Indonesian Journal of English Language Teaching and Applied Linguistics

Vol. 11(2), November 2026
www.ijeltal.org
e-ISSN: 2527-8746; p-ISSN: 2527-6492

From Role-Play to AI Interaction: Speaking Confidence Development among EFL Nursing Students at AKPER YPTK Solok

Kristian Burhan[1]*, Nofri Yuhelman[2]

[1] *Universitas Negeri Padang. e-mail: kristianburhan@unp.ac.id**
[2] *Universitas Negeri Padang. e-mail: nofriyuhelman@unp.ac.id*

Received 07 July 2026 | Received in revised form 30July 2026 | Accepted 08 August 2026

ARTICLE INFO

Keywords:
ChatGPT, EFL, conversational AI, role-play, speaking confidence, willingness to communicate

DOI:
http://dx.doi.org/10.21093/ijeltal.v11i2.2720

ABSTRACT

Conversational artificial intelligence is now a common sight in classrooms for foreign languages, but evidence of it improving speaking confidence compared to established techniques like role-play is still not available. With a sample of 60 EFL nursing students who were part of the program AKPER YPTK Solok, in West Sumatra, Indonesia, the explanatory sequential mixed-methods study looked into how combining classroom role-play with structured ChatGPT-assisted rehearsal performed against a control group role-playing only. Speaking confidence, willingness to communicate, speaking anxiety, and oral performance were the four factors assessed using modified scales and an analytic speaking rubric, together with interviews of eight students to elaborate upon the quantitative results. A significant increase in spoken confidence was detected in the group with AI assistance by a very small margin. The gains in willingness to communicate and oral performance were relatively small, and as for speaking anxiety, no significant difference was found. The interview data suggest that the learners experienced greater confidence and readiness in class after repeatedly practicing dialogue that was free of any judgment, and they found that a conversation with a person and with AI resulted in significantly different reactions. Based on these findings, it seems that a good use of conversational AI is as a rehearsal tool that develops learners' confidence in speaking and prepares them for authentic human interaction it has not replaced.

How to cite:
Burhan, K. & Yuhelman, N. (2026). From Role-Play to AI Interaction: Speaking Confidence Development among EFL Nursing Students at AKPER YPTK Solok. *Indonesian Journal of English Language Teaching and Applied Linguistics*, 11(2), 183-202

1. Introduction

Speaking is the skill that makes English learners, as a foreign language, the most vulnerable. It is done in front of an audience whose views the learners cannot control or foresee. In Indonesian higher education, where English often functions as a service subject layered onto

a demanding professional curriculum, these pressures are sharpened by limited classroom hours and few authentic opportunities to use the language outside the timetable (Poedjiastutie & Oliver, 2017). Vocational students in professions like nursing have a specific kind of issue: at some point they will have to communicate with patients, colleagues, and documentation in English. However, the actual speaking practice they receive is just a couple of sessions per week, which are usually shared with thirty or more people. The question of how to give such learners more, and more confidence-building, speaking practice is therefore a practical concern.

Role-play remains one of the best ways to address that question. By placing learners inside a structured scenario with assigned identities and goals, it lowers the stakes of performance, supplies a communicative purpose, and rehearses the kind of interaction learners will meet beyond the classroom (Richards, 2006). Indonesian classroom research has repeatedly associated role-play with gains in fluency and a more positive affective climate, although the same studies note that learners who are already anxious can find the prospect of performing in front of classmates intimidating rather than liberating (Dwiyanti & Lolita, 2023; Katemba & Grace, 2023). In fact, role-play makes effective use of the actual social environment, which for many learners means their biggest nightmare, but this way a place for other methods that enable rehearsals is given.

Conversational artificial intelligence has emerged, almost overnight, as a candidate for filling that gap. Since the public release of ChatGPT in late 2022, large language models have given learners access to an interlocutor that is available at any hour, responds in fluent and contextually appropriate English, and offers correction without the social cost of being corrected in front of others (Kohnke et al., 2023; Xiao & Zhi, 2023). Early synthesis work suggests that such tools can support personalised, low-stakes practice across language skills, though much of the evidence concerns writing and perceptions rather than measured oral outcomes (Crompton & Burke, 2023; Teng, 2024). If students do language drilling and role-play with a chatbot, they probably get familiar with the language and the situation, so when they finally do the role-play in the classroom, they might be happier and speak more easily.

Whether that intuition holds for the affective dimension of speaking is less settled than enthusiastic commentary implies. A useful distinction here is between speaking confidence, willingness to communicate (WTC), and speaking anxiety, which are related but not interchangeable. Confidence concerns a learner's belief in their capacity to perform; WTC describes the readiness to initiate communication when free to do so (MacIntyre, 2007; MacIntyre et al., 1998); and anxiety captures the apprehension that accompanies, and can degrade, oral performance (Horwitz et al., 1986). These constructs do not always move together. A learner may grow more confident in rehearsal while remaining anxious about live assessment, or become more willing to speak with a machine without that willingness transferring to a human audience. Treating the affective outcome of AI practice as a single undifferentiated quantity risks obscuring exactly the contrasts that matter for teaching. The significance of the distinction is theoretical before it is practical. If confidence, willingness, and anxiety are separable appraisals with different antecedents and different degrees of stability, then any account of how AI-mediated practice works has to specify which appraisal a given tool can plausibly reach, and why. A theory predicting a uniform affective benefit is, in that respect, under-specified, and an intervention study that reports a single composite cannot test it. The design reported here is built around that requirement.

Recent experimental work has begun to map these contrasts, with results that are encouraging but uneven. Fathi et al. (2024), in a well-controlled study, found that AI-mediated interaction outperformed conventional instruction on both speaking skills and WTC among EFL learners. Wang et al. (2024) reported that different generative AI chatbots influenced WTC, self-perceived communicative competence, and foreign-language speaking anxiety in distinct ways, with effects on anxiety proving the least consistent. Studies using voice-based agents have likewise documented improvements in speaking ability and motivation, while flagging cognitive load, technological dependence, and the artificiality of machine talk as recurrent concerns (Klimova et al., 2024; Yang et al., 2022).

Instead, the contrary evidence should be given due attention since it has a direct bearing on the claim of emotion, the main subject of the paper. Çakmak (2022) conducted a study of ninety Turkish EFL learners using Replika over twelve weeks and found that speaking performance increased speaking anxiety remained unchanged. Learners even mentioned that they felt awkward and were quite often misunderstood. The researchers thus made a recommendation that chatbot interaction in general may not contribute effectively to anxiety reduction. Ballıdağ & Aydın (2025) extended their analysis to another method of language learning and showed that, after chatting with the chatbot, users were more affected by anxiety than those who were paired with classmates. This may have been the result of the feedback that the user received from the chatbot being too impersonal, and, moreover, communication difficulties that happened quite easily and affected particularly those whose level was not very high. Meta-analytic evidence tempers the picture from the other direction rather than contradicting it: pooled effects on second language learning are positive but moderate rather than transformative, and they are markedly heterogeneous across tools, skills, and settings (Li et al., 2025; Lyu et al., 2025; Wu & Li, 2026). In general, results are mostly good with regard to engagement and willingness, a bit less so in relation to performance measured objectively, and truly mixed with some negative elements with respect to anxiety. Confidence, despite being central to how teachers talk about speaking, has seldom been isolated as an outcome in its own right, and that omission is one reason the divergence among these affective outcomes has gone largely untested.

