Journal of Information Technology Education: Innovations in Practice

An Official Publication of the Informing Science Institute
InformingScience.org

JITEiip.org

Volume 25, 2026

# TUNING IN WITH TECHNOLOGY: AI-ENHANCED LISTENING INSTRUCTION IN THE JORDANIAN EFL CLASSROOM

| | | |
|---|---|---|
| Ruba Fahmi Bataineh* | Yarmouk University, Irbid, Jordan | rubab@yu.edu.jo |
| Salameh Fleih Obeiah | Ministry of Education, Mafraq, Jordan | Loujein_salama@yahoo.com |
| Rula Fahmi Bataineh | Jordan University of Science and Technology, Irbid, Jordan | rula@just.edu.jo |

* Corresponding author

## ABSTRACT

| | |
|---|---|
| Aim/Purpose | To evaluate whether a coordinated, multi-tool AI instructional design, chatbots, LingQ gamification, Google Speech-to-Text, and AI-driven virtual reality improve listening comprehension for Jordanian ninth-grade learners. |
| Background | Listening is underdeveloped across many EFL primary and lower-secondary classrooms in the MENA region, where classrooms rarely sustain theory-informed, technology-rich scaffolding; this study responds by pairing Vandergrift's metacognitive model and Vygotsky's sociocultural lens with practicable AI tools. |
| Methodology | A quasi-experimental comparison of two intact ninth-grade sections (experimental n = 24; control n = 24) contrasted AI-enhanced lessons with textbook activities; analyses used ANCOVA to control for pre-test differences while lesson logs and platform analytics monitored fidelity. |
| Contribution | The study provides experimental evidence that an integrated, multi-tool AI design can produce meaningful listening gains and models how affective, cognitive, motivational, and situational scaffolds function together in classroom practice. |
| Findings | The AI-enhanced group demonstrated higher adjusted post-test listening scores than the control group after controlling for pre-test performance, F(1, 45) = |

Accepting Editor Stamatis Papadakis | Received: October 24, 2025 | Revised: December 12, December 18, 2025; January 8. 2026 | Accepted: January 13, 2026.
Cite as: Bataineh, Ruba F., Obeiah, S. F., & Bataineh, Rula 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

| | |
|---|---|
| | 42.39, $p < .001$, partial $\eta^2 = .49$. Adjusted means favored the experimental group ($M = 16.56$, $SE = .29$) over the control group ($M = 13.94$, $SE = .29$), indicating a substantial advantage for learners receiving coordinated, multi-tool AI instruction. These gains were observed within the B1 range of the assessment. |
| Recommendations for Practitioners | Sequence and blend multiple affordable AI tools (e.g., chatbots, LingQ, automated transcription, short VR scenarios) in short in-class rotations, and use platform analytics to tailor difficulty and feedback rather than relying on a single app or textbook. |
| Recommendations for Researchers | Replicate and scale the design across multiple schools and larger samples; use mixed methods to isolate which tool components drive gains; examine moderators (proficiency, digital literacy, teacher training); and, when possible, apply formal standard-setting to link outcomes to CEFR levels. |
| Impact on Society | Effective, cost-sensitive AI scaffolding can broaden access to international media, higher education, and employment for secondary learners in under-resourced MENA settings, thereby supporting educational inclusion and social mobility. |
| Future Research | Pursue longitudinal, multi-site trials to test durability and transfer to spontaneous spoken interaction, compare single-tool versus multi-tool architectures, evaluate teacher professional development needs, and assess cost-effectiveness across diverse Jordanian and regional contexts. |
| Keywords | AI tools, EFL listening, multi-tool integration, metacognitive strategies, sociocultural theory |

## INTRODUCTION AND BACKGROUND

Listening comprehension is widely acknowledged as a foundational skill in foreign language learning, underpinning the development of speaking, reading, and writing abilities. It is not merely a passive reception of sound but a dynamic, cognitively demanding process that requires the integration of bottom-up decoding of linguistic forms with top-down activation of prior knowledge, inferencing, and contextual interpretation (Buck, 2001; Richards, 2008; Vandergrift & Goh, 2012). Yet, in many EFL contexts, listening instruction has historically been marginalized, often due to limited exposure to authentic spoken English, insufficient training in active listening strategies, and the persistence of grammar-translation and teacher-centered pedagogies (Nowrouzi et al., 2015; Osada, 2004).

In the Arab world, these challenges are compounded by contextual constraints, as learners frequently lack access to authentic listening materials that reflect natural speech patterns in speed, accent, and register (Hwaider, 2017). Instead, classroom listening practice often revolves around scripted textbook dialogues and discrete-item comprehension questions, which prioritize recall over holistic comprehension and strategic listening. While the traditional three-stage sequence of pre-listening, while-listening, and post-listening (Davies & Pearse, 2000; Lindsay & Knight, 2006) offers a clear structure, it is often applied in ways that neither adapt to learners' needs nor provide timely feedback. More recent attempts to modernize listening pedagogy have incorporated podcasts, YouTube videos, and extensive listening activities to create more authentic experiences (e.g., Karkera & Chamundeshawari, 2018). However, these approaches still frequently lack interactivity, adaptive scaffolding, and sustained learner motivation (e.g., Ordoñez Procel et al., 2024).

Over the past decade, advances in artificial intelligence (AI) have introduced tools that have the potential to address longstanding gaps in EFL listening instruction. AI-powered chatbots can simulate authentic conversations; gamified platforms, such as LingQ, offer interactive tasks with immediate feedback; speech-to-text applications, such as Google Speech-to-Text, facilitate pronunciation and

comprehension analysis; and immersive AI-driven virtual reality environments can recreate real-world communicative contexts (Dong et al., 2024; Panagiotidis, 2025; Tolstykh & Oshchepkova, 2024; Wiboolyasarin et al., 2025). These technologies align with learner-centered and constructivist approaches, enabling adaptive, multimodal, and context-rich learning while helping to mitigate persistent challenges, such as performance anxiety, limited speaking opportunities, and lack of personalized support (Cooray et al., 2024; Fathi et al., 2024; Kim & Su, 2024; Wu, 2024).

