design, duration, AI tool, listening construct, teacher actions, principal findings, and methodological limitations. Duplicate files were checked before coding. Screening and extraction were completed in two sequential verification passes rather than by independent parallel reviewers. Teacher-mediation statements were coded only when an action could be located in the full text, such as selecting materials, recording explanations, sequencing activities, training learners, demonstrating interaction, monitoring progress, facilitating reflection, or responding to technical instability.

Methodological quality was appraised with the 2018 Mixed Methods Appraisal Tool (MMAT), using criteria appropriate to randomized, non-randomized, quantitative descriptive, qualitative, or mixed methods designs (Hong et al., 2018). Mixed-method studies were evaluated against three sets: qualitative criteria, the relevant quantitative category, and mixed-method integration criteria. The two screening questions and all applicable design-specific items were recorded as yes, no, or cannot tell with a full-text justification. No numerical total or universal confidence label was calculated; interpretation considered the weakest relevant component and retained null or contradictory findings.

Data synthesis

Analysis followed an iterative comparison process. Extracted teacher actions were first coded at the level of concrete practice. Related codes were then grouped into higher-order functions and compared against listening, engagement, and affective outcomes. Negative and equivocal cases were examined, including the non-significant between-group listening results in Tai and Chen (2024) and Yang et al. (2024), and the uneven

performance of mobile and computer delivery in Kim (2022). The five final functions are author-generated interpretive categories, not statistically tested moderators. Each was retained only when supported by at least two core or contextual studies; the evidence workbook records the study-to-theme links.

Results and discussion

Characteristics and methodological quality of the evidence

The 11 core studies were conducted in Jordan, China, Saudi Arabia, South Korea, Oman, and Taiwan. Samples ranged from 22 to nearly 500 learners and included upper-elementary pupils, adolescents, and university students. Designs included a randomized trial, non-randomized comparisons, and mixed-method investigations. Interventions lasted from six to sixteen weeks where duration was reported. The AI configurations included multi-tool ecosystems, chatbots, speech-recognition or text-to-speech systems, smart voice assistants, intelligent personal assistants, and adaptive computer- or mobile-assisted platforms.

Table 1: Characteristics and appraisal of the 11 core listening-outcome studies

Study / context	Design and sample	AI configuration and teacher mediation	Main listening finding	Quality considerations
Bataineh et al. (2026)	Quasi-experimental; n = 48; 6 weeks	Chatbots, LingQ, speech-to-text and VR. Teacher sequenced pre-, while-, and post-listening scaffolds and monitored fidelity.	Higher adjusted post-test scores for the AI group; large instructional advantage.	Strong measurement and fidelity; intact classes and bundled tools limit causal isolation.
Huang et al. (2024)	Mixed methods; main study n = 89; 12 weeks	Purpose-built chatbot with teacher-recorded explanations. Teacher-designed content was retained while reminders and repeated support shifted to the chatbot.	Chatbot and flipped groups achieved comparable decoding outcomes; chatbot group showed stronger engagement and less teacher dependence.	Integrated evidence; intact classes and embedded teacher videos complicate attribution to automation.
Zhou et al. (2025)	Mixed-method quasi-experiment; n = 67; 16 weeks	ChatGPT-4 for informal digital listening. Instructor introduced the tool, assigned practice, monitored participation, and curated resources.	Substantial listening gains and personalized practice were reported for the experimental class.	Different intact classes/majors and informal exposure leave residual confounding.
Xiao (2025)	Randomized controlled trial; n = 84; 8 weeks plus follow-up	AI speech technology embedded in a matched curriculum. Instructors facilitated use and monitored self-study logs.	Sustained listening gains, increased flow, and reduced anxiety in the AI group.	Random allocation and follow-up are strengths; assessor blinding was unreported and the technology description conflates TTS and ASR.

