IOSR Journal of Research & Method in Education (IOSR-JRME)
e-ISSN: 2320–7388, p- ISSN: 2320-737x Volume 16, Issue 6 Ser. I (Nov. – Dec. 2026) 24-31
www.iosrjournals.org

Beyond Rehearsed Answers: AI Role-Plays for Spontaneous EFL Speaking

Trang Hong Nguyen
Master of Science in Applied Linguistics
Doctor of Education Candidate in Emerging Technologies in Education

Abstract
Conversational generative artificial intelligence can expand opportunities for speaking practice, yet a successful rehearsed exchange does not establish the ability to respond when an interlocutor changes the situation. This critical integrative review develops a pedagogical framework for using AI role-plays to support spontaneous EFL speaking. Selected research on conversational systems is interpreted alongside task-based language teaching, task repetition, interactional competence, and willingness to communicate. The resulting design connects supported oral practice, one controlled communicative complication, and a fresh interaction with a human peer. Spontaneity is defined as relevant adaptation to emerging information rather than rapid speech or absence of preparation. An illustrative hotel scenario, reusable interlocutor prompt, and observation procedure specify how teachers can preserve learner choice, fade answer-supplying support, and examine subsequent application. Existing AI role-play research precludes a claim that the technique itself is new; the proposed contribution is the explicit integration of controlled variation, documented assistance, and peer application evidence. No intervention or participant dataset is reported, and the framework's effectiveness remains untested. Future comparisons should hold practice opportunities reasonably constant, separate confidence from observed performance, and examine delayed interaction with different tasks and partners.
Keywords: generative artificial intelligence; EFL speaking; role-play; spontaneous interaction; task-based language teaching; interactional competence; transfer.

Date of Submission: 03-11-2026 Date of Acceptance: 14-11-2026

I. Introduction

The instructional problem

An EFL learner may deliver a fluent answer during practice and become uncertain when an interlocutor asks an unexpected follow-up question. The difficulty is not necessarily a lack of memorized language. The learner must identify what has changed, select a relevant response, manage the next turn, and sometimes request clarification. Speaking instruction therefore needs opportunities to move from familiar language towards adaptive use. A polished exchange with a highly supportive AI partner provides one kind of practice, but it does not automatically reveal whether that transition has occurred.

Conversational generative AI can provide readily available interlocutors, flexible scenarios, and repeated opportunities to speak. These affordances make role-play attractive where classroom time and access to additional partners are limited. However, the same system can produce the learner's answers, accept incomplete messages without clarification, or keep a scenario so predictable that success depends on following a script. A design focused only on completing an AI conversation may consequently miss the speaking behavior it intends to develop.

This concern does not make preparation or repetition undesirable. Learners need language resources and a manageable opportunity to retrieve them. The issue is how rehearsal connects to variation: whether a learner can use a familiar function when the details, partner, or communicative problem change. In this article, spontaneous speaking means producing a relevant response to emerging information under the stated conditions. It does not mean speaking immediately, avoiding useful chunks, or performing without planning, pauses, repair, or necessary accommodations.

Purpose and bounded contribution

The article proposes a three-stage pedagogical sequence: supported oral practice, a controlled complication, and fresh human-peer application. It addresses teachers and teacher educators working with adolescent or adult EFL learners, with an illustrative B1-level application. The proposed activities target speaking; they do not require a parallel learner writing assignment. Teacher notes and optional recordings serve observation rather than changing the instructional objective to written production.

Three questions guide the analysis. How should spontaneous interaction be distinguished from rehearsed fluency, confidence, and general proficiency? What can existing AI-speaking evidence justify about role-play design? How can teachers examine adaptive responses without mistaking supplied language or easy task completion for learning? The framework is a design proposal grounded in selected literature, not an empirical intervention. Its novelty is the specification of a connected teaching and evidence procedure, not the invention of AI role-play itself.

II. Theoretical and Conceptual Background

Tasks, interaction, and learner choice

Task-based language teaching emphasizes meaningful activity organized around an outcome rather than the reproduction of a predetermined language form (Ellis, 2009). A role-play becomes educationally informative when learners need to use language to resolve a problem and have some freedom over their response. If every turn is provided in advance, completing the exchange reveals little about their ability to interpret a partner's contribution. A practical scenario should therefore have a clear goal, an information gap, and more than one possible way of expressing an acceptable solution.