However, in tandem with these advancements, researchers caution that overreliance on technology may reduce opportunities for teacher-mediated interaction, raise ethical and privacy concerns, and exacerbate inequities if access is uneven (Al-Zahrani, 2024; Davar et al., 2025; Farooqi et al., 2024; Hınız, 2026; Hussein et al., 2025; Umoke et al., 2025; Vesna et al., 2025). These concerns underscore the need for balanced integration strategies that position technology as a complement rather than a substitute for pedagogical expertise, ensuring that digital innovations enhance rather than diminish the quality, equity, and human dimensions of language education. Nonetheless, a growing body of empirical research highlights the pedagogical potential of AI-driven tools in EFL listening instruction when applied strategically and aligned with established learning frameworks.

Empirical evidence suggests that these tools can act as both cognitive scaffolds, by supporting prediction, monitoring, and problem-solving strategies central to Vandergrift's (2004) metacognitive model, and affective enablers that reduce anxiety and foster learner autonomy (Kukulska-Hulme, 2020; Miao et al., 2021). Speech recognition technologies have been shown to improve fluency and self-regulation (Abdalkader, 2023), while conversational AI offers interactive, low-stakes listening practice (Baxramova, 2025), albeit with noted limitations in prosodic authenticity (Suvorov, 2022). Gamified and mobile-assisted AI platforms sustain engagement by offering adaptive challenges and real-time feedback (Kukulska-Hulme, 2020).

Evidence from the MENA region not only reinforces these findings but also underscores structural barriers. In Jordan, AI-supported listening activities were reported to improve young learners' comprehension (Al-mawaly & AL-Jamal, 2022) and enhance university students' motivation and performance (Hazaymeh et al., 2025), yet rural–urban disparities in access persist. In Saudi Arabia, chatbot-based listening training was reported to yield significant comprehension gains (Alrasheedi, 2024) while self-regulated dynamic assessment approaches improved both listening performance and learner autonomy (Abdolrezapour & Ghanbari, 2021). Across contexts, however, infrastructural constraints, uneven digital literacy, and limited teacher preparedness seem to be persistent obstacles (AlAli & Wardat, 2024; Alshahrani & Qureshi, 2024), underscoring the need for sustained professional development, targeted infrastructure investment, and policy interventions that ensure equitable and effective integration of AI in EFL listening instruction.

Despite the growing body of research on AI in language education, much of it has focused on reading, writing, or speaking skills (Fathi et al., 2024; Hidayatullah, 2024; Li et al., 2024), with relatively few studies examining the effect of AI on listening comprehension (Goh & Aryadoust, 2025). Studies in the Arab region remain scarce, and those that exist typically examine single-tool interventions rather than integrated, multi-tool designs (e.g., Al-Barakat et al., 2025; Bataineh & Al-Ghareeb, 2025; Zghoul & Bataineh, 2024). Hence, the potential of a coordinated, strategy-aligned AI approach to listening instruction in primary-stage EFL contexts remains underexplored.

Despite mounting evidence for the value of AI in language learning, existing research rarely tests multi-tool, strategy-driven designs in primary-stage EFL contexts, particularly in the Arab region. Where studies do exist, they tend to focus on isolated tools, short-term gains, or tertiary education, leaving unanswered questions about scalability, contextual fit, and integrated pedagogical design. This study addresses these gaps by embedding four complementary AI modalities (viz., chatbots, the LingQ gamification platform, Google Speech-to-Text, and AI-driven VR) within a unified, metacognitively grounded listening cycle tailored for Jordanian ninth-grade students. By explicitly aligning each tool with a core listening strategy and situating practice in authentic, varied communicative

contexts, the research aims to test whether a coordinated AI approach can yield statistically significant performance gains and/or deeper shifts in learner confidence, strategic behavior, and engagement with real-world spoken English.

## *Conceptual Framework*

This study is anchored in sociocultural theory, specifically drawing on Vygotsky's (1978) concept of the Zone of Proximal Development, which highlights the distance between what learners can achieve independently and what they can accomplish with guidance from a more knowledgeable other, whether that be a teacher, a peer, or an AI-mediated system. In the context of listening instruction, AI tools can serve as scaffolding agents, offering timely prompts, corrective feedback, and adaptive tasks that extend learners' communicative abilities.

The study also draws on Vandergrift's (2004) metacognitive model of listening comprehension, which foregrounds three interrelated strategies that guide effective listening. The first is *prediction*, whereby learners anticipate content before listening by drawing on contextual clues, such as topic, setting, or speaker identity. The second is *monitoring*, which involves actively checking comprehension during listening and adjusting strategies in real time to maintain understanding. The third is *problem-solving*, in which learners address comprehension gaps after listening through clarification, selective re-listening, or the targeted application of specific strategies. Together, these processes foster greater learner autonomy, strategic awareness, and resilience in navigating authentic spoken language.

The AI tools used in this study (viz., chatbots, the LingQ gamification platform, Google Speech-to-Text, and AI-driven VR) operationalize these theoretical constructs as follows:

1. AI-powered chatbots facilitate prediction by introducing topic-specific vocabulary, eliciting prior knowledge, and prompting pre-listening discussions, hence preparing learners within their ZPD.
2. LingQ gamification supports monitoring by embedding interactive comprehension checks, vocabulary tasks, and adaptive feedback loops that enable learners to track understanding in real time.
3. Google Speech-to-Text aids problem-solving by allowing learners to compare AI-generated transcripts with authentic scripts, identify discrepancies, and refine listening strategies.
4. AI-driven virtual reality simulations extend scaffolding into immersive, task-based environments, enabling learners to apply comprehension strategies in authentic communicative contexts that closely mirror real-world demands.

By aligning these technological affordances with sociocultural and metacognitive principles, the instructional design seeks to bridge the gap between textbook-based listening activities and the complex, unpredictable nature of authentic communication. In this study, AI-powered tools served as scaffolds, systematically extending learners' listening abilities. Vandergrift's metacognitive model highlights prediction, monitoring, and problem-solving as three interrelated strategies that enable effective listening. These strategies were explicitly operationalized through AI tools and their effectiveness was measured through learners' performance scores.

Figure 1 presents the current conceptual research model, which integrates AI-powered tools within the EFL listening instruction framework. The model aligns with Vygotsky's sociocultural theory and Vandergrift's metacognitive model of listening comprehension, illustrating how specific AI tools operationalize prediction, monitoring, and problem-solving strategies to scaffold learner engagement and improve listening outcomes.