Study / context	Design and sample	AI configuration and teacher mediation	Main listening finding	Quality considerations
Alqarni and Alhramelah (2025)	Experimental comparison; n = 22	Researcher-designed smart voice assistant structured through planning, implementation, evaluation, and adaptable lesson design.	Experimental students improved on listening and speaking measures.	Very small sample; randomization and confounder control are unclear; skill outcomes are bundled.
Alrasheedi (2024)	Quasi-experimental pre/post; n = 100	Chatbot and Duolingo supported by a teacher manual, repeated feedback, and reset opportunities.	Experimental students improved overall and across listening subskills.	Allocation wording is inconsistent; confounder adjustment and fidelity reporting are limited.
Kim (2022)	Non-randomized class comparison; methods n = 486	AI TOEIC platform delivered by computer or mobile phone; teacher guided scheduled classes and taught the control group.	All groups improved; AI-CALL outperformed teacher-led control, whereas AI-MALL was not consistently superior.	Arbitrary section assignment, repeated testing, and a 496/486 sample discrepancy reduce confidence.
Yang et al. (2024)	Comparative intervention; n = 34; 6 weeks	Google Assistant used at home. All learners received weekly commands; one group also completed teacher-designed evaluation forms.	The guided group improved over time, but listening gains did not differ significantly between groups.	Small sample and short duration; teacher guidance was bundled with additional reflective activity.
Abdellatif et al. (2024)	Quasi-experimental; n = 57; 8 lessons	Both groups used Duolingo and the same listening course; the experimental group completed Nearpod-based listening assessments.	Adjusted post-test results favored the AI-assessment condition for listening and affective outcomes.	Convenience, all-male sample; limited technical validation of the claimed AI assessment component.
Soliman et al. (2025)	Quasi-experimental; n = 64; 8 weeks	Text Blaze, Loom, and Otter.ai were integrated into a teacher-organized program targeting four listening subskills.	Post-test performance favored the experimental group across the reported listening subskills.	Allocation and confounder control are insufficiently detailed; reporting relies heavily on gain ratios.
Tai and Chen (2024)	Mixed methods; n = 92; 10 weeks	Google Nest Hub or Mini. Researcher-teacher trained learners, demonstrated interaction styles, briefed each task, provided worksheets, and offered help.	Only the multimodal Hub group improved significantly over time; post-test differences among groups were not significant.	Baseline equivalence and triangulation are strengths; intact classes and within-group emphasis limit causal inference.

Note. ASR = automatic speech recognition; CALL = computer-assisted language learning; MALL = mobile-assisted language learning; TTS = text-to-speech. Quality judgments are criterion-based MMAT interpretations; no summed score was calculated.

The MMAT appraisal did not support treating the evidence as methodologically uniform. The strongest designs combined clear outcome measurement with fidelity monitoring, baseline adjustment, or integrated qualitative evidence. Nevertheless, even the stronger studies often used intact classes or bundled teacher-designed materials with the AI system, making it difficult to isolate a purely technological effect. Confidence was lower where samples were very small, group allocation was unclear, confounders were not addressed, or internal reporting was inconsistent. These limitations do not erase the studies' pedagogical value, but they restrict claims of general effectiveness.

Four contextual studies broadened the interpretation of mediation. Sadykova and Kayumova (2025) examined AI-generated images and audio with 40 children aged four to six. Listening comprehension was not measured directly; however, classroom video, teacher journals, interviews, and vocabulary tests showed teachers orchestrating interaction with, through, and around AI while counteracting unstable outputs. Lee (2025) compared AI assistants, an AI learning application, peer interaction, and textbook instruction with 201 seventh-grade students. The study focused on engagement rather than listening achievement, but it demonstrated that AI offered repetition, humor, and immediate feedback, whereas peers supplied socioemotional connection and collaborative support that the tools did not reproduce. Chiu et al. (2024) examined how teacher support conditioned motivation during chatbot learning, and Borges et al. (2026) documented how explicit teacher scaffolding shifted university students from accepting generated output to evaluating it. These studies were used only to interpret mediation processes, not to infer listening achievement.

Thematic synthesis of teacher mediation

Five interrelated functions explain how teacher mediation operated across the corpus. They are analytically distinct but often occurred together within the same lesson. Their overlap is important: an intervention that appears to be an AI effect may also be an effect of sequencing, task design, teacher-created explanations, monitoring, or peer organization.

Table 2: Teacher-mediation functions identified in the integrative synthesis.