Two limitations in this body of work shape what the researchers set out to do. The first is a design problem. Most experimental studies pit AI practice against ordinary instruction rather than against an active, theoretically comparable technique, so an advantage for the AI group may amount to little more than extra practice time or the novelty of a new tool. The researchers regard that as too weak a test. A fairer one sets AI-supported practice against an established communicative method, role-play, so that both groups receive the same amount of communicative activity and the contribution of the AI component alone becomes visible. The second limitation is one of population. Recent studies have leaned heavily on general or pre-service EFL learners, while vocational and professional learners in Indonesian institutions, whose motivation is tied to concrete occupational demands, have rarely been studied (Munirah et al., 2023; Nurlaelawati et al., 2022).

The contribution the researchers claim follows directly from these gaps. Rather than asking the familiar question of whether AI helps, the researchers ask where its affective leverage actually falls, and the researchers answer it by measuring confidence, willingness to communicate, anxiety, and rated performance as separate outcomes under an active-control comparison in a vocational nursing cohort. Treating the affective domain as a single bundle,

the researchers argue, is precisely what has kept the field from seeing that these outcomes need not move together. To our knowledge, no prior study has tracked these four outcomes side by side within a single role-play-versus-role-play-plus-AI design in an Indonesian vocational ESP setting; that is the gap the present design is built to expose.

With that aim in view, the study examined how speaking confidence develops when classroom role-play is augmented with structured ChatGPT interaction, set against role-play conducted in the conventional way. The researchers situated the work in a nursing academy in West Sumatra, Indonesia, the same professional setting in which our earlier needs analysis identified speaking as the skill these students most needed and least practised (Burhan & Putra, 2021). Confidence was tracked alongside, but kept distinct from, willingness to communicate and anxiety, so that any divergence among the three would be plain to see. Oral performance was rated as well, to tie the affective measures to observable language behaviour, and learner interviews were used to interpret the numbers. Three questions guided the work:

(1) Does combining role-play with ChatGPT interaction produce greater gains in EFL learners' speaking confidence than conventional role-play alone?

(2) How do any effects on confidence compare with effects on willingness to communicate, speaking anxiety, and rated oral performance?

(3) How do learners account for the role of AI-mediated rehearsal in their developing confidence to speak?

On the basis of the sociocultural and interactionist reasoning developed below, the researchers anticipated a confidence advantage for the AI-supported group and expected the affective outcomes to diverge rather than move in unison, but the researchers treated the direction of the anxiety effect as an open empirical question given the inconsistency of prior findings.

2. Literature Review

2.1 Role-play and the Social Demands of Speaking Development

Role-play has earned a fixed position in communicative language teaching because it provides the need to communicate and thus naturally gives rise to a communication situation. Assigning learners roles, a situation, and an objective converts abstract language practice into purposeful interaction, and the resulting negotiation of meaning is widely held to drive acquisition (Richards, 2006). In Indonesian secondary and tertiary settings, classroom studies have linked role-play to improvements in fluency, vocabulary use, and participation, and to a warmer affective climate in which learners are more prepared to take risks (Dwiyanti & Lolita, 2023; Katemba & Grace, 2023). The technique's strength, however, doubles as its constraint. As performing before an audience, the learners who most need speaking practice are precisely those who are either shy or stressed; also, these same learners are those who feel the role-play's exposure most negatively. The evidence therefore includes the contradiction that the study emphasizes: playing a role enhances communication skills by exposing one to the public, but the exposure itself is the very reason anxious learners hesitate to take part.

2.2 Confidence, Willingness to Communicate, and Anxiety

Affective variables sit at the centre of speaking development, but they are not a single variable. Willingness to communicate was conceptualised by MacIntyre et al. (1998) as the product of a layered set of situational and enduring influences, with self-perceived communicative competence and low anxiety among its more immediate antecedents. MacIntyre (2007) later stressed that the decision to speak is volitional and situated, shaped moment to moment by context. Speaking anxiety, theorised by Horwitz et al. (1986) as a situation-specific apprehension, can depress performance independently of competence and tends to be comparatively stable. Confidence, finally, is the learner's appraisal of their own capacity to carry out a speaking task. Empirical work confirms that these constructs are correlated yet separable: learners with higher communicative competence and lower anxiety tend to be more willing to communicate (Peng & Woodrow, 2010), and digital or informal exposure to English has been shown to raise WTC partly by building confidence (Lee & Drajati, 2019; Reinders & Wattana, 2015). The implication for intervention research is methodological as much as theoretical

2.3 Conversational AI and ChatGPT in Speaking Practice

Dialogue-based computer-assisted language learning predates the current generation of chatbots, and earlier syntheses already identified its promise for low-pressure conversational rehearsal (Bibauw et al., 2019; Golonka et al., 2014). The behaviour of early bots as task partners suggested that learners would sustain interest and practice when the interlocutor was patient and non-evaluative (Fryer & Carpenter, 2006; Fryer et al., 2017). Generative models have sharpened these affordances. ChatGPT produces coherent, contextually responsive English, adapts to a prompt, and supplies on-demand correction, which has prompted a wave of studies on its classroom use (Kohnke et al., 2023; Xiao & Zhi, 2023). These potential benefits mainly concern the aspects of involvement, freedom, and being open to trying out new language skills without the fear of others assessing them negatively (Ghafouri, 2024; Teng, 2024; Tram et al., 2024). The speaking outcomes, however, are presented in a more restrained fashion. A number of voice and chatbot studies have revealed a marked enhancement of oral ability and motivation (Fathi et al., 2024; Yang et al., 2022). However, different types of language use, such as speaking, anxiety, and tool effectiveness, can lead to very different outcomes, as found in the research of Klimova et al.(2024) and Wang et al. (2024). What is repeatedly mentioned in the texts is that interaction with machines does not really cover natural elements present in human conversation as for instance spontaneity, physical presence and social expectations meaning that a person's development with a chatbot may not directly be reflected in how a person speaks to others face-to-face. Three recently made reviews highlight the issue of such reservations in figures. Wu & Li (2026) brought together results of twenty-one works on the usage of AI chatbots as language tools and got an overall, average-level effectiveness. Lyu et al. (2025) presented results of thirty-one comparative studies and mentioned that the level of accessibility was a major factor besides modality when the medium was a language aid, while the overall level of efficiency was also quite similar; and Li et al. (2025) collected outcomes of forty-one experiments and quasi-experiments with a focus on the impact of generative AI chatbots and they came up with an outcome between moderate and large effects in terms of second language learning but this depended on many factors like targeted skills, learner demographics and length of the program. The three of them are in agreement as regards the fact that the differences

among the results, rather than the typical, are the most prominent points. A few of the papers support the matter from the point of affective variables by showing that speaking anxiety was unchanged by chatbot use (Çakmak, 2022) or even increased compared to talking to peers (Ballıdağ & Aydın, 2025). Therefore, in a fairly accurate and balanced way the field shows results with different benefits which will depend, to a large degree, on the situation and the targeted language skill, as well as that the change of the anxiety is a separate matter and not one that can be predicted or caused by any of the activities.