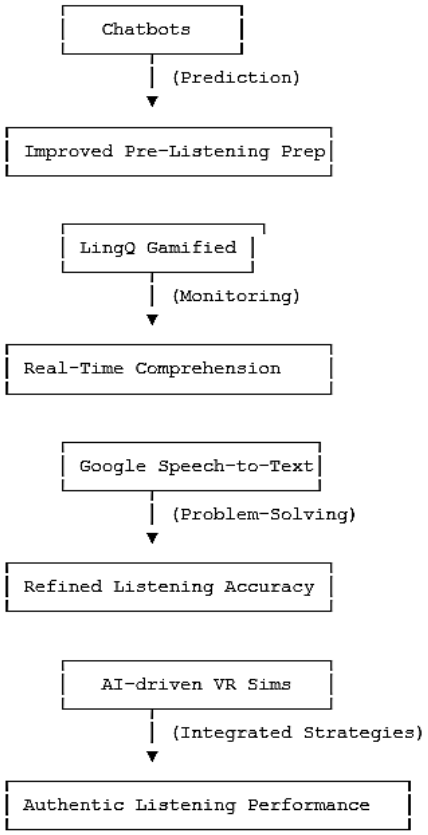


**Figure 1**. **Conceptual research model of AI-enhanced EFL listening instruction**

Figure 1 depicts the theoretical and procedural foundations of the study. It illustrates how each AI tool directly operationalizes a metacognitive strategy and contributes to measurable listening comprehension outcomes. The flowchart layout emphasizes the progression from specific AI tools to strategies and quantifiable performance gains, keeping the model visually clear while anchored in a quantitative orientation.

The instructional design deliberately aligned each AI tool with a distinct metacognitive strategy to enhance learners' engagement and comprehension. AI-powered chatbots were employed to facilitate prediction by introducing contextual cues and key vocabulary before listening activities. The gamified features of LingQ reinforced monitoring through adaptive comprehension checks and continuous vocabulary practice. Google Speech-to-Text supported problem-solving by allowing learners to compare their listening with generated transcripts and refine their understanding accordingly. Finally, AI-driven VR simulations extended the application of these strategies into immersive, authentic communicative contexts, thereby consolidating skills in environments that mirror real-world demands.

In this study, Vygotskian sociocultural mediation and Vandergrift's metacognitive model operate as complementary rather than competing frameworks. AI tools function both as social mediators that extend learners' participation in guided, scaffolded interactions and as cognitive supports that activate prediction, monitoring, and problem-solving strategies. Chatbots and VR simulations approximate Vygotsky's conception of assisted performance within the ZPD, while the metacognitive routines embedded in LingQ and Speech-to-Text enable learners to regulate comprehension independently. By connecting socially mediated guidance with strategy-driven cognitive work, the instructional design establishes a hybrid model in which interactional scaffolding and individual strategic behavior reinforce one another.

### *Purpose of the Study*

Grounded in the sociocultural and metacognitive perspectives outlined above, this study aims to investigate the combined effect of a multi-tool AI design (viz., chatbots, the LingQ gamification platform, Google Speech-to-Text, and AI-driven VR) on Jordanian EFL ninth-grade listening instruction. While listening comprehension is a critical, albeit generally neglected, skill in local and regional EFL contexts, there remains limited empirical evidence on how AI technologies can be systematically used to enhance young learner performance.

More specifically, the study examines the combined effect of AI-powered chatbots, the LingQ gamification platform, Google Speech-to-Text, and AI-driven VR simulations on learners' listening comprehension. These tools were selected because each operationalizes core aspects of scaffolding and metacognitive strategy use: chatbots for prediction, gamification for monitoring, speech-to-text for post-listening problem-solving, and virtual reality for authentic contextual application.

Even though AI tools have been recognized for their potential to enrich language instruction, much of the existing research in the MENA region has examined them as standalone interventions. This study departs from that trend by integrating four AI modalities (viz., chatbots, the LingQ gamification platform, Google Speech-to-Text, and AI-driven VR) into a single instructional cycle aligned with Vandergrift's (2004) metacognitive model. By explicitly mapping each tool to a distinct listening strategy, prediction, monitoring, or problem-solving, and embedding authentic speech conditions into VR tasks, the design operationalizes strategy use in ways not previously tested in regional primary EFL contexts. This multi-tool, strategy-driven approach aims to generate both measurable performance gains and deeper learner engagement in authentic listening tasks.

## METHODOLOGY

### *Research Design*

This study adopted a quasi-experimental, pre-test/post-test control-group design to examine the combined effect of a multi-tool AI design on the listening comprehension of Jordanian ninth-grade EFL learners. While this design enabled a meaningful comparison between an experimental group receiving AI-enhanced instruction and a control group taught through conventional textbook-based listening tasks, the use of intact ninth-grade sections rather than randomly assigning participants represents a notable limitation. Random allocation was not feasible due to administrative policies in the public school system, where class rosters are set at the beginning of the academic year and cannot be reorganized mid-term without disrupting both instructional continuity and institutional regulations. This arrangement raises the possibility that pre-existing differences between the two sections (e.g., prior exposure to technology, general academic ability, levels of learner motivation) could affect the findings. To minimize these potential confounds, both groups were taught by the same EFL teacher, using parallel lesson objectives, identical time allocations, and closely matched activity sequences. Pre-test scores were statistically controlled using ANCOVA to account for any initial differences in listening ability.

### *Participants*

The participants comprised 48 ninth-grade students enrolled in a public school in Irbid, Jordan. Two intact sections were assigned as the experimental group (n = 24) and the control group (n = 24). Both groups were taught by the same teacher, who received targeted training on the integration of the AI tools to ensure consistent delivery. The training emphasized the sequencing of activities, equal distribution of instructional time, and adherence to a structured lesson plan.