Mediation function	Evidence pattern	Representative evidence	Interpretive implication
Instructional orchestration	Teachers selected tools and materials, aligned tasks with listening objectives, sequenced phases, and monitored implementation.	Bataineh et al. (2026); Huang et al. (2024); Soliman et al. (2025); Tai & Chen (2024)	AI value depended on pedagogical architecture rather than tool presence alone.
Scaffolding and feedback	Teacher or teacher-designed AI supplied modeling, prompts, explanations, corrective feedback, strategy support, and repeated practice.	Huang et al. (2024); Zhou et al. (2025); Xiao (2025); Abdellatif et al. (2024)	Immediate feedback was useful when it remained connected to explicit learning goals and interpretive support.

Mediation function	Evidence pattern	Representative evidence	Interpretive implication
Division of labor and guided autonomy	Routine repetition, reminders, personalization, and access shifted toward AI, while teachers retained design, diagnosis, and interpretation.	Huang et al. (2024); Zhou et al. (2025); Kim (2022); Yang et al. (2024)	AI redistributed responsibility; it did not eliminate teacher expertise.
Affective-relational mediation	Teachers supported confidence, motivation, anxiety management, relatedness, safe participation, and productive peer interaction.	Xiao (2025); Sadykova & Kayumova (2025); Lee (2025); Chiu et al. (2024)	Private AI practice may reduce pressure, but human relationships remain central to sustained engagement.
Verification and risk management	Teachers identified breakdowns, unstable content, superficial feedback, recognition errors, and inappropriate reliance.	Tai & Chen (2024); Sadykova & Kayumova (2025); Borges et al. (2026)	Human judgment remained indispensable for accuracy, contingency management, and critical use.

Instructional orchestration: from tool selection to listening sequence

The most consistent pattern was that AI was embedded in an instructional architecture. Bataineh et al. (2026) did not expose students to a single application; the teacher coordinated chatbots, gamification, transcription, and virtual scenarios through pre-listening, while-listening, and post-listening scaffolds. The substantial adjusted difference therefore represents a multi-tool pedagogical design rather than an isolated app effect. This distinction matters for replication: adopting the same platforms without the sequencing, monitoring, and fidelity checks would not reproduce the intervention.

Tai and Chen (2024) provide a parallel example with a more differentiated outcome. Both AI groups received training, demonstrations, short briefings, worksheets, and access to teacher help, but only the group using a multimodal Google Nest Hub improved significantly over time. The audio-only group and conventional group did not. Moreover, the post-test difference among the three groups was not statistically significant. The finding suggests that teacher orchestration can create favorable conditions, but the representational affordances of the tool and the design of the comparison still influence what learners gain.

The smaller study by Alqarni and Alhramelah (2025) also embedded a smart voice assistant in a planning, implementation, and evaluation process. Its positive results are promising but should not be generalized from 22 students. Across these studies, the teacher's first contribution was not real-time explanation; it was the upstream design work that determined which listening problem the AI would address, how difficulty would progress, and what evidence of learning would be collected.

Soliman et al. (2025) similarly organized Text Blaze, Loom, and Otter.ai within an eight-week program targeting main ideas, details, frequent words, and discourse questions. Reported gains favored the experimental group, but allocation and confounder control were insufficiently described. Abdellatif et al. (2024) held the listening content and Duolingo practice constant while varying the assessment condition through Nearpod. Although adjusted outcomes favored the experimental condition, the study did not technically demonstrate which Nearpod functions were artificially intelligent. Both studies therefore support

structured technology-enhanced listening more strongly than they support an isolated artificial-intelligence effect.

Scaffolding and feedback: immediacy is not sufficient

AI systems supplied functions that are difficult to provide continuously in conventional classes: immediate responses, repeated trials, individualized pacing, and feedback outside scheduled contact time. Zhou et al. (2025) used ChatGPT-4 to extend listening practice into informal digital learning, while the instructor introduced the tool, assigned tasks, monitored participation, and contrasted chatbot resources with curated materials. Listening gains were substantial, but the intact-class design and different student majors make it impossible to attribute the entire effect to chatbot personalization.