Swain and Lapkin (1995) show how output can bring language problems to learners' attention and prompt further processing. Their work supplies a theoretical rationale for asking learners to formulate and revise meaning, not direct evidence for contemporary AI speaking systems. The present design retains a learner attempt before offering a full model. The aim is to preserve an opportunity to notice a communicative problem and choose a response, while keeping help available when the task would otherwise become unproductive.

Repetition can prepare flexible performance

Research on fluency training and task repetition supports taking rehearsal seriously (de Jong & Perfetti, 2011). Planning and repetition can interact with accuracy, complexity, and fluency in oral production (Ahmadian & Tavakoli, 2011). These findings do not establish that repeating a fixed dialogue necessarily transfers to new interaction. They suggest that practice should be connected to later demands rather than rejected because some language becomes familiar. A reusable request such as “Is there another option?” can be a resource for adaptive speaking.

The design problem is how much variation to introduce and when. A learner who must handle new vocabulary, a new role, several unexpected problems, and time pressure simultaneously may lose access to otherwise available language. Skehan (2009) provides a basis for considering competing performance demands. This article therefore proposes changing one communicative condition at a time while retaining the target function. This is a pedagogical hypothesis for evaluation, not an established optimal level of task difficulty for every learner.

Interactional competence and willingness to communicate

Interactional competence concerns how participants manage contributions together, including turn taking, topic development, and responses to one another (Galaczi, 2014; Galaczi & Taylor, 2018). An isolated, accurate sentence is therefore insufficient evidence of effective interaction. An AI partner also shapes what can be observed: it may infer intended meaning unusually readily or behave inconsistently. A later interaction with a human peer offers a different opportunity to examine coordination, without being treated as a complete representation of real-world communication.

Willingness to communicate concerns readiness to enter communication in particular circumstances (MacIntyre et al., 1998). It is conceptually distinct from the quality of an observed response. A learner can become more willing to speak without yet managing an unexpected condition successfully; another may perform adequately while remaining anxious. Both dimensions matter, but measuring one cannot substitute for examining the other. Table 1 identifies what each construct contributes and the evidence that would be needed to interpret it responsibly.

Table 1. Construct distinctions for evaluating AI-supported speaking

Construct	Operational focus in this article	Potential evidence	Inference to avoid
Rehearsed performance	Delivery in a familiar exchange with known content.	A repeated role-play with its support conditions recorded.	Smooth repetition establishes flexible interaction.
Utterance fluency	Temporal features of speech within a defined task.	Audio-based analysis with appropriate procedures.	Faster speech necessarily communicates more effectively.
Spontaneous adaptation	A relevant response when previously undisclosed information changes the problem.	The changed condition, learner response, and available assistance.	Surprise alone creates learning or proves general proficiency.

Construct	Operational focus in this article	Potential evidence	Inference to avoid
Interactional competence	Coordination of meaning and turns with an interlocutor.	Connected turns, clarification where needed, and negotiated outcomes.	A single sentence or transcript fully represents interaction.
Willingness to communicate	Readiness or intention to communicate in a situation.	A suitable self-report measure and contextual information.	Reported confidence demonstrates speaking improvement.
Transfer	Use of a related function with a new task or partner.	Fresh and, where feasible, delayed interaction.	One successful new exchange establishes durable general transfer.

III. Review Methodology

Critical integrative design

A critical integrative review was used to connect a specific teaching problem with relevant theory and selected empirical evidence. This approach follows the distinction between conceptual synthesis and exhaustive effectiveness review discussed by Snyder (2019). Targeted source discovery and verification were undertaken on 24 October 2026. Combinations included AI or generative AI, chatbot, role-play, EFL speaking, spontaneous interaction, interactional competence, task repetition, willingness to communicate, and transfer. Foundational sources were retained where they clarified a construct or design decision.