### *Instruments*

Listening comprehension was assessed using a researcher-developed parallel pre-test/post-test aligned with the B1 level of the Common European Framework of Reference for Languages (CEFR,

Council of Europe, 2020). The test comprised 20 points in total: multiple-choice items (4 × 2 points = 8), short-answer items (2 × 3 points = 6), and fill-in-the-blank items (2 × 3 points = 6). Items were drawn from authentic spoken sources and adapted to reflect natural speech features (varied accents, natural pacing, and limited background noise). Two EFL assessment specialists reviewed the materials for content validity, and a pilot with a separate group of 20 ninth-grade students (excluded from the main sample) produced acceptable internal consistency (Cronbach's $\alpha = 0.78$ for the pre-test and $\alpha = 0.82$ for the post-test). Scoring procedures were designed to be transparent and consistent: partial credit was awarded for partially correct short-answer and fill-in responses, and minor orthographic errors that did not change meaning were accepted. To safeguard scoring reliability, two experienced EFL raters independently scored a 30% random subsample of scripts; interrater agreement was high (ICC[2,1] = 0.91 for the pre-test; ICC[2,1] = 0.88 for the post-test; Cohen's $\varkappa = 0.86$), indicating substantial-to-excellent agreement under standard reporting conventions.

To make the alignment with CEFR transparent and reproducible, each item was independently mapped to CEFR listening descriptors by two CEFR-trained raters (both senior EFL assessment specialists). Mapping used an analytic checklist linking each item to one or more illustrative CEFR B1 descriptors (e.g., listening for gist, specific information, and attitude) drawn from the CEFR Companion Volume; raters recorded the target descriptor, the observable response expected, and a provisional difficulty judgement for each item. Discrepancies were resolved through a calibration meeting and documented in an item-by-item mapping log. The mapping process followed the operational alignment procedures recommended for CEFR-based test development.

Item-level statistics (item difficulty as proportion correct, item discrimination as corrected item–total point-biserial) and distractor-function analyses for the multiple-choice items were calculated during the piloting phase. Aggregate reliability was estimated with Cronbach's α (reported above for the pilot) and McDonald's ω with 95% bootstrap confidence intervals. Items with discrimination values of $r_{pb} < 0.20$ or with non-functioning distractors were revised or omitted prior to the main data collection.

For short-answer and fill-in items, a three-point rubric was used: 2 = fully correct response (target meaning and acceptable lexical form), 1 = partial credit (partial comprehension, missing non-essential detail, or minor orthographic error not altering meaning), 0 = incorrect or no response. Rater training consisted of: (a) a 90-minute calibration workshop reviewing rubric anchors and worked examples, (b) independent scoring of a 15% training set followed by discrepancy meetings to align interpretations, and (c) final double-scoring of 30% of scripts for interrater-reliability estimation (ICC and $\varkappa$ reported above). These procedures follow recommended practice for rater training and reliability reporting.

## *Instructional Materials and Fidelity of Implementation*

To safeguard instructional fidelity, the teacher maintained detailed lesson logs of the activities conducted, the AI tools used, the duration of each segment, and observations of learner engagement. Additionally, usage analytics were retrieved from the LingQ platform and Google Speech-to-Text, including completion rates, time-on-task, and accuracy scores. These logs and analytics were reviewed weekly by the research team to confirm that the planned integration of AI tools was implemented as intended. The control group's sessions were also documented to ensure that no AI tools were inadvertently introduced.

## *VR Task Design and Examples*

The AI-driven virtual reality component was incorporated once a week for the experimental group, providing immersive simulations of communicative contexts with authentic speech rates, diverse accents (e.g., British, American, Australian), and occasional background noise to mirror real-world conditions.

The VR component incorporated a series of interactive role-play tasks designed to simulate authentic communicative scenarios. These included hotel booking (reserving rooms, confirming details, and handling special requests), airport check-in (interacting with airline staff to check in, confirm flight details, and address baggage issues), restaurant ordering and complaint resolution (placing orders, clarifying menu items, and lodging polite complaints about incorrect orders), job interview role-play (responding to common and follow-up questions from a virtual interviewer), public transport navigation (interpreting announcements, requesting directions, and confirming routes in a simulated train station), and school open day interactions (asking and answering questions about facilities, events, and schedules during an interactive guided tour). These scenarios were purposefully selected for their relevance to learners' likely real-world encounters and were sequenced to increase in both linguistic demand and situational complexity over the six-week duration of the treatment.

## *Procedures*

The intervention lasted six weeks, with each group receiving two 45-minute listening lessons per week. In the experimental group, AI tools were systematically embedded at each stage of listening instruction. During the pre-listening stage, AI chatbots introduced topic-specific vocabulary and activated background knowledge through short, interactive exchanges. While listening, the LingQ gamification platform provided comprehension checks, vocabulary reinforcement, and adaptive feedback in real time. In the post-listening stage, Google Speech-to-Text enabled participants to compare AI-generated transcripts with the original scripts, prompting reflection on comprehension gaps and potential misinterpretations. In addition, a weekly VR immersion allowed learners to apply listening strategies in simulated, context-rich environments.

The control group followed the same lesson structure in terms of timing and sequence, but relied exclusively on the listening materials and activities provided in the prescribed textbook per the guidelines in the teacher's book. Pre-listening activities consisted of brief teacher-led discussions to introduce the topic and relevant vocabulary. While listening, students engaged with the textbook's recorded audio tracks, completing comprehension questions and vocabulary exercises without the aid of adaptive feedback or gamification. Post-listening tasks involved reviewing answers as a class and discussing key points, with the teacher providing oral corrective feedback. No virtual reality or AI-based tools were used in the control group, ensuring that any differences in outcomes could be attributed to the combined effect of the multi-tool AI treatment.

## *Data Analysis*

Post-test listening scores were analyzed using analysis of covariance (ANCOVA), with instructional group as the between-subjects factor and pre-test score entered as a covariate to adjust for baseline differences. This approach is appropriate for quasi-experimental designs using intact groups and allows post-test comparisons while accounting for initial variation in listening ability (Field, 2018). Adjusted means, standard errors, and 95% confidence intervals are reported. Given the bounded nature of the listening score scale, results are interpreted conservatively with primary emphasis on adjusted mean differences rather than standardized effect sizes.

## *Ethical Considerations*

Ethical approval for the study was obtained from the relevant institutional authority and the participating school. Written parental consent and student assent were secured prior to data collection. The AI tools used in the intervention generated limited interaction data (e.g., text responses and usage logs) for instructional purposes only; no biometric data was collected. All data were anonymized prior to analysis, stored on password-protected devices accessible only to the research team, and used solely for research purposes in accordance with institutional guidelines.