Huang et al. (2024) make the relationship between automation and teacher expertise particularly visible. Their chatbot-supported course retained teacher-selected content and prerecorded teacher explanations while shifting reminders, repetitive support, and portions of feedback to a 24-hour conversational agent. The chatbot group achieved decoding outcomes comparable to a teacher-intensive flipped course and demonstrated stronger behavioral and emotional engagement. The result is not evidence that the teacher became unnecessary. It shows that teacher knowledge can be designed into a resource, allowing automation to handle high-frequency support while the teacher concentrates on decisions that require professional judgment.

The same caution applies to Xiao (2025). The randomized design and follow-up support confidence that the intervention was associated with sustained listening, flow, and anxiety benefits. Yet the description alternates between speech recognition and text-to-speech, and feedback conditions differed between groups. These ambiguities illustrate why AI terminology must be reported precisely. Teachers and researchers need to specify what the system receives, what it generates, how feedback is triggered, and how learners are expected to act on it.

Division of labor and guided autonomy

The corpus does not support a replacement narrative. It supports a redistribution-of-labor narrative. AI was most suited to repetition, access, reminders, rapid feedback, and the adaptation of routine practice. Teachers remained responsible for defining outcomes, selecting content, diagnosing misconceptions, interpreting progress, and deciding when the automated response was insufficient. Huang et al. (2024) and Yang et al. (2024) provide especially clear examples because automation carried routine prompts or practice while teacher-designed explanations or reflection continued to organize learning.

Kim (2022) complicates assumptions about anytime-anywhere learning. The computer-based AI-CALL group outperformed teacher-led instruction in listening and reading, but the mobile AI-MALL group did not consistently outperform the computer group. Portability alone did not guarantee better learning. Scheduled teacher guidance, screen characteristics, task engagement, or implementation conditions may have shaped the result. Yang et al. (2024) likewise found no significant difference in listening gains between the group receiving weekly teacher-designed evaluation forms and the comparison group. These null or uneven findings show that additional guidance or mobility may expand and structure practice without necessarily changing the assessed listening construct.

Guided autonomy therefore requires a balance. Students need room to control pace, repeat input, and explore personalized content, but they also need teacher-set goals, monitoring, and reflection that keep activity connected to listening development. Excessive control can undermine autonomy, as Chiu et al. (2024) observed, whereas insufficient guidance can encourage superficial use. Effective mediation calibrates support rather than maximizing it.

Affective and relational mediation

Listening performance is partly shaped by emotional conditions. AI can create a private practice space, permit repetition without embarrassment, and offer predictable access to feedback. Xiao (2025) reported reduced anxiety and greater flow alongside listening gains, while Huang et al. (2024) found stronger emotional and behavioral engagement in the chatbot-supported condition. However, these benefits should not be interpreted as evidence that automated interaction can replace human relatedness.

Lee (2025) found that AI assistants and applications supported engagement through humor, gamification, repetition, and immediate feedback, whereas peer interaction produced collaborative support and socioemotional connection that the technologies did not replicate. Sadykova and Kayumova (2025) similarly showed teachers capitalizing on children's enjoyment while maintaining a safe and productive classroom around unpredictable generative tools. Teacher mediation thus connects affect to pedagogy: the teacher notices anxiety or confusion, organizes peer support, adjusts demands, and determines whether novelty is serving a learning purpose.

Verification, limitations, and heterogeneous outcomes

The synthesis also identified verification and contingency management as underreported teacher functions. AI systems may misunderstand accented speech, return unstable content, provide superficial explanations, or encourage learners to accept fluent output uncritically. Borges et al. (2026) showed that teacher mediation can transform a verification deficit into evaluative judgment. In listening classrooms, the analogous work includes checking transcripts, validating feedback, explaining recognition errors, distinguishing language difficulty from system failure, and preventing the tool from becoming the sole authority.

Outcome heterogeneity confirms the need for this work. The multi-tool Jordanian intervention produced a large, adjusted advantage, the Chinese chatbot and randomized speech-technology studies reported substantial gains, and the multimodal Google Assistant group improved over time. Yet Tai and Chen found no significant post-test difference among groups, Yang et al. found no significant between-group difference in listening gains, and Kim found delivery-mode differences. The defensible conclusion is therefore descriptive and conditional: several structured AI-supported arrangements were associated with better listening outcomes, but current designs do not establish teacher mediation as an independent causal moderator.