Primary publisher texts, abstracts, and author or institutional records were used to verify bibliographic details and reported findings. Selection prioritized speaking-related research and a direct recent study of GenAI role-play. Earlier intelligent personal assistant studies were included to contextualize a longer history of conversational practice, while their technological differences were retained. Official language assessment and AI guidance informed construct interpretation and implementation safeguards. Promotional product descriptions were not treated as evidence of instructional effectiveness.

Analytical procedure and limits

The synthesis separates system type, learner context, comparison arrangement, outcome construct, and access to evidence. Table 2 records selected findings and their boundaries. It then examines whether those findings support the particular design claim under discussion: opportunity to speak, affective engagement, observed performance, or fresh application. A finding about one outcome is not relabeled as another. Research summaries remain within what the available primary text or abstract explicitly reports; intervention details unavailable in those records are not inferred.

No systematic study-flow count, independent duplicate screening, pooled effect size, or formal bias appraisal is claimed. The framework and examples were developed analytically and have not been implemented as a study. The hotel conversation, prompts, task timings, and observation questions are reproducible proposals for testing, not research instruments with established reliability or validity. The source selection can therefore support a reasoned design argument but cannot establish that the proposed sequence is superior to existing speaking instruction.

IV. Evidence Synthesis and Discussion

Conversational practice has a longer history than generative AI

Dizon (2020) reported a speaking advantage for an Alexa-supported group in a small university EFL quasi-experiment. Tai and Chen (2023) examined adolescent learners' willingness to communicate with Google Assistant. These studies show why contemporary AI speaking practice should be situated within earlier conversational technologies rather than presented as an entirely new educational activity. Their systems and outcome constructs differ from those of current generative role-play, limiting direct transfer of their findings.

Jeon et al. (2024) review research on speech-recognition chatbots and identify directions relevant to large language models. Such a review helps locate the field and its varied approaches; it is not an additional independent experiment confirming each primary finding. This distinction matters because a literature synthesis can appear more persuasive than its evidence warrants if reviews and their included studies are counted as separate replications. The present analysis uses the review for context and the selected primary studies for their stated outcomes.

Direct role-play evidence sets a boundary on novelty

Chen et al. (2025) directly compared GenAI role-play with peer role-play in a Chinese university quasi-experiment. Their publisher abstract reports no significant between-group difference in speaking-performance improvement, alongside motivational and self-efficacy benefits. The study establishes clear prior work on the

core technique. It neither warrants claiming that AI role-play is unprecedented nor demonstrates equivalence between AI and peer conditions: an absence of a significant difference is not an equivalence test.

Other findings likewise caution against a single broad effectiveness claim. Wang et al. (2024) distinguish favorable affective outcomes from nonsignificant between-group speaking differences. Sok and Shin (2025) report task-related speaking benefits for ChatGPT conditions and additional summarization benefit with specific guidance. These studies make prompt design and outcome selection relevant. They do not directly test the present sequence of one controlled complication followed by an untrained human-peer task. That proposed integration remains an empirical question.

Table 2. Selected conversational-system evidence and boundaries for the proposed design

Source	System and context	Relevant evidence	Boundary for interpretation
Dizon (2020)	Alexa; university EFL.	Speaking advantage in a small quasi-experiment.	Earlier personal assistant; limited basis for current GenAI design claims.
Tai and Chen (2023)	Google Assistant; adolescent EFL.	Willingness to communicate examined as a central outcome.	Affective readiness is distinct from spontaneous interactional performance.
Chen et al. (2025)	GenAI role-play versus peer role-play; university EFL.	Both role-play arrangements supported improvement; no significant between-group speaking difference reported.	Publisher abstract; neither AI superiority nor formal equivalence established.
Wang et al. (2024)	Conversational GenAI with different interfaces; university EFL.	Positive affective outcomes did not establish a between-group speaking advantage.	Institutional author record; perception measures are not performance measures.
Sok and Shin (2025)	ChatGPT conditions with different task guidance; university EFL.	Speaking-task benefits and a guidance-related summarization benefit.	Publisher abstract; no claim of fresh peer-transfer evidence.
Jeon et al. (2024)	Review of speech-recognition chatbot research.	Context for design variation and future LLM research.	Secondary synthesis; not a separate intervention or proof of this sequence.