The same body of work has begun to examine AI-mediated feedback and interaction in writing and argumentation, where automated correction and dialogic support improve accuracy and revision (Guo et al., 2023, 2024; Yan, 2023). These findings matter for speaking because the mechanism learners describe, that of private and repeated judgement-free correction, is the same one hypothesised to build oral confidence. What is largely missing is a design that holds communicative activity constant across conditions, so that the AI contribution to confidence can be distinguished from the contribution of practice itself. The contrast between an AI-augmented role-play group and a conventional role-play group, adopted here, responds to that gap.

2.4 Theoretical Framing: Sociocultural Mediation and Interaction

Two major perspectives provide support to the predictions. From the point of view of sociocultural theory, learning happens through the mediation of tools and through interaction with more knowledgeable others, and the person's development is achieved through learning by being provided assistance in performing tasks that are just slightly above the person's independent level (Vygotsky, 1978). The dialogue between the user and the computer program can also be viewed as a mediating element: the agent helps learners to practice; it offers correction; it enables learners to work at the level which would be a little challenging for them in front of their peers. Besides, the Interaction Hypothesis provides another line of explanation. Long (1996) pointed out that the use of comprehensible input, mutual clarification, and giving feedback during communication are the main ways in the development of acquisition; other, output-oriented theories also maintain that creating language and the awareness of language gaps contribute to development (Swain, 1995). Thus, a chatbot would allow multiple communication turns and quick feedback, and in doing so, response formulation would become a kind of meaningful output. When one takes these two perspectives as a whole, it would seem that AI-driven rehearsal could really help the learners in developing the trait of their emotional evaluation that is a result of the most frequent practice in a safe, threat-free environment, i.e., confidence. However, these two frameworks do have quite a weak theoretical basis for predicting effects on personality traits like dispositional anxiety, which, as the authors rightly state, are rather not susceptible to the effects of short-term mediation. As such a line of reasoning is the basis for different emotional states being put to a test by the researchers.

3. Research Methodology

3.1 Research Design

This study employed a quasi-experimental, pre-test-post-test design with a non-equivalent control group and was augmented by the use of qualitative interviews arranged in a sequential explanatory design. Given that the students came in two classes taught at fixed

times, it was not possible to randomly assign each student; accordingly, the allocation was at the class level: one whole class was placed in the experimental condition while the other was in the control condition, and the lecturer decided which class went to which condition prior to any data collection. This class-level assignment is a recognised constraint of school-based research and the reason a quasi-experimental rather than a true experimental design was adopted. The quantitative strand established whether and where the conditions differed, and the qualitative strand interpreted that pattern from the learners' own perspective. The reason why pre-test scores were kept throughout as covariates is because of the fact that the two groups were originally unequal. Therefore, baseline similarities to every measure were established by means of explicit, direct testing instead of simply being taken for granted (Section 4.1).

3.2 Participants

There were sixty participants in this study, of which students who took the English course at Akademi Keperawatan (AKPER) YPTK Solok, a nursing academy in West Sumatra, Indonesia were chosen. Each of the two study groups had 30 members. The students in the role-play group were asked to do role-play tasks while also using ChatGPT. The second group of students who used only face-to-face role-play for their practice was called the control group. The group was predominantly female, consistent with enrolment patterns in Indonesian nursing programmes, and learners were in the early years of tertiary study with English functioning as a foreign language used mainly within coursework. Participation followed convenience sampling at the class level, constrained by the academy's existing groupings. Because intact groups were used, baseline equivalence on all outcome measures was examined directly rather than assumed, and pre-test scores were used as covariates in the main analyses. Two features of this sampling frame bound the claims that follow. Because the experimental and control groups were randomly assigned across classes (i.e., to a whole class), our method does not support the identification of participants as representative of a national or even a regional or institutional sample of Indonesian vocation nursing students, nor can one be sure even after covariates that there was no differences like prior teachers, timetable position, or peer make-up which the other cohort would not have if only they were able to be exposed to the same condition. In this case, the study is quite rightly regarded as a one-time, one-place finding, and not as a population estimation. To collect the qualitative data, the first eight learners who participated from the experimental group were selected; this is a logical choice following the third research question, and that one asks the learners themselves to describe what AI-mediated rehearsal meant to them, an opportunity that was not given to the control group. In effect, the consequence of the researchers' decision not to recruit controls for the qualitative interviews is that they are not able, in one piece of evidence, to show how differently the two groups experienced the situation and thus the interview findings can at most shed light on the process inside the experimental condition and cannot be taken as a cross-condition comparison. This limitation is revisited in Section 5.

3.3 Instruments

Four instruments generated the data, and each self-report scale was adapted to the speaking context of this cohort from established sources. Speaking confidence was measured with a fifteen-item scale on which learners rated their belief in their capacity to handle spoken English tasks (1 = strongly disagree, 5 = strongly agree); the items drew on self-efficacy

formulations of L2 confidence (MacIntyre et al., 1998), with a representative item reading, "I feel confident when I have to speak English in front of others." Willingness to communicate was measured with a twelve-item scale capturing readiness to initiate English communication across classroom and out-of-class situations, adapted from the WTC tradition (MacIntyre, 2007; MacIntyre et al., 1998); a representative item reading, "I am willing to start a conversation in English with a classmate." Speaking anxiety was measured with a fifteen-item scale indexing apprehension during oral performance, adapted from the Foreign Language Classroom Anxiety Scale (Horwitz et al., 1986), with a representative item reading, "I feel nervous when I speak English without preparation," so that higher scores denote greater anxiety. All three used the same five-point Likert format. Oral performance was evaluated through an analytic speaking rubric assessing fluency, vocabulary, grammar, comprehensibility, and pronunciation. A five-point scale was used to rank individual performance on each speaking dimension. The approach mirrored commonly practiced ways of communicative speaking assessment (Richards, 2006). Each tape recording of the student's performance was assessed separately by two raters, after the raters had agreed substantially and the raters' scores were averaged.

A semi-structured interview protocol with ten open-ended questions was used during which the learner's prior speaking experience was inquired about, as well as their routines of using ChatGPT for rehearsals; they reflected on the practice differences between machine and peer, expressed their uncertainties about the AI output, and finally, they were asked about their changing perceptions of a live performance. Adaptation followed a documented sequence rather than direct borrowing. Items from the source instruments were first screened for relevance to a vocational nursing cohort, and wording was localised so that the situations described referred to classroom and clinical-communication contexts these learners recognise; items describing settings irrelevant to the cohort were removed.