Implications for teaching and research

For practice, AI-supported listening should be planned as a sequence rather than an add-on. A teacher can establish a listening objective, select or generate level-appropriate input, prepare pre-listening orientation, define the AI's role during practice, and follow automated activity with interpretation, reflection, or transfer. Feedback should be checked for accuracy and linked to a strategy the learner can apply. Multimodal support should be used deliberately, since visual responses may aid comprehension but can also shift the task away

from listening if they reveal too much information. Finally, AI practice should be combined with peer and teacher interaction to preserve relational support and authentic negotiation of meaning.

For research, intervention reports should separate the AI component from the surrounding pedagogy. Authors should identify the system and version, describe input and output modalities, report teacher training and implementation fidelity, document prompts or task sequences, and specify whether feedback was automated, teacher-produced, or hybrid. Larger multi-site studies should compare different levels of mediation while holding content and practice time constant. Mixed-methods designs are especially valuable when qualitative evidence explains why the same tool produces gains, null effects, or unequal engagement across learners.

Limitations of the review

This review has five limitations. First, identification counts represent saved full-text records rather than all unstable results displayed by web search interfaces, and subscription-level Scopus or Web of Science exports were unavailable; the corpus is therefore a structured integrative sample, not a complete census. Second, screening and extraction were verified sequentially rather than by independent parallel reviewers. Third, the core corpus was geographically concentrated in Asian and Middle Eastern settings; no Latin American listening-outcome study was found among the screened records. Fourth, teacher mediation was usually reconstructed from methods because it was rarely measured as an independent construct. Fifth, outcomes ranged from aural decoding and classroom tests to standardized listening components, limiting direct comparison. These constraints require an interpretive synthesis and preclude causal claims about mediation.

Conclusions

The evidence supports a qualified conclusion. Several structured artificial-intelligence-supported programs reported better English listening, aural-decoding, engagement, or affective outcomes, but effects were not uniform and two studies found no significant between-group listening advantage. Because most interventions bundle technology with task design, teacher guidance, content, and practice time, the observed outcomes cannot be attributed exclusively to artificial intelligence.

Teacher mediation was interpreted through five functions: instructional orchestration, scaffolding and feedback, division of labor and guided autonomy, affective-relational support, and verification and risk management. These categories describe recurrent practices; they do not demonstrate a statistically tested mediation effect. The corpus nevertheless indicates that routine repetition, reminders, and some feedback can be delegated, while diagnosis, curricular alignment, interpretation, ethical judgment, and human connection remain professional responsibilities.

For educational practice, the central question is therefore not whether AI or the teacher should control listening instruction. The more productive question is which responsibilities can be assigned to each, at what stage of the task, and with what safeguards. For research, stronger causal evidence will require larger and more diverse samples, precise descriptions of AI functionality, explicit measurement of teacher mediation, fidelity evidence, and comparisons that separate technological affordances from instructional design. Until such comparisons are available, teacher mediation should be treated as a plausible and pedagogically important condition, not as a proven causal moderator.

References

Abdellatif, M. S., Alshehri, M. A., Alshehri, H. A., Hafez, W. E., Gafar, M. G., & Lamouchi, A. (2024). I am all ears: Listening exams with AI and its traces on foreign language learners' mindsets, self-competence, resilience, and listening improvement. Language Testing in Asia, 14, Article 54. https://doi.org/10.1186/s40468-024-00329-6

Alqarni, A., & Alhramelah, A. (2025). Designing a smart voice assistant based on artificial intelligence to develop listening and speaking skills in English for elementary school students. Educational Process: International Journal, 17(1), Article e2025403. https://doi.org/10.22521/edupij.2025.17.403

Alrasheedi, S. (2024). The effect of using AI applications to develop EFL listening comprehension skills among university students. Conhecimento & Diversidade, 16(44), 601–637. https://doi.org/10.18316/rcd.v16i44.12346