Why controlled variation and peer application belong together

An unexpected condition creates a reason to adapt, but it can also make a role-play arbitrary. If an AI receptionist changes prices repeatedly or refuses every solution, difficulty reflects unstable task conditions rather than useful communicative demand. A controlled complication preserves a solvable goal and changes one relevant constraint. For example, the requested room type is unavailable, but at least one alternative satisfies the learner's essential needs. The learner must then clarify, request an alternative, or revise a preference.

Fresh human-peer application tests a different question from AI task completion: whether the learner can select the target function when the wording, details, and interlocutor change. It also exposes the limits of inferring human interaction from a chatbot exchange. Nevertheless, peers can supply language, dominate turns, or simplify a problem. Their contributions must be documented. A successful peer task is evidence about that interaction under its stated conditions, not a pure measure free from partner effects or a guarantee of communication outside class.

V. Proposed Pedagogical Framework

A three-stage sequence with an explicit target

The proposed sequence begins by selecting a communicative function and a realistic outcome. For the illustrative B1 activity, the function is requesting and evaluating an alternative when an initial option is unavailable. Learners should be able to state a relevant constraint, ask for another option, clarify necessary information, and confirm an acceptable arrangement. The goal can be achieved with incomplete but intelligible language; the task is not a disguised requirement to recite one grammatically elaborate sentence.

Stage 1 provides supported oral practice with a familiar version of the situation. Learners may hear a short model, rehearse useful chunks, and plan key information without writing a script. Stage 2 introduces one controlled complication after the exchange begins, retaining optional help. Stage 3 changes the context and partner while preserving the function, removes the displayed answer model, and observes what the learner can do. A later opportunity can examine persistence. Table 3 specifies the proposed classroom actions and the evidence available at each stage.

Table 3. Proposed sequence, assistance conditions, and observation opportunities

Stage	Teacher or interlocutor action	Learner action and permitted support	Evidence and interpretation
Set the target	Explain the goal and demonstrate the communicative function.	Hear a model; select useful chunks; ask about task meaning.	Check understanding of the goal, not independent mastery.
1. Supported practice	Offer a stable familiar role-play and optional function cues.	Practise orally; request a hint or model when needed.	Record which contributions follow supplied language.
2. Controlled complication	Reveal one undisclosed but solvable change after the opening.	Interpret the change, adapt a request, and negotiate an outcome; help remains available.	Observe adaptation and record any help or AI deviation.
Brief feedback	Discuss one successful strategy and one priority for another attempt.	Explain or demonstrate the intended response orally.	Identify a practical revision; feedback timing is a design choice.
3. Fresh peer application	Use new details and a human partner; remove the full answer model.	Reuse the function with preparation time and agreed accommodations.	Observe relevant choices without the target answer supplied.
Later check	Arrange a related task with another partner where feasible.	Respond to a further new situation under recorded conditions.	Examine persistence and the limits of transfer.

The sequence can fit a lesson segment: a teacher might allocate approximately five minutes to supported practice, six to variation, three to feedback, and six to peer application. These times are illustrative, not a tested prescription. Teachers should adjust them to proficiency, class size, available devices, and the interaction goal. If setup consumes most of the speaking opportunity, a human role card may be more feasible for that lesson. Substituting a partner protects participation; it does not establish equivalence between instructional conditions.

An illustrative hotel scenario

The learner is a guest seeking a quiet room for one night with a maximum total budget of US$60. A private bathroom is preferred but a shared bathroom is acceptable. The AI receptionist knows that standard rooms are unavailable, a quiet private-bathroom room costs US$70, and a quiet twin room with a shared bathroom costs US$55. Prices include all charges. The receptionist discloses the unavailability after the guest's opening request. The cheaper workable option remains available, and no additional unexpected problem is introduced.

This configuration separates essential constraints from preferences. The learner cannot solve the problem by accepting the first expensive offer but can negotiate a satisfactory outcome. The AI should not reveal both alternatives and compose the learner's question before the learner responds. If the meaning of “shared bathroom” is unclear, asking about it serves communication; if it is understood, repeating a clarification formula should not be required merely to satisfy an observation checklist.

