Curriculum and Teaching Methodology (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/curtm.2026.090520
ISSN 2616-2261 Vol. 9 Num. 5

Integrating AI into Student-Centered Speaking Assessment Design for Large EFL Classes

Fengjunzi Wang*

Department of Foreign Languages, Hubei Business College, Xiongchu Avenue No. 632, Hongshan District, Wuhan, 430079, Hubei Province, China
wangfengjunzi@hbc.edu.cn
**Corresponding author*

Keywords: Artificial intelligence; Speaking assessment; Learning-oriented assessment; College English; Formative feedback

Abstract: Speaking remains the least-assessed skill in Chinese College English: large classes and the absence of an oral component in high-stakes examinations leave most non-English majors without routine speaking evaluation. Drawing on design research methodology, this paper develops an AI-integrated, student-centered speaking assessment in which Doubao, ByteDance's conversational AI, serves as scenario interlocutor, simulated rater against the four IELTS criteria, and immediate feedback generator, complemented by a teacher analytic checklist within a dual-track feedback system. A one-semester pilot quasi-experimental questionnaire study (experimental n = 127, control n = 128) examined whether the system could operate in routine teaching: the experimental class completed weekly AI-mediated assessment sessions, and students responded positively across all seven perceived speaking-gain dimensions (all $p < .001$, $d = 0.50$–0.67), with high dialogue-extension and feedback-revision rates, indicating that the assessment process was operable at scale and accepted by students. The study offers application-oriented universities a low-cost, replicable pathway for embedding speaking assessment into College English curricula.

1. Introduction

Assessment powerfully steers what teachers teach and what learners learn [1]. The digital-intelligence empowerment of education and the rapid maturation of artificial intelligence (AI) have created unprecedented conditions for language assessment: speech-enabled large language models can now sustain spoken dialogue, produce immediate written feedback, and run on ordinary student devices at near-zero marginal cost. Yet within College English—the compulsory English course for non-English majors in Chinese universities—speaking remains the most weakly assessed skill, because the high-stakes examinations shaping the curriculum privilege listening, reading, and writing, while ordinary classroom hours cannot support routine oral testing.

The problem is especially acute in application-oriented universities, such as the institution in Hubei Province where the present study was conducted: College English is delivered in classes of forty to sixty staffed by a single teacher, so interviewing, rating, and feeding back to every student

individually would consume more time than the syllabus affords. Meanwhile, the spread of blended teaching—combining in-class instruction with online, out-of-class learning on students' own devices—offers a structural interface through which AI tools can enter routine teaching, allowing speaking practice and assessment to migrate outside class hours without forfeiting teacher oversight.

Three bodies of prior work inform such an effort, yet each leaves a gap. First, research on automated speaking assessment has concentrated on validating machine scoring for high-stakes standardized tests [2]; the resulting systems are proprietary and neither accessible nor adaptable for ordinary classroom teachers. Second, studies of conversational AI in English as a foreign language (EFL) learning have documented improved speaking performance, greater willingness to communicate (WTC), and reduced anxiety when chatbots serve as practice partners [3-6], but they examine learning effects rather than assessment design, rarely specifying how AI judgments should be produced, checked, or used. Third, classroom-based formative assessment possesses a mature theoretical apparatus, most notably learning-oriented assessment (LOA) [1,7], and Chinese university students respond positively to formative practices [8]; however, operable designs that integrate AI tools into routine speaking assessment for large non-English-major classes remain scarce.

The present study addresses this intersection by designing conversational AI into a complete, learning-oriented classroom speaking assessment. The tool adopted is Doubao, ByteDance's conversational AI, chosen for its free voice-call function and configurable agent platform. One fact must be stated at the outset: Doubao has no official, standardized speaking assessment engine. Its scoring is prompt-driven, large language model (LLM)-simulated rating—the teacher prompts the model to act as an International English Language Testing System (IELTS) examiner and rate performances against the four IELTS criteria—and it offers no phoneme-level acoustic pronunciation evaluation. The design treats these constraints as structuring conditions, answered by teacher–AI dual rating in which teacher judgment remains decisive and by explicitly bounded validity claims.

Following a design-research logic, the paper proceeds in four steps. It first reviews the three bodies of literature that inform the design (Section 2) and derives test specifications from the instructional context (Section 3). It then constructs the complete assessment scheme—a three-phase workflow, unit-anchored scenario tasks, classroom-adapted IELTS criteria, a teacher-authored agent prompt, and a dual-track feedback loop (Section 4