The adapted scales were administered in a bilingual format to ensure that participants clearly understood each item while preserving the meaning of the original instruments. Content validity was confirmed after the instrument was reviewed by experts, who consisted of two lecturers in English for Specific Purposes (ESP) and one lecturing nurse. The experts were asked to assess whether each one of the items was relevant and clear enough for vocational nursing students as well as whether it was a good fit with real-life nursing situations. After they had given their comments, the content team did a revamp of the items that were identified as needing changes. The analytic speaking rubric was reviewed by the same panel to ensure adequate coverage of the speaking constructs assessed. Prior to scoring the participants' performances, the two raters underwent a calibration session using sample recordings until acceptable agreement in scoring was achieved. Given the relatively small sample size, construct validity was supported through expert validation and internal consistency evidence rather than factor-analytic procedures.

Accordingly, evidence for construct validity in the present study was based on expert review and content validation, complemented by the internal consistency evidence reported below, rather than factor-analytic procedures, which were not appropriate for the present sample size. Cross-instrument and internal consistency reliabilities, as well as inter-judge reliability on the subject population, have also been calculated. All instruments show high reliability coefficients ranging from good to excellent.

3.4 Treatment Procedures

The intervention was conducted over the course of eight weeks as part of the regular course schedule. The communicative themes that both groups dealt with were drawn from the nursing curriculum in the same order and included greeting and assessing patients, taking a history, giving health education, and handover communication. In each weekly cycle, the two groups received equivalent classroom role-play instruction and an equal allocation of practice time, so that the conditions differed only in the rehearsal that preceded live role-play. Students from the experimental group had used ChatGPT to prepare their roles in the classroom. Having a set of teacher's prompts, the learners created dialogues for different situations, the learner role of one of the dialogues and the model was the one for the other, asked for grammatical correction, word choice and word phrasing, and kept reenacting until they were ready. They were guided to treat the model's output critically and to verify uncertain language against course materials. The control group prepared for the same in-class role-plays through conventional means: they reviewed materials, planned with peers, and rehearsed without AI support. All learners then performed the role-play in class, where the recorded performances used for rating were collected at pre-test and post-test.

3.5 Data Analysis Procedures

Item responses were aggregated into mean composite scores for each scale, and the two rubric ratings were averaged into a single oral-performance score per learner at each time point. The analytic sequence was chosen so that each step answered a specific question. Descriptive statistics summarised the distributions; independent-samples t-tests on pre-test scores checked baseline equivalence, since a quasi-experimental design cannot assume it. Paired-samples t-tests then established whether each group changed over time, with Cohen's d expressing the size of that change. The principal between-group question, namely whether the experimental group ended higher after accounting for where it started, was addressed with one-way analysis of covariance (ANCOVA) on post-test scores using the pre-test as a covariate, with partial η^2 as the effect-size index; ANCOVA was preferred over a simple comparison of post-test means precisely because the groups were non-equivalent at baseline. The assumptions underlying ANCOVA were tested rather than presumed: the normality of residuals (Shapiro–Wilk), the homogeneity of variances (Levene's test), the linearity of the covariate–outcome relationship, and, critically, the homogeneity of regression slopes (the pre-test × group interaction).

Because those assumptions were not met for every variable (Section 4.1), two procedures that do not depend on them were added as the basis for inference where ANCOVA was compromised: bias-corrected bootstrap 95% confidence intervals (10,000 resamples) for the between-group difference in change scores, and Mann–Whitney U tests on the change scores. Convergence across these methods, rather than any single p-value, was treated as the test of a robust effect, and divergence among them was read as a signal of fragility rather than ignored. Qualitative data were translated and analysed through the reflexive thematic analysis of Braun & Clarke (2006), moving from familiarisation and open coding to iterative theme construction, and the resulting themes were used to interpret the quantitative pattern. Computation was carried out in Python's scientific stack, with analytic choices consistent with standard treatments of educational data.

4. Results

4.1. Preliminary Analyses, Reliability, and Baseline Equivalence

The measurement instruments performed reliably on this sample. Internal consistency was high for speaking confidence and willingness to communicate at both time points and acceptable for speaking anxiety, and the two raters showed strong agreement on the oral-performance rubric (Table 1). Independent-samples t-tests on the pre-test scores returned no significant differences between the groups on any measure, although willingness to communicate approached the conventional threshold; the groups were therefore treated as adequately equivalent at baseline, and pre-test scores were retained as covariates to adjust for residual differences (Table 2).

Table 1. *Internal consistency (Cronbach's α) and inter-rater reliability*

Measure	Pre-test	Post-test	Reliability type
Speaking confidence (15 items)	.88	.90	Cronbach's α
Willingness to communicate (12 items)	.90	.92	Cronbach's α
Speaking anxiety (15 items)	.85	.74	Cronbach's α
Oral performance rubric (2 raters)	.91	.92	Inter-rater r

Table 2. *Descriptive statistics by group and time, and baseline equivalence*

Measure / Group	Pre M (SD)	Post M (SD)	Gain	Baseline t	p
Speaking confidence				1.24	.222
Experimental	3.44 (0.84)	3.76 (0.65)	+0.32		
Control	3.20 (0.64)	3.33 (0.56)	+0.13		
Willingness to communicate				1.98	.053
Experimental	3.96 (0.83)	4.17 (0.62)	+0.21		
Control	3.57 (0.71)	3.64 (0.69)	+0.07		
Speaking anxiety				1.00	.319
Experimental	3.59 (0.82)	3.44 (0.39)	−0.15		
Control	3.40 (0.58)	3.33 (0.41)	−0.07		
Oral performance				0.89	.379
Experimental	2.78 (0.27)	3.05 (0.30)	+0.27		
Control	2.73 (0.18)	2.87 (0.17)	+0.14		

Note. Composite scores range from 1 to 5. For speaking anxiety, higher scores indicate greater anxiety. Baseline t and p are from independent-samples t-tests on pre-test scores (df = 58).

Before the covariance analyses were interpreted, their assumptions were tested for each variable, and the results were not uniform (Table 3). The covariate was related linearly to the outcome in every case, and the homogeneity of variances held throughout. The two remaining assumptions, however, were met only in part. For speaking confidence, the regression slopes were homogeneous (p = .056), and the model was sound apart from mild non-normality of residuals, so ANCOVA is defensible and its result can be read at face value. For oral performance, the slopes were homogeneous as well, but variances were unequal, and the covariate–outcome relationship was comparatively weak. For willingness to communicate and for speaking anxiety, the homogeneity-of-slopes assumption was violated, which means the single covariate-adjusted group effect that ANCOVA reports does not capture the way the groups differ across the range of pre-test scores. The researchers therefore treat ANCOVA as the primary inference only for speaking confidence and lean on the assumption-free gain-score analyses (the bootstrap intervals and Mann–Whitney tests) as the basis for judging the other three outcomes.