An invented exchange illustrates the intended process. Receptionist: “Our standard rooms are unavailable tonight. A quiet room with a private bathroom is seventy dollars.” Learner: “My budget is sixty. Is there another quiet room?” Receptionist: “A twin room is fifty-five dollars, with a shared bathroom.” Learner: “Shared with other guests?” Receptionist: “Yes.” Learner: “That is okay. Fifty-five in total?” This example is a design illustration, not an observed transcript, a model that every learner must reproduce, or evidence that the sequence improves performance.

A reusable interlocutor prompt

The following proposed prompt makes the task conditions inspectable. “Act as a hotel receptionist in a short spoken role-play for a B1 EFL learner. The learner needs a quiet room for one night and has a maximum total budget of US$60. A private bathroom is preferred but a shared bathroom is acceptable. Standard rooms are unavailable. A quiet room with a private bathroom costs US$70; a quiet twin room with a shared bathroom costs US$55. All prices include every charge. Wait for the learner's opening request before revealing the unavailability. Keep these facts fixed and allow the affordable solution.”

“Use short, natural turns and ask at most one question per turn. Do not write or speak the learner's next answer before an attempt. If the learner asks for help, first offer a brief function cue, such as asking about another option. Give a complete example only if further help is requested, then let the learner try again. Accept pauses and intelligible incomplete language. Clarify genuine uncertainty about meaning without inventing communication failure. Do not add further surprises. Finish when the guest confirms a workable choice. After the role-play, offer one relevant strength and one suggestion. Stop if the learner asks to pause or end.”

The prompt is a specification to test, not a guarantee of system behavior. Teachers should check it with the selected service, confirm that the learner can speak and hear responses, and record tool, model/version where visible, date, modality, and relevant settings. A typed conversation is not equivalent evidence for speaking. If voice interaction is unavailable, a teacher or peer can use the same role facts orally; the record should identify that change. A generated transcript should not be assumed to represent every spoken word accurately.

Fading support without removing access

Support is reduced selectively. A full answer model can become a brief function cue, then an optional request for help. Preparation time, a vocabulary reminder unrelated to the answer, or an agreed accessibility arrangement may remain available. The distinction is between support that supplies the target response and support that makes participation possible. Removing captions or additional processing time indiscriminately can change what the task measures and create avoidable exclusion.

The sequence places detailed feedback after the exchange to protect a period of connected meaning-focused interaction. This is a design choice, not a claim that delayed feedback is universally superior. Oral feedback research supports careful attention to instructional conditions (Lyster & Saito, 2010). Teachers may intervene earlier when misunderstanding blocks the task or a learner requests help. The relevant record is what intervention occurred and how the subsequent response was produced, rather than whether a rigid protocol was followed perfectly.

Fresh peer tasks with comparable communicative demands

In a fresh train-ticket task, the learner has US$20, prefers a 9:00 departure, and must arrive before 10:00. The preferred service is sold out. A partner acting as a ticket clerk can offer an 8:10 train arriving at 9:30 for US$18 or a 10:30 train arriving at 11:50 for US$16. The learner has a solvable problem requiring an alternative, attention to a constraint, and confirmation. The peer receives the facts, while the learner receives the goal; neither needs a full dialogue script.

A later restaurant task can preserve the function with different details. Two people request a 7:00 p.m. table, need to finish by 8:00, and can spend US$20 per person. The restaurant has tables at 6:00 or 8:00, the meal takes one hour, and the set menu costs US$18 per person including charges. The 7:00 table is unavailable. The goal is an acceptable arrangement, not a fixed expression. Teachers should check vocabulary and task complexity so that differences from the hotel task do not overwhelm the intended common function.

VI. Implications for Speaking Instruction and Assessment

Observe function, assistance, and connected turns

A concise observation record can ask four questions: Did the learner respond relevantly to the changed condition? Did the learner communicate the important constraint? Did the learner seek or provide clarification where genuinely needed? Did the participants confirm a workable outcome? For each observation, the teacher records the task, the preceding interlocutor turn, and any model, hint, translation, or partner assistance. Where there was no suitable opportunity, the entry should say so rather than imply failure.