### *DATA AND MATERIALS AVAILABILITY*

De-identified data, scoring rubrics, item-level statistics from the pilot phase, and analysis scripts used to generate the reported results are available from the corresponding author upon reasonable request. Supplementary analyses conducted during manuscript development yielded convergent patterns and are not reported in detail here due to space constraints.

## FINDINGS

This section reports the findings of the study on the effects of AI-powered tools on listening comprehension compared with conventional textbook-based instruction. The analysis sample comprised 48 participants; at baseline, the two groups were broadly comparable, with pre-test means differing by only 0.22 points.

Prior to examining group differences, the psychometric properties of the listening measure were evaluated. Item-level analyses indicated acceptable functioning of the final instrument: item difficulty (proportion correct) ranged from 0.39 to 0.75 (mean $p \approx 0.54$), and most items showed corrected item-total point-biserial correlations ($r_{pb} \geq 0.20$), consistent with expected discrimination for a short, mixed-format listening measure. Internal consistency was satisfactory (Cronbach's $\alpha = 0.78$ at pre-test; 0.82 at post-test), and McDonald's $\omega$ was 0.79 (pre-test; 95% CI [0.72, 0.86]) and 0.83 (post-test; 95% CI [0.76, 0.89]). Inter-rater agreement for the 30% double-scored subsample was high (ICC[2,1] = 0.91 at pre-test; ICC[2,1] = 0.88 at post-test; Cohen's $\kappa = 0.86$). Full item-level outputs (item p, corrected r_pb, distractor-frequency tables), the complete scoring rubric, and rater-calibration logs are available from the corresponding author upon request.

Following the intervention, both groups showed higher post-test scores; descriptively, the experimental group had a higher post-test mean than the control group. Group means and standard deviations are reported in Table 1.

**Table 1. Means, standard deviations, and score ranges for pre-test and post-test listening scores by group**

| Group | n | Pre-test | | Pre-test Min–Max | Post-test | | Post-test Min–Max |
|---|---|---|---|---|---|---|---|
| | | Mean | SD | | Mean | SD | |
| Control | 24 | 10.96 | 3.00 | 2–18 | 13.38 | 3.05 | 1–17 |
| Experimental | 24 | 11.08 | 2.73 | 2–15 | 15.96 | 3.34 | 1–20 |

*Note:* Listening scores were based on a raw-score scale ranging from 0 to 20. Values reported are raw (unadjusted) means and standard deviations calculated directly from observed scores. Min–max values indicate the observed score range within each group. Two experimental participants and one control participant achieved the maximum post-test score (20).

A small number of learners reached the maximum post-test score, indicating limited compression at the upper end of the scale. Inspection of the score distributions indicated approximately symmetric post-test distributions with no extreme outliers.

Raw post-test means are reported for descriptive purposes, whereas adjusted post-test means represent model-based estimates obtained from an analysis of covariance (ANCOVA) that controlled for pre-test listening performance (Table 2). Results indicated a statistically significant effect of instructional condition on post-test performance, $F(1, 45) = 42.39$, $p < .001$, partial $\eta^2 = .49$. Adjusted means favored the experimental group ($M = 16.56$, $SE = .29$) over the control group ($M = 13.94$, $SE = .29$), indicating that learners receiving AI-enhanced instruction achieved higher listening scores after adjustment for baseline differences.

**Table 2. ANCOVA results for post-test listening scores**

| Source | SS | df | MS | F | P (exact) | Partial $\eta^2$ | N |
|---|---|---|---|---|---|---|---|
| Pre-test (covariate) | 30.12 | 1 | 30.12 | 15.48 | < .001 | 0.26 | 48 |
| Group (control vs. experimental) | 82.47 | 1 | 82.47 | 42.39 | < .001 | 0.49 | 48 |
| Error | 87.55 | 45 | 1.95 | | | | |
| Corrected total | 203.00 | 47 | | | | | |

*Note:* ANCOVA = analysis of covariance. Pre-test listening score was entered as a covariate. Listening scores were based on a raw-score scale ranging from 0 to 20. Adjusted means favored the experimental group (M = 16.56, SE = 0.29) over the control group (M = 13.94, SE = 0.29). Adjusted means represent estimated marginal means from the ANCOVA model and are not expected to correspond numerically to the raw means reported in Table 1.

Model diagnostics indicated no major violations of ANCOVA assumptions. The Group × Pre-test interaction was not significant ($F(1,45) = 0.29$, $p = 0.592$); residuals approximated normality (Shapiro–Wilk $W = 0.982$, $p = 0.432$) and Levene's test showed no evidence of unequal variances ($F = 0.41$, $p = 0.525$). Influence diagnostics (Cook's D) identified no case whose omission materially altered the adjusted effect.

The adjusted advantage observed for the experimental group is consistent with substantive improvement in listening performance within the CEFR B1 band for the descriptors sampled by the instrument. Categorical claims that participants migrated between lower and upper B1 sub-bands would require formal linking or standard-setting evidence (for example, rater-anchored cut scores or empirical concordance studies) together with documented item-by-item mapping and rater calibration; documentation of the mapping procedures, panel reports, and reconciliation logs is provided in the project supplement to permit independent verification.

Data and materials: the scoring rubric, item keys, item-level output tables, rater-calibration notes, and analysis codes (R scripts used for item statistics, ω bootstrap, and partial-$\eta^2$ CI computations) are available from the corresponding author upon request.

## DISCUSSION

The study examined whether integrating AI-powered instructional tools into an EFL listening curriculum produced larger gains in listening comprehension than conventional textbook-based instruction. The covariate-adjusted analysis indicated a clear advantage for the experimental group, with higher adjusted post-test listening scores after controlling for baseline differences. The adjusted mean difference between groups indicates a meaningful instructional advantage within the present instructional context; the associated effect size should be interpreted cautiously, given the bounded score scale and the sample size.