Table 3. *ANCOVA assumption checks by outcome variable*

Assumption	Confidence	WTC	Anxiety	Performance
Homogeneity of slopes (p)	.056 ✓	<.001 ✗	<.001 ✗	.391 ✓
Levene homogeneity of var. (p)	.361 ✓	.829 ✓	.974 ✓	.037 ✗
Shapiro–Wilk residuals (p)	.001	.001	.194 ✓	.041
Covariate linearity (r)	.94	.96	.90	.52
ANCOVA status	Primary	Caution	Caution	Caution

Note. ✓ = assumption met; ✗ = assumption violated. Where slopes were non-homogeneous (WTC, anxiety) or variances unequal (performance), inference rests on the assumption-free gain-score procedures in Table 6 rather than on the ANCOVA F.

4.2. Speaking Confidence

Both groups improved in speaking confidence from pre-test to post-test, and both within-group gains were statistically significant (Table 4). The change was larger in the AI-supported group. With the pre-test score held constant, the effect of group on post-test confidence was significant and large, $F(1, 57) = 26.48$, $p < .001$, partial $\eta^2 = .317$, and the adjusted post-test mean for the experimental group (3.67) exceeded that of the control group (3.42) (Table 5). The robustness checks agreed: the bias-corrected bootstrap interval for the between-group difference in confidence gains excluded zero, and the Mann–Whitney test on change scores was significant (Table 6). The advantage in speaking confidence therefore held across parametric, resampling, and rank-based procedures.

4.3. Willingness to Communicate, Anxiety, and Oral Performance

The remaining outcomes patterned differently. Willingness to communicate rose in both groups and favoured the experimental group in the covariance analysis, $F(1, 57) = 19.27$, $p < .001$, partial $\eta^2 = .253$, with adjusted means of 4.01 against 3.80. The advantage was corroborated by a significant independent-samples comparison of gain scores and a bootstrap interval that excluded zero, yet the rank-based Mann–Whitney test on the same gains did not reach significance ($p = .095$). The willingness advantage therefore held under the parametric and resampling tests but attenuated under the rank-based check.

Oral performance followed a similar but more moderate pattern. Both groups improved, with a significant group effect in the covariance analysis, $F(1, 57) = 7.96$, $p = .007$, partial $\eta^2 = .123$, and adjusted means of 3.04 for the experimental group against 2.88 for the control group; the bootstrap interval for the gain difference excluded zero, while the Mann–Whitney test was again borderline ($p = .077$). The advantage was not distributed evenly across the rubric (Table 7). The experimental group showed a significant gain in vocabulary, with an average of +0.33 for them versus +0.08 for the control group. They also performed very well in the fluency dimension, with 0.28 as their score against the control group's score of 0.13. Grammar and pronunciation of English improved only slightly. It was the comprehension skill in English that witnessed the least difference as an increase of 0.17 for the control group and 0.22 for the experimental group.

Speaking anxiety behaved quite differently. Both groups experienced a modest decrease in anxiety; however, the within-group decrease was not significant in either case. The group difference was also absent according to the covariance analysis ($F(1, 57) = 0.06$, $p = .804$, partial $\eta^2 = .001$), with adjusted means that were practically identical (3.39 and 3.38). The bootstrap interval covering the difference in the level of anxiety change crossed zero and

therefore did not provide any convincing evidence in favor of one or the other intervention. In addition, the comparison of anxiety levels through the Mann-Whitney test was non-significant. Across every analysis, the AI-supported condition produced no differential effect on anxiety. The anxiety scale was internally consistent in this sample (Table 1), so the null is not attributable to unreliable measurement.

Table 4. *Within-group pre–post comparisons (paired-samples t-tests)*

Measure / Group	t(29)	p	Cohen's d	Direction
Speaking confidence – Exp.	5.58	<.001	1.02	Increase
Speaking confidence – Ctrl.	4.57	<.001	0.83	Increase
WTC – Exp.	3.85	<.001	0.70	Increase
WTC – Ctrl.	3.53	.001	0.64	Increase
Speaking anxiety – Exp.	−1.68	.104	−0.31	n.s.
Speaking anxiety – Ctrl.	−1.63	.113	−0.30	n.s.
Oral performance – Exp.	5.54	<.001	1.01	Increase
Oral performance – Ctrl.	4.12	<.001	0.75	Increase

Table 5. *ANCOVA on post-test scores with pre-test as covariate*

Measure	Adj. M Exp.	Adj. M Ctrl.	F(1,57)	p	η^2p
Speaking confidence	3.67	3.42	26.48	<.001	.317
Willingness to communicate	4.01	3.80	19.27	<.001	.253
Speaking anxiety	3.39	3.38	0.06	.804	.001
Oral performance	3.04	2.88	7.96	.007	.123

Table 6. *Between-group comparison of change scores with robustness checks*

Measure	Gain Exp.	Gain Ctrl.	Cohen's d	Bootstrap 95% CI / M–W p
Speaking confidence	+0.32	+0.13	0.77	[+0.07, +0.31]; p = .017
Willingness to communicate	+0.21	+0.07	0.60	[+0.03, +0.26]; p = .095
Speaking anxiety	−0.15	−0.07	−0.20	[−0.27, +0.12]; p = .063
Oral performance	+0.27	+0.14	0.56	[+0.01, +0.25]; p = .077

Note. Bootstrap intervals use 10,000 bias-corrected resamples of the between-group difference in change scores; M–W p is the two-sided Mann–Whitney U test on change scores. A confidence interval excluding zero indicates a reliable between-group difference.

Table 7. *Oral-performance gains by rubric dimension and group*

Rubric dimension	Exp. pre→post	Exp. gain	Ctrl. pre→post	Ctrl. gain
Fluency	2.77 → 3.05	+0.28	2.75 → 2.88	+0.13
Pronunciation	2.78 → 3.03	+0.25	2.75 → 2.90	+0.15
Vocabulary	2.77 → 3.10	+0.33	2.80 → 2.88	+0.08
Grammar	2.72 → 3.00	+0.28	2.70 → 2.88	+0.18
Comprehensibility	2.87 → 3.08	+0.22	2.63 → 2.80	+0.17

Note. Each dimension was rated 1–5 by two raters and averaged. Gains are post-test minus pre-test composite means.

4.4. Learner Accounts of AI-mediated Rehearsal

Interviews were conducted with eight learners from the experimental group. Four themes recurred across the accounts.

The first and most prominent theme was the value of a non-judgmental rehearsal space. Learners repeatedly described the freedom from peer evaluation as the feature that changed how they practised. One explained that the absence of a watching face and the certainty of

not being laughed at made him willing to attempt difficult structures and unfamiliar vocabulary, so that his trial and error became faster. Another said simply that not having to feel embarrassed in front of the machine made her braver to try, even when unsure of her grammar. Several others echoed that being able to make mistakes privately, and to be corrected without an audience, was what lowered the fear of speaking.