Bataineh, R. F., Obeiah, S. F., & Bataineh, R. F. (2026). Tuning in with technology: AI-enhanced listening instruction in the Jordanian EFL classroom. Journal of Information Technology Education: Innovations in Practice, 25, Article 7. https://doi.org/10.28945/5699

Borges, I., Caridade, C. M. R., Sebastião, C., & Pereira, V. (2026). From ChatGPT to classroom learning: Exploring the role of teacher mediation in AI-supported education. Australasian Journal of Educational Technology, 42(2), 116–137. https://doi.org/10.14742/ajet.11056

Chiu, T. K. F., Moorhouse, B. L., Chai, C. S., & Ismailov, M. (2024). Teacher support and student motivation to learn with artificial intelligence (AI) based chatbot. Interactive Learning Environments, 32(7), 3240–3256. https://doi.org/10.1080/10494820.2023.2172044

Crompton, H., Edmett, A., Ichaporia, N., & Burke, D. (2024). AI and English language teaching: Affordances and challenges. British Journal of Educational Technology, 55(6), 2503–2529. https://doi.org/10.1111/bjet.13460

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2018). Mixed Methods Appraisal Tool (MMAT), version 2018: User guide. McGill University. https://mixedmethodsappraisaltoolpublic.pbworks.com/

Huang, W., Jia, C., Hew, K. F., & Guo, J. (2024). Using chatbots to support EFL listening decoding skills in a fully online environment. Language Learning & Technology, 28(2), 62–90. https://doi.org/10.64152/10125/73572

Kim, N.-Y. (2022). AI-integrated mobile-assisted language learning: Is it an effective way of preparing for the TOEIC test in classroom environments? English Teaching, 77(3), 79–102. https://doi.org/10.15858/engtea.77.3.202209.79

Lee, J. (2025). EFL adolescents' engagement in artificial intelligence and peer interaction. Language Learning & Technology, 29(1), 1–21. https://doi.org/10.64152/10125/73643

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, Article n71. https://doi.org/10.1136/bmj.n71

Sadykova, G., & Kayumova, A. (2025). AI-powered image and audio generators for very young EFL learners. Iranian Journal of Language Teaching Research, 13(2), 1–24. https://doi.org/10.30466/ijltr.2025.55564.2821

Soliman, S. A., Aladini, A., Jalambo, M., & Abdelkarim, R. (2025). The effectiveness of artificial intelligence tools in enhancing listening skills among Omani EFL 12th graders: A quasi-experimental study. Educational Sciences Journal, 33(1), 1–19. https://doi.org/10.21608/ssj.2025.424588

Tai, T.-Y., & Chen, H. H.-J. (2024). The impact of intelligent personal assistants on adolescent EFL learners' listening comprehension. Computer Assisted Language Learning, 37(3), 433–460. https://doi.org/10.1080/09588221.2022.2040536

Whittemore, R., & Knafl, K. (2005). The integrative review: Updated methodology. Journal of Advanced Nursing, 52(5), 546–553. https://doi.org/10.1111/j.1365-2648.2005.03621.x

Xiao, Y. (2025). The impact of AI-driven speech recognition on EFL listening comprehension, flow experience, and anxiety: A randomized controlled trial. Humanities and Social Sciences Communications, 12, Article 425. https://doi.org/10.1057/s41599-025-04672-8

Yang, C. T. Y., Lai, S. L., & Chen, H. H. J. (2024). The impact of intelligent personal assistants on learners' autonomous learning of second language listening and speaking. Interactive Learning Environments, 32(5), 2175–2195. https://doi.org/10.1080/10494820.2022.2141266

Zhang, R., Zou, D., & Cheng, G. (2023). A systematic review of technology-enhanced L2 listening development since 2000. Language Learning & Technology, 27(3), 41–64. https://doi.org/10.64152/10125/73531

Zhou, Q., Hashim, H., & Sulaiman, N. A. (2025). Integrating AI chatbots in informal digital English learning: Impacts on listening competencies in Chinese higher education. Education and Information Technologies, 30(18), 27031–27059. https://doi.org/10.1007/s10639-025-13811-2