Several plausible, non-mutually exclusive mechanisms may account for the observed advantage. AI systems routinely provide adaptive sequencing of practice items, matching difficulty to current learner performance. This adaptive pacing concentrates practice where it is most informative and thereby increases the efficiency of learning. In addition, AI-enabled tools can increase the quantity and diversity of comprehensible input while allowing the provision of automated, immediate feedback that directs attention to segmental and suprasegmental features of speech (e.g., pronunciation variability, reduced forms, and prosody), which focuses attention on diagnostic features of the input and supports error correction on retrieval practice, ultimately accelerating perceptual attunement and decoding skills. Furthermore, many AI tools expose learners to a broader, more variable set of speech samples (different accents, speaking rates, reductions, and noise conditions) than a static textbook can provide, potentially supporting stable perceptual tuning and transfer to novel listening contexts.

These mechanisms operate together within an instructional ecology. Technology amplifies pedagogical design: AI affordances produce learning gains when integrated into coherent sequences (clear objectives, scaffolded tasks, and teacher mediation) that translate algorithmic feedback into classroom practice. In the absence of curricular alignment and scaffolding, adaptive algorithms risk producing isolated improvements that do not generalize; when coupled with teacher guidance, however, system output becomes actionable and supports progressive consolidation of strategy-based listening skills.

Empirical syntheses and contemporary reviews report that AI-based speech technologies, adaptive practice, and generative tutoring can enhance listening practice opportunities and scaffold learner attention in ways that are difficult to achieve with static textbook materials alone (e.g., Cooray et al., 2024; Dong et al., 2024; Wiboolyasarin et al., 2025; Wu, 2024). These affordances are consistent with broader meta-analytic evidence that AI-based instructional interventions often yield positive effects on language learning outcomes, especially where systems provide adaptive, feedback-rich practice and opportunities for repeated retrieval (e.g., Dong et al., 2024; Wiboolyasarin et al., 2025; Wu, 2024).

The pattern of results also speaks to the nature of the learning gains observed. Rather than merely increasing time-on-task, AI-assisted instruction seems to have altered how practice time is used, promoting selective attention to diagnostic cues, supporting repeated retrieval of critical forms, and offering corrective cycles closely tied to performance.

From a theoretical perspective, these findings sit comfortably within frameworks that emphasize deliberate practice and variability of input as engines of perceptual learning. They extend prior work by showing that algorithmic adaptation and automated feedback can operationalize these principles at scale in classroom contexts, thereby bridging laboratory evidence on perceptual learning with classroom practice. Meta-analytic evidence on AI and GenAI interventions in language learning corroborates the general direction and boundary of the effect reported here, while underscoring heterogeneity across features and contexts.

Finally, the findings carry measured pedagogical significance. When AI tools are selected and configured to align with curricular aims, and when teachers are supported to interpret and embed system feedback, the technology can materially improve listening outcomes. This implies that investment in teacher training and curricular integration, rather than the deployment of standalone tools, is a necessary condition for the benefits demonstrated here to appear in routine practice.

## CONCLUSIONS, LIMITATIONS, IMPLICATIONS, AND DIRECTIONS FOR FUTURE RESEARCH

The findings indicate that AI-enhanced listening, when deliberately integrated into a coherent curriculum and accompanied by transparent assessment practices, produces clear and meaningful improvements in listening comprehension in the present sample. The pattern of results is held under alternative analytic specifications, suggesting that the advantage reflects substantive learning-related change rather than an idiosyncrasy of a single analytic choice. These outcomes align with contemporary syntheses indicating that systems that offer adaptive practice, targeted feedback, and varied speech input are well-positioned to accelerate L2 listening development.

Several qualifications circumscribe the breadth of inference. One limitation concerns the imbalance in engagement between groups. The experimental condition introduced multiple novels, high-interest tools (viz., chatbots, gamification, and VR), while the control group relied solely on textbook activities. This disparity introduces the possibility of a novelty effect, as increased motivation and curiosity, rather than the AI tools themselves, may have contributed to the observed performance gains. Although typical in early-stage intervention research, this design asymmetry limits inferences about whether the advantage stems from AI-mediated scaffolding or from heightened learner engagement associated with new technologies.

Before extending the interpretation of these results, it is necessary to acknowledge that the post-test scores displayed some compression near the upper score range, with a subset of learners in both groups approaching the 20-point maximum. Such compression of the score range constrains variance and challenges core ANCOVA assumptions, particularly the expectation of normally distributed residuals and unbounded outcomes (Field, 2018). Although the diagnostic checks suggested reasonable robustness, the restricted scale likely inflates both adjusted means and the apparent magnitude of effect sizes. Therefore, the gains are interpreted as improvements within bounded scoring conditions rather than as definitive indicators of large population-level effects, and only cautious generalization beyond the sampled cohort is attempted.

The sample size is relatively modest, which warrants caution when generalizing beyond the present instructional context. Although the results indicate meaningful differences within the present sample, replication with larger and more diverse cohorts is required to assess the stability of the observed effects. In addition, the CEFR-referenced interpretation offered here is indicative rather than classificatory and should not be taken as evidence of categorical movement between proficiency bands without formal standard-setting procedures.

The findings point to several implications for practice. First, AI is most effective when it amplifies sound pedagogy rather than replacing it: systems should be aligned with curricular objectives, item specifications, and explicit practice routines, and teachers should be supported in interpreting and mediating system feedback. Second, where human scoring is required, clear scoring rules and rater training are necessary to preserve score validity. Third, institutions contemplating use of scores for placement or certification must not substitute informal mapping for formal linking; any categorical use of scores requires dedicated standard-setting or concordance studies. Investment in teacher professional development that covers interpretation of system diagnostics, scaffolding AI feedback, and monitoring fidelity is likely to yield better classroom outcomes than deployment without support.

Priority areas for further work are (1) larger, preferably pre-registered and stratified or cluster-randomized studies to assess generalisability and moderators of effect; (2) longitudinal and transfer studies to determine whether gains persist and extend to broader communicative tasks; (3) formal CEFR linking and concordance research (rater-anchored panels or empirical concordance with external CEFR-referenced anchors) to permit defensible categorical interpretation; and (4) component and mechanism studies (factorial trials, mediation analyses and log-file research) that isolate which features of AI systems, adaptive sequencing, corrective feedback, speech variability, or dialogic practice, drive learning. Methodologically, future studies should emphasise pre-registration, transparent reporting of reliability and linking procedures, and the sharing of analytic scripts and item-level outputs to facilitate cumulative progress.