The second theme was rehearsal and the transfer of preparedness to live performance. Learners described a routine of generating a scenario, taking the nursing role while ChatGPT played the patient, requesting correction, and repeating the exchange until they felt ready. They then connected this rehearsal to how they felt when performing in class. One reported that, having simulated the situation beforehand, he came to the classroom role-play with a stock of vocabulary and sentence patterns and felt the talk flow more easily by mid-semester. Another recalled being nervous the first time she performed before classmates but becoming accustomed to it after repeated practice.

The third theme was an unblinking acknowledgment of the limitations of mechanical interaction. Users who were highly satisfied with the system were still of the view that trying out a chatbot is quite different from having a real conversation with one's interlocutor. A user noticed that communication with AI was more like one-way and roundabout, whereas a discussion with a classmate would enable genuine expression; A student who valued face-to-face interactions as human partners' means of expression with their natural flow of conversation, body language and facial expressions argued that the best possible way would be combining some chatbot rehearsals with actual speaking. A minority voice was more sceptical still: one learner reported recurring frustration when the model's answers drifted from what had been asked and said she did not want to depend on such tools.

The fourth theme was critical evaluation of AI output. Instead of taking for granted the answers generated by the model, the students explained that they verified the uncertain language against their lecturer's feedback, textbooks, and other sources, and re-prompted when the answer appeared to be wrong. One said she compared the model's suggestions with course material and asked a classmate or lecturer when still unsure; another stressed that AI should be a learning aid rather than a sole source, and that practice with people remained necessary.

Across the four themes, learners located the benefit of AI-mediated rehearsal in private, repeatable practice and located its limit at the point where practice becomes live interaction with people. How this maps onto the quantitative pattern is taken up in the discussion.

5. Discussion

The central result of this study is a dissociation among affective outcomes that are too often bundled together. Augmenting role-play with structured ChatGPT interaction produced a large and robust advantage in speaking confidence, a real but less stable advantage in willingness to communicate and rated oral performance, and no advantage at all in speaking anxiety. The different performances cannot be accounted for by extra practice or novelty, since the groups performed the same communicative role-plays for the same duration. This suggests that something specific about the rehearsal that the AI component provided, other than extra practice or novelty, is responsible for the diverging results.

Both theoretical lenses make sense of this. Sociocultural theory frames the chatbot as a mediational means that scaffolds performance within the learner's zone of proximal development (Vygotsky, 1978), while the interaction hypothesis treats the dense, feedback-rich, output-generating turns of a chat exchange as development-supporting interaction (Long, 1996; Swain, 1995). Both predict that the appraisal tied most closely to repeated, low-threat practice, namely confidence, should respond first and most strongly, which is what researchers found. By presenting the main subject matter of an interview as a non-judgmental environment, which is actually the way the whole thing functions, the interviewer can take it one step further: if the price of committing a mistake is practically nil, students will make more attempts, be corrected more, and will gradually develop a sense of self-reliance and a kind of own capacity. These echo reports that AI tools lower the fear of attempting language (Ghafouri, 2024; Teng, 2024; Tram et al., 2024); our contribution is to isolate confidence as a measured outcome under an active control rather than infer it from perceptions.

The willingness-to-communicate result both agrees with and complicates prior work. As in Fathi et al's. (2024), the AI group became more willing to communicate; unlike a clean replication, however, the effect thinned under the rank-based test. The interviews explain why. Learners drew a sharp line between willingness to speak with a machine and willingness to speak with people, several insisting that human spontaneity and social presence could not be reproduced by a chatbot. Wang et al. (2024) likewise found conversational agents acting unevenly on willingness. If willingness is a situated, volitional readiness (MacIntyre, 2007), machine-mediated practice may raise it within that situation while transferring only partly to the face-to-face classroom, a reading that fits both the borderline statistic and the learners' own reservations.

The anxiety null is worth dwelling on. It sits awkwardly beside studies reporting that AI lowers speaking anxiety, yet the contrast is softer than it looks: reviews find anxiety the least consistent affective outcome of AI use, shifting with some tools and contexts but not others (Klimova et al., 2024; Wang et al., 2024), and the two studies closest to ours in design report the same absence or even a reversal (Ballıdağ & Aydın, 2025; Çakmak, 2022). We found two ways to read our data, and they do not exclude each other. First, anxiety as Horwitz et al. (1986) theorised it is a comparatively stable apprehension that an eight-week change may be too brief to move, especially when the high-stakes event, performing before classmates, stayed in both conditions; the learner who felt safe rehearsing yet still froze the first time she faced the class captures this exactly. Second, the comfort of the AI setting may dampen state apprehension during rehearsal without touching the trait-like anxiety the scale measures. Either way, the data warn against assuming that whatever raises confidence must lower anxiety. Here the two simply did not co-vary with condition, and reporting the confidence gain without the anxiety null would have misrepresented the intervention's reach.

Rated oral performance moved in step with confidence, which matters because it grounds the affective gains in observable language. Fathi et al. (2024) likewise found AI-mediated speaking tasks outperforming conventional instruction on oral skills, and the sub-skill profile reported above sharpens that picture. The dimensions that gained the most, vocabulary and fluency, are the ones that repeated, correction-rich text rehearsal is best placed to build; the dimension that gained least, comprehensibility, depends on live spoken delivery the chatbot never trained.

Rating oral performances changed in parallel with confidence since this feature is relevant as it anchors emotional development in physical changes of speech production. Similarly, Fathi et al. (2024) reported that the use of AI tools in spoken exercises leads to better results than conventional teaching methods, which means focusing more on the oral sub-skill, as reported above, enhances clarity of the whole picture. The two dimensions that developed most rapidly, "vocabulary" and "fluency," happen to be the ones where repetitive and highly corrective text rehearsals are the most suitable; "comprehensibility," though, only developed to a very small extent, and it is quite dependent on being able to speak live while a chatbot is always the only interlocutor. Read against the interaction hypothesis, this is what one would expect: abundant turns and immediate feedback enlarge a learner's usable repertoire and speed its retrieval, but they do not rehearse the prosody and real-time coordination of speech. The performance data thus carry the same message as the affective data. The tool strengthens what can be rehearsed in text and stops short of what cannot.

The divergence between the parametric and rank-based results is itself informative. Where the assumptions held, as for confidence, all three approaches agreed, and the large effect can be asserted with some security. Where they failed, a covariate-adjusted F can overstate a difference that distribution-free tests then decline to confirm, which is the situation for willingness, anxiety, and performance. The borderline Mann–Whitney results therefore correct the covariance analysis rather than contradict it, since the conservative verdict is the safer one when assumptions are not met. Confidence rests on firm ground; willingness and performance are probably real but smaller and less certain than a lone ANCOVA would imply; anxiety did not move.

Taken together, the four outcomes describe a layered effect. The intervention acted most directly on learners' appraisal of their own capacity, less on their readiness to communicate and observable performance, and not at all on dispositional anxiety. That ordering is coherent: mediated rehearsal should shape beliefs about capability before a situated willingness that depends on the social setting, and well before a stable disposition. The lesson for the field is methodological. Studies that fold these constructs into one affective index, or that pick anxiety as the lone affective outcome, risk reporting muted or null effects simply because the movement sits in confidence. Disaggregating the affective domain is what makes it possible to see what AI-mediated practice actually does.