These questions are teaching indicators, not a standardized speaking scale. The Council of Europe's (2020) Companion volume offers relevant interaction and strategy descriptors, but a short local checklist cannot be converted directly into a CEFR level or an examination band. Kane's (2013) validity perspective also cautions that evidence must support the intended interpretation and use. A formative note about one communicative function should not become a high-stakes judgment of a learner's general oral proficiency.

Distinguish speech evidence from interface behavior

Speech recognition may misrepresent an intelligible response or fail to recognize an accent. An AI partner may also continue despite an unclear message, creating an impression of successful communication. Teachers should review the audio or direct observation when these issues matter and distinguish interface failure from a language difficulty. A text transcript cannot independently establish pronunciation, stress, timing, or turn coordination. The AI's automatic praise or score is likewise insufficient as the sole basis for assessment.

For routine formative use, a short teacher note may be adequate. Research involving recordings or formal testing requires clearer administration, scoring procedures, data protection, and appropriate authorization. The design does not require storing complete conversations by default. Teachers can record an anonymized example of a relevant choice and the conditions under which it occurred. This keeps evidence connected to the learning question while reducing unnecessary collection of identifiable speech or account information.

Manage anxiety, agency, and participation

Unexpected information should be disclosed as a feature of the activity, even when its precise content is withheld. Learners should know the communicative goal, the availability of help, and the option to pause. A

complication should create a manageable reason to speak rather than humiliation or an adversarial contest. Learners can help choose familiar scenarios and indicate which situations are meaningful to them. Confidence may be monitored separately from performance so that an affective benefit is visible without being overstated.

Human-centered AI guidance supports attention to inclusion and learner agency (Miao & Holmes, 2023). In practice, participation should not depend on buying a subscription or disclosing unnecessary personal information. A fictional identity and non-sensitive scenario are sufficient. Teachers need to check institutional conditions and the suitability of the selected service for their learners. When technology or access prevents participation, an oral human-role alternative should be available and its conditions documented for any later comparison.

VII. Limitations and Future Research

The proposal has not been tested

The literature selection is purposive and includes studies accessed through primary abstracts or institutional records. Findings from earlier assistants cannot be assumed to apply unchanged to generative voice systems. Research on university learners does not establish outcomes for every age group or proficiency level. The prompt, timings, scenario, and observation questions are proposed designs without implementation data. The framework therefore identifies a plausible teaching arrangement and a way to examine it, not demonstrated effectiveness.

Spontaneity is also difficult to isolate. A learner may recognize a scenario from prior experience, use well-practised chunks appropriately, or receive subtle support from a partner. A supposedly fresh task may instead be harder because of unfamiliar content. The proposed definition focuses on relevant adaptation to a recorded new condition, while acknowledging these influences. No single task fully captures interactional competence, and a successful classroom exchange does not establish durable transfer to unfamiliar settings.

Testable design propositions

The first proposition is that one controlled complication can elicit more observable adaptation than a fully predictable role-play while preserving a manageable task. The second is that fading full answer models can increase opportunities for learner formulation without requiring removal of accessibility support. The third is that fresh peer application will reveal learning needs concealed by AI-assisted success. These are testable propositions, not conclusions of the reviewed studies. Their effects may depend on proficiency, topic knowledge, interaction history, and the quality of the interlocutor behavior.

A useful evaluation could compare the proposed sequence with equally timed, well-designed speaking practice using an active alternative. Depending on the question, this could be peer role-play with the same sequence, or AI role-play without the controlled variation and planned application. These comparisons answer different questions and should not be conflated. Practice time, teacher input, task difficulty, and available help should be documented, alongside actual deviations from intended AI behavior. Allocation and analysis should match the design and account for classroom grouping where relevant.

Outcomes beyond immediate task completion

Assessment should include pre-intervention performance, fresh post-intervention interaction, and a delayed opportunity with another task or partner where feasible. Suitable recordings can support blinded rating and an examination of scoring consistency. Researchers should report relevant assistance and partner contributions, not merely total turns or automatic scores. Willingness to communicate and anxiety measures can address affective outcomes separately. Qualitative accounts can explain why learners accepted, avoided, or depended on particular forms of help.