In sum, AI-enhanced listening shows promise as a pedagogical tool when embedded in well-designed instruction and paired with rigorous assessment practices. The present evidence encourages continued, rigorous investigation while underscoring that technological promise must be matched by careful curricular integration, teacher development, and transparent evaluation before widespread, high-stakes adoption.

## REFERENCES

Abdalkader, S. M. A. (2023). *The impact of using artificial intelligence on enhancing EFL language fluency and self-regulation for the preparatory stage students in distinguished governmental language schools* [Doctoral Dissertation, Ain Shams University]. http://files.eric.ed.gov/fulltext/ED630026.pdf

Abdolrezapour, P., & Ghanbari, N. (2021). Enhancing learning potential score in EFL listening comprehension and self-regulation through self-regulated dynamic assessment procedures. *Language Testing in Asia, 11*, Article 10. https://doi.org/10.1186/s40468-021-00126-5

AlAli, R., & Wardat, Y. (2024). Enhancing classroom learning: ChatGPT's integration and educational challenges. *International Journal of Religion*, *5*(6), 971–985. https://doi.org/10.61707/znwnxd43

Al-Barakat, A. A., AlAli, R. M., Al-Hassan, O. M., Bataineh, R. F., Al-Saud, K. M., & Aboud, Y. Z. (2025). Beyond the pencil: Blogging for better written expression in the primary classroom. *International Journal of Information and Education Technology, 15*(7), 1460–1467. https://doi.org/10.18178/ijiet.2025.15.7.2347

Al-mawaly, H. M., & AL-Jamal, D. A. H. (2022). The effect of artificial intelligence application on Jordanian EFL sixth-grade students' listening comprehension and their attitudes towards it. *Journal of Positive School Psychology*, *6*(6), 8781–8791.

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

Alshahrani, K., & Qureshi, R. J. (2024). Review the prospects and obstacles of AI-enhanced learning environments: the role of ChatGPT in education. *International Journal of Modern Education and Computer Science*, *16*(4), 71–86. https://doi.org/10.5815/ijmecs.2024.04.06

Al-Zahrani, A. M. (2024). Unveiling the shadows: Beyond the hype of AI in education. *Heliyon*, *10*(9), e30696. https://doi.org/10.1016/j.heliyon.2024.e30696

Bataineh, R. F., & Al-Ghareeb, M. B. (2025). Starfall as a catalyst for Kuwaiti EFL young learners' reading comprehension: A teacher's reflections. *Journal of Ethnic and Cultural Studies, 12*(5), 141–153. https://doi.org/10.29333/ejecs/2338

Baxramova, M. M. (2025). Integrating ChatGPT as a listening assistant in secondary EFL classes: Benefits and limitations. *Pedagogik Islohotlar va Ularning Yechimlari, 15*(1), 76–78. https://www.wosjournals.com/index.php/medical/article/view/2973/3587

Buck, G. (2001). *Assessing listening*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732959

Cooray, S., Hettiarachchi, C., Nanayakkara, V., Matthies, D., Samaradivakara, Y., & Nanayakkara, S. (2024). Kavy: Fostering language speaking skills and self-confidence through conversational AI. *Proceedings of the Augmented Humans International Conference* (pp. 226-236). Association for Computing Machinery. https://doi.org/10.1145/3652920.3652944

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Davar, N. F., Dewan, M. A. A., & Zhang, X. (2025). AI chatbots in education: Challenges and opportunities. *Information*, *16*(3), 235. https://doi.org/10.3390/info16030235

Davies, D., & Pearse, E. (2000). *Success in English teaching*. Oxford University Press.

Dong, W., Pan, D., & Kim, S. (2024). Exploring the integration of IoT and generative AI in English language education: Smart tools for personalized learning experiences. *Journal of Computational Science, 82*, 102397. https://doi.org/10.1016/j.jocs.2024.102397

Farooqi, M. T. K., Amanat, I., & Awan, S. M. (2024). Ethical considerations and challenges in the integration of artificial intelligence in education: A systematic review. *Journal of Excellence in Management Sciences*, *3*(4), 35–50. https://doi.org/10.69565/jems.v3i4.314

Fathi, J., Rahimi, M., & Derakhshan, A. (2024). Improving EFL learners' speaking skills and willingness to communicate via artificial intelligence-mediated interactions. *System, 121*, 103254. https://doi.org/10.1016/j.system.2024.103254

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage.

Goh, C. C. M., & Aryadoust, V. (2025). Developing and assessing second language listening and speaking: Does AI make it better?. *Annual Review of Applied Linguistics*, *45*, 179–199. https://doi.org/10.1017/S0267190525100111

Hazaymeh, W. A., Alshaikhi, T., Abdul Wahab, M. O., & Khasawneh, M. A. (2025). Evaluating AI applications for enhancing listening comprehension among EFL students in English language in real-world scenarios. *World Journal of English Language, 15*(6), 209–223. https://doi.org/10.5430/wjel.v15n6p209

Hidayatullah, E. (2024). The impact of Talkpal.AI on English speaking proficiency: An academic inquiry. *Journal of Insan Mulia Education*, *2*(1), 19–25. https://doi.org/10.59923/joinme.v2i1.98

Hınız, G. (2026). Using AI to enhance receptive foreign language skills. In A. Shukla, B. Meeprom, K. Khunasathitchai, & N. Yadav (Eds.), *AI's role in language learning and communication* (pp.65–100). IGI Global. https://doi.org/10.4018/979-8-3373-5681-5.ch003

Hussein, S., Safina, Qureshi, S. S., & ul Emman, S. K. (2025). Analyzing how AI can both exacerbate and help overcome digital inequalities in education, especially in underserved regions. *The Critical Review of Social Sciences Studies*, *3*(2), 143–157. https://doi.org/10.59075/4964v051

Hwaider, S. M. (2017). Problems of teaching the listening skill to Yemeni EFL learners. *International Journal of Scientific and Research Publications, 7*(6), 140–148.