Beyond this tool and this outcome, the result bears on how AI-assisted language learning should be theorised. Mediation accounts of technology in second language acquisition have tended to ask whether a tool supports learning; the more tractable question is which psychological process a given mediational means can plausibly reach. Conversational agents of the text-based variety are able to present users with an unlimited number of private correction-laden interaction turns. It is these features that allow the development of appraisal through the repetition of successful attempts that build mainly self-efficacy, rather than dispositions formed over a long period which are tied up with social evaluation, like trait-like anxiety. Viewed in that light, the dissociation detected in this study is not just any deviation to be justified but is a consequence that naturally flows from treating the affordances as real factors: the scope of a gadget is what the affordances of that tool actually provide a person with. The principle generalises past ChatGPT. It anticipates that voice-enabled agents, which do rehearse prosody and real-time coordination, should reach outcomes text-based agents cannot, and that agents deployed in socially exposed settings should behave differently

again. It also suggests a reading of the heterogeneity that recent meta-analyses report as their most consistent finding (Li et al., 2025; Lyu et al., 2025; Wu & Li, 2026): pooled averages taken across tools with different affordances and outcomes with different degrees of malleability may be averaging over precisely the distinctions that matter. For theory building, the implication is that the field needs process-level models specifying tool affordance, target construct, and time-scale together, in place of global claims about whether AI helps.

The pedagogical implication is specific. Conversational AI suits the rehearsal phase that precedes performance, where its private, correction-rich interaction builds the confidence anxious learners need before facing an audience; it is far less obviously a substitute for that audience. The learners themselves came up with the reasonable principle: the use of AI rehearsal to get oneself ready for human interaction in reality, but not to substitute such interaction. Considering Indonesian vocational programmes like nursing, where speaking practice is limited and communicating at the workplace is what one is aimed at, using AI to teach learners the scenario through role-play and keeping classroom time for that human interaction, which AI is not capable of delivering, results in the most suitable and efficient use of AI and human interaction. In fact, the critical-use guidance reflected in the interviews teachers had with AI could become a valuable teaching resource as part of what distinguishes this tool from a crutch it is verifying AI's output against authoritative sources.

While enthusiasm may naturally come, one should not be carried away and forget about mistakes made, since the results could not be accepted unanimously by the learners. One reported steady frustration when the model missed her question and decided she would rather not rely on it; others caught it offering stiff, over-formal phrasing, and several routinely checked its output against textbooks and lecturers before trusting it. These complaints matter. Generative models produce fluent text whether or not it is accurate, and a learner who cannot yet judge the language is poorly placed to catch a confident error, which raises a real risk of quietly internalising mistakes. There is also an equity dimension a single-site study cannot settle, since dependable practice presumes a stable device and data connection, resources unevenly distributed among Indonesian students. A contrasting opinion serves as a helpful reminder: AI was a great help to most of the students in making them more confident; however, it is still an unreliable partner, and its benefit mostly relies on the learner being critically questioning and on the instructor presenting it as a step towards independence, rather than as the end, which the learner needs to trust.

Several limitations bound these claims. The design was quasi-experimental with intact groups, so despite baseline checks and covariate adjustment, unmeasured group differences cannot be ruled out. The study sample was collected from just one nursing school with mostly women involved; hence the results cannot be generalised across genders, disciplines, or different institutions. The checking of the adapted scales was carried out mainly through content-oriented review and internal consistency, and not via factor-analytic confirmation, which a sample of that size would not have supported. The affective measures were self-reported and, as such, subject to respondent tendencies, and only a two-and-a-half-month period would not be able to show if the confidence advantage would be lasting or whether anxiety might fluctuate within that longer timeframe. One limitation is specific to the tool as used here: rehearsal was text-based, so learners typed rather than spoke. That gap is not cosmetic. Typed practice trains lexical and syntactic retrieval but bypasses pronunciation, intonation, listening, and the real-time pressure of speech, the most likely reason

comprehensibility gained least and anxiety held flat; claims about "speaking" built on text rehearsal must be read with that caveat. Lastly, the interviews focused on eight learners from the experimental group only, which made it evident that it was just explaining the internal processes of that group, whereas comparing different groups was not the idea. In the future, research may be carried out that uses random assignment of participants into the same active-control condition, tracks their development over a long enough time to check how durable the changes are and whether the new communication skills are transferable to job-related professional communication. Additionally, if one of the participants requested it to do so, researchers might close the pronunciation and spontaneity gaps by using voice-based AI, which are left unaddressed by text rehearsal alone.

6. Conclusion

The combination of traditional role-plays in the classroom and ChatGPT-assisted rehearsals has led to remarkable improvements in English-speaking confidence in EFL nursing students compared to a situation where only role plays are used. Although the willingness to communicate and the perceived performance by the students were also enhanced as a result, the effects were not as great as the increase in speaking confidence. The researchers also failed to notice any influence of this intervention technique on the students' speaking anxiety. One of the ways this study is original is that it demonstrates that AI is capable of triggering emotion-related change mainly in self-confidence, which is the feeling of being able to carry out a speaking task successfully. Theoretically, this suggests conversational AI as both a mediational and interactional tool that increases self-assessment by enabling practice in a very low-risk environment; it also presents the view that the emotional domain is different and, therefore, shouldn't be treated as one thing only. In terms of teaching, this result places AI as a rehearsal tool that supports, but does not replace, human-to-human interaction at the level of the student, which is exactly the students' perception. As far as a curriculum, the result is more specific than a general invitation of using AI. In Indonesia, English courses for professional purposes tend to set aside only a limited number of contact hours per week for a subject that, nevertheless, has to train graduates for their workplace communication. In this context, the limiting factor is the time spent in rehearsal, not the quality of instruction. When AI is integrated as another component of the work cycle that comes before a human role play is staged, a resource issue changes into one of the order of steps. This change implies a commitment to specific aspects of the curricula such as the teaching of prompting techniques, checking of AI outputs, and continued emphasis on assessment criteria based on live interactive skills rather than rehearsed accuracy. Furthermore, attention must be devoted to the disparity in access to equipment and internet, which determines students' ability to practice. If a program treats AI merely as an additional feature, it will probably be difficult to prove its worth. The ones that include AI as part of the task cycle, however, with the human interaction as the final stage, have sound reasons for doing so, if not just that. The most reasonable way of moving ahead would be to conduct large-scale double-blind experiments that record learners' spoken conversations, in which researchers can see whether the increased confidence learners experience during AI rehearsals actually results in a successful interaction in the workplace, which the curriculum is aiming to equip them with.