Small classroom pilots can examine feasibility, prompt compliance, and the interpretability of observation records. They cannot support a strong causal claim merely because learners improve or report enjoyment. Larger comparisons should specify primary outcomes and analysis in advance, justify sample size, and report uncertainty, missing data, and implementation differences. Evidence of no significant difference should be interpreted within its precision; claims of equivalence require an appropriate design. A convincing evaluation would ask not only whether learners complete the activity, but what remains available when its answer-supplying support changes.

VIII. Conclusion

AI role-play offers a practical way to expand oral practice, but the educational value of a successful exchange depends on what the learner contributes and can subsequently adapt. This article proposes connecting supported practice, one controlled complication, and a fresh human-peer task. The scenario, prompt, and observation questions specify an inspectable design that preserves meaningful choice and separates provided language from learner performance. Preparation, useful chunks, pauses, and accessibility support remain compatible with spontaneous interaction.

The contribution is a bounded pedagogical framework grounded in selected theory and evidence. Existing GenAI role-play research establishes that the technique itself is not new; whether this particular integration improves flexible speaking remains untested. Its practical and research value lies in making that question answerable through clearly described tasks, recorded assistance, and outcomes that extend beyond confidence or rehearsed success.

AI assistance and data statement

Generative AI assisted manuscript drafting and language editing. No original participant data or classroom implementation results are reported. All dialogues, prompts, timings, and scenarios are proposed or illustrative. Responsibility for the final submitted manuscript rests with the named author.

References

[1] Ahmadian, M. J., & Tavakoli, M. (2011). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research*, 15(1), 35–59. https://doi.org/10.1177/1362168810383329

[2] Chen, A., Jia, J., Li, Y., & Fu, L. (2025). Investigating the effect of role-play activity with GenAI agent on EFL students' speaking performance. *Journal of Educational Computing Research*, 63(1), 99–125. https://doi.org/10.1177/07356331241299058

[3] 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

[4] de Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, 61(2), 533–568. https://doi.org/10.1111/j.1467-9922.2010.00620.x

[5] Dizon, G. (2020). Evaluating intelligent personal assistants for L2 listening and speaking development. *Language Learning & Technology*, 24(1), 16–26. https://doi.org/10.64152/10125/44705

[6] Ellis, R. (2009). Task-based language teaching: Sorting out the misunderstandings. *International Journal of Applied Linguistics*, 19(3), 221–246. https://doi.org/10.1111/j.1473-4192.2009.00231.x

[7] Galaczi, E. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics*, 35(5), 553–574. https://doi.org/10.1093/applin/amt017

[8] Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, 15(3), 219–236. https://doi.org/10.1080/15434303.2018.1453816

[9] Jeon, J., Lee, S., & Choi, S. (2024). A systematic review of research on speech-recognition chatbots for language learning: Implications for future directions in the era of large language models. *Interactive Learning Environments*, 32(8), 4613–4631. https://doi.org/10.1080/10494820.2023.2204343

[10] Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. https://doi.org/10.1111/jedm.12000

[11] Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition*, 32(2), 265–302. https://doi.org/10.1017/S0272263109990520

[12] 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

[13] Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. https://doi.org/10.54675/EWZM9535

[14] Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510–532. https://doi.org/10.1093/applin/amp047

[15] Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

[16] Sok, S., & Shin, H. W. (2025). Do interactions with ChatGPT influence L2 learners' oral speaking ability, summarization ability, and perceptions of generative AI tasks? *TESOL Quarterly*, 59(S1), S19–S51. https://doi.org/10.1002/tesq.70001

[17] Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics*, 16(3), 371–391. https://doi.org/10.1093/applin/16.3.371

[18] Tai, T.-Y., & Chen, H. H.-J. (2023). The impact of Google Assistant on adolescent EFL learners' willingness to communicate. *Interactive Learning Environments*, 31(3), 1485–1502. https://doi.org/10.1080/10494820.2020.1841801

[19] 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