Karkera, S., & Chamundeshawari, C. (2018). YouTube: A teaching tool to improve listening skills. *International Journal of Creative Research Thoughts, 6*(2), 1311–1316. https://ijcrt.org/papers/IJCRT1813041.pdf

Kim, A., & Su, Y. (2024). How implementing an AI chatbot impacts Korean as a foreign language learners' willingness to communicate in Korean. *System, 122*, 103256. https://doi.org/10.1016/j.system.2024.103256

Kukulska-Hulme, A. (2020). Mobile-assisted language learning and AI: Future directions. In C.A Chapelle (Ed.), *The concise encyclopedia of applied linguistics*. Wiley. https://doi.org/10.1002/9781405198431.wbeal0768.pub2

Li, J., Zong, H., Wu, E., Wu, R., Peng, Z., Zhao, J., Yang, L., Xie, H., & Shen, B. (2024). Exploring the potential of artificial intelligence to enhance the writing of English academic papers by non-native English-speaking medical students-the educational application of ChatGPT. *BMC Medical Education*, *24*, Article 736. https://doi.org/10.1186/s12909-024-05738-y

Lindsay, C., & Knight, P. (2006). *Learning and teaching English: A course for teachers*. Oxford University Press.

Miao, F., Holmes, W., Huang, R., & Zhang, H. (2021). *AI and education: Guidance for policy-makers*. United Nations Educational, Scientific and Cultural Organization. https://doi.org/10.54675/PCSP7350

Nowrouzi, S., Tam, S. S., Zareian, G., & Nimehchisalem, V. (2015). Iranian EFL students' listening comprehension problems. *Theory and Practice in Language Studies, 5*(2), 263–269. https://doi.org/10.17507/tpls.0502.05

Ordoñez Procel, G. J., Freire Medina, M. L., Sotomayor Sanchez, D. J., & Poma Tacuri, M. A. (2024). *Using technology in English teaching*. Centro de Investigación y Desarrollo. https://doi.org/10.37811/cli_w1048

Osada, N. (2004). Listening comprehension research: A brief review of the past thirty years. *Dialogue, 3*, 53–66. https://www.talk-waseda.net/dialogue/no03_2004/2004dialogue03_k4.pdf

Panagiotidis, P. (2025). AI-driven applications for language learning: Transformative technologies and educational benefits. *Proceedings of the 17th Annual International Conference on Education and New Learning Technologies* (pp. 747–758). IATED. https://doi.org/10.21125/edulearn.2025.0282

Richards, J. C. (2008). *Teaching listening and speaking: From theory to practice*. Cambridge University Press.

Suvorov, R. (2022). Technology and listening in SLA. In R. Suvorov (Ed.), *The Routledge handbook of second language acquisition and technology* (pp. 136–147). Routledge. https://doi.org/10.4324/9781351117586

Tolstykh, O. M., & Oshchepkova, T. (2024). Beyond ChatGPT: Roles that artificial intelligence tools can play in an English language classroom. *Discover Artificial Intelligence, 4*, Article 60. https://doi.org/10.1007/s44163-024-00158-9

Umoke, C. C., Nwangbo, S. O., & Onwe, O. A. (2025). AI-driven educational policy design: Enhancing equity and access through intelligent data analytics. *International Journal of Computer Science and Mathematical Theory, 11*(3), 1–19.

Vandergrift, L. (2004). Listening to learn or learning to listen? *Annual Review of Applied Linguistics*, *24*, 3–25. https://doi.org/10.1017/S0267190504000017

Vandergrift, L., & Goh, C. (2012). *Teaching and learning second language listening: Metacognition in action*. Routledge.

Vesna, L., Sawale, P. S., Kaul, P., Pal, S., & Murthy, B. S. N. V. R. (2025). Digital divide in AI-powered education: Challenges and solutions for equitable learning. *Journal of Information Systems Engineering and Management*, *10*(21), 300–308. https://doi.org/10.52783/jisem.v10i21s.3327

Vygotsky, L. S. (1978). *Mind in society: Development of higher psychological processes*. Harvard University Press.

Wiboolyasarin, W., Wiboolyasarin, K., Tiranant, P., Jinowat, N., & Boonyakitanont, P. (2025). AI-driven chatbots in second language education: A systematic review of their efficacy and pedagogical implications. *Ampersand*, *14*, 100224. https://doi.org/10.1016/j.amper.2025.100224

Wu, X.-Y. (2024). Artificial intelligence in L2 learning: A meta-analysis of contextual, instructional, and social-emotional moderators. *System, 126*, 103498. https://doi.org/10.1016/j.system.2024.103498

Zghoul, W. M., & Bataineh, R. F. (2024). Flipgrid: Unlocking the English speaking potential of Jordanian adolescent EFL learners. *Journal of Information Technology Education: Innovations in Practice, 23*, Article 17. https://doi.org/10.28945/5407

## AUTHORS

**Dr Ruba Fahmi Bataineh** is a professor of TESOL at the Department of Curriculum and Methods of Instruction at Yarmouk University, Jordan. Professor Bataineh is currently the vice president of Yarmouk University. She has also been the dean of the Faculty of Arts and Sciences and the director of the Language Center at Al-Ahliyya Amman University, 2022/2023 (on sabbatical leave from Yarmouk University). Formerly, Professor Bataineh was the founding executive director of the National Center for Curriculum Development, Jordan (2018-2020), and the director of the Prince Salman Center for Research and Translation, Prince Sultan University, Saudi Arabia (2011-2013), where she received the Distinguished Researcher Award (2010). Professor Bataineh has conducted workshops and seminars, delivered keynote speeches, conducted university program assessments, and served as a consultant in language teaching and teacher preparation, both locally and internationally. She has also published extensively on pragmatics, literacy, CALL, and teacher education in renowned international and regional journals. She is also an affiliate of professional organizations and a member of the advisory, editorial, and/or review boards of several regional and international journals.

**Dr Salameh Fleih Obeiah** is currently an EFL supervisor in the Jordanian Ministry of Education. He holds a PhD from the TEFL Program at Yarmouk University (Irbid, Jordan). He is also the recipient of the Sheikh Faisal Bin Qassim Al Thani Award for Educational Research (2023). Dr Obeiah has published extensively in renowned regional and international journals on contrastive analysis, teacher education, and technology-enhanced pedagogy.

**Rula Fahmi Bataineh** is an Assistant Professor of Rhetoric and Linguistics at Jordan University of Science and Technology, Jordan. Her research spans pragmatics, discourse analysis, language development, and multimodal communication in language learning. She has a strong record of interdisciplinary collaboration, with recent publications examining politeness, digital transformation, and language development.