References

Ballıdağ, M., & Aydın, S. (2025). A comparison of the effects of AI-based chatbots and peer interactions on speaking anxiety among EFL learners. *Future in Educational Research*, *3*(2), 224–238. https://doi.org/10.1002/fer3.70005

Bibauw, S., François, T., & Desmet, P. (2019). Discussing with a computer to practice a foreign language: Research synthesis and conceptual framework of dialogue-based CALL. *Computer Assisted Language Learning*, *32*(8), 827–877. https://doi.org/10.1080/09588221.2018.1535508

Burhan, K., & Putra, H. (2021). Developing English materials for nursing students through students' need analysis. *Journal of English Education and Teaching*, *5*(3), 413–424. https://doi.org/10.33369/jeet.5.3.413-424

Çakmak, F. (2022). Chatbot-human interaction and its effects on EFL students' L2 speaking performance and speaking anxiety. *Novitas-ROYAL (Research on Youth and Language)*, *16*(2), 113–131.

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, *20*, 22. https://doi.org/10.1186/s41239-023-00392-8

Dwiyanti, A., & Lolita, Y. (2023). The Effectiveness of Role Play in Improving Speaking Skill of EFL Students. *Prosodi*, *17*(2), 138–148. https://doi.org/10.21107/prosodi.v17i2.17064

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

Fryer, L., & Carpenter, R. (2006). Bots as language learning tools. *Language Learning & Technology*, *10*(3), 8–14.

Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and Human task partners. *Computers in Human Behavior*, *75*, 461–468. https://doi.org/10.1016/j.chb.2017.05.045

Ghafouri, M. (2024). ChatGPT: The catalyst for teacher-student rapport and grit development in L2 class. *System*, *120*, 103209. https://doi.org/10.1016/j.system.2023.103209

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, *27*(1), 70–105. https://doi.org/10.1080/09588221.2012.700315

Guo, K., Li, Y., Li, Y., & Chu, S. K. W. (2023). Understanding EFL students' chatbot-assisted argumentative writing: An activity theory perspective. *Education and Information Technologies*, *29*(1), 1–20. https://doi.org/10.1007/s10639-023-12230-5

Guo, K., Pan, M., Li, Y., & Lai, C. (2024). Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability. *The Internet and Higher Education*, *63*, 100962. https://doi.org/10.1016/j.iheduc.2024.100962

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, *70*(2), 125–132. https://doi.org/10.1111/j.1540-4781.1986.tb05256.x

Katemba, C. V, & Grace, R. M. (2023). Role-playing improves speaking proficiency skills. *Acuity: Journal of English Language Pedagogy, Literature and Culture*, *8*(2), 244–268. https://doi.org/10.35974/acuity.v8i2.3153

Klimova, B., Pikhart, M., & Al-Obaydi, L. H. (2024). Exploring the potential of ChatGPT for foreign language education at the university level. *Frontiers in Psychology*, *15*, 1269319. https://doi.org/10.3389/fpsyg.2024.1269319

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for Language Teaching and Learning. *RELC Journal*, *54*(2), 537–550. https://doi.org/10.1177/00336882231162868

Lee, J. S., & Drajati, N. A. (2019). Affective variables and informal digital learning of English: Keys to willingness to communicate in a second language. *Australasian Journal of Educational Technology*. https://doi.org/10.14742/ajet.5177

Li, M., Wang, Y., & Yang, X. (2025). Can generative AI chatbots promote second language acquisition? A meta-analysis. *Journal of Computer Assisted Learning*, *41*(4), e70060. https://doi.org/10.1111/jcal.70060

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press.

Lyu, B., Lai, C., & Guo, J. (2025). Effectiveness of chatbots in improving language learning: A meta-analysis of comparative studies. *International Journal of Applied Linguistics*, *35*(2), 834–851. https://doi.org/10.1111/ijal.12668

MacIntyre, P. D. (2007). Willingness to communicate in the second language: Understanding the decision to speak as a volitional process. *The Modern Language Journal*, *91*(4), 564–576. https://doi.org/10.1111/j.1540-4781.2007.00623.x

MacIntyre, P. D., Clément, R., Dörnyei, Z., & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal*, *82*(4), 545–562. https://doi.org/10.1111/j.1540-4781.1998.tb05543.x

Munirah, M., Syahruddin, S., & Yusuf, A. B. (2023). The development of cultural integrated Indonesian speaking e-module for higher education students in Indonesia. *Indonesian Journal of Applied Linguistics*, *13*(1), 191–203. https://doi.org/10.17509/ijal.v13i1.58266

Nurlaelawati, I., Gunawan, W., & Lengkanawati, N. S. (2022). Enhancing student participation in learning to write a recount text: Learning from EFL pre-service teachers in implementing R2L pedagogy. *Indonesian Journal of Applied Linguistics*, *12*(2), 374–384. https://doi.org/10.17509/ijal.v12i2.51086

Peng, J.-E., & Woodrow, L. (2010). Willingness to Communicate in English: A Model in the Chinese EFL Classroom Context. *Language Learning*, *60*(4), 834–876. https://doi.org/10.1111/j.1467-9922.2010.00576.x

Poedjiastutie, D., & Oliver, R. (2017). English learning needs of ESP learners: Exploring stakeholder perceptions at an Indonesian university. *TEFLIN Journal*, *28*(1), 1–21. https://doi.org/10.15639/teflinjournal.v28i1/1-21

Reinders, H., & Wattana, S. (2015). Affect and willingness to communicate in digital game-based learning. *ReCALL*, *27*(1), 38–57. https://doi.org/10.1017/S0958344014000226

Richards, J. C. (2006). *Communicative language teaching today*. Cambridge University Press.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics* (pp. 125–144). Oxford University Press.

Teng, M. F. (2024). "ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education: Artificial Intelligence*, *7*, 100270. https://doi.org/10.1016/j.caeai.2024.100270

Tram, N. H. M., Nguyen, T. T., & Tran, C. D. (2024). ChatGPT as a tool for self-learning English among EFL learners: A multi-methods study. *System*, *127*, 103528. https://doi.org/10.1016/j.system.2024.103528

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

Wang, C., Zou, B., Du, Y., & Wang, Z. (2024). The impact of different conversational generative AI chatbots on EFL learners: An analysis of willingness to communicate, foreign language speaking anxiety, and self-perceived communicative competence. *System*, *127*, 103533. https://doi.org/10.1016/j.system.2024.103533

Wu, X., Li, R. (2026). Unraveling Effects of AI Chatbots on EFL Learners' Language Skill Development: A Meta-analysis. *Asia-Pacific Edu Res 35*, 703–714. https://doi.org/10.1007/s40299-024-00853-2

Xiao, Y., & Zhi, Y. (2023). An Exploratory Study of EFL Learners' Use of ChatGPT for Language Learning Tasks: Experience and Perceptions. *Languages*, *8*(3), 212. https://doi.org/10.3390/languages8030212

Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Educ Inf Technol 28*, 13943–13967. https://doi.org/10.1007/s10639-023-11742-4

Yang, H., Kim, H., Lee, J. H., & Shin, D. (2022). Implementation of an AI chatbot as an English conversation partner in EFL speaking classes. *ReCALL*, *34*(3), 327–343. https://doi.org/10.1017/S0958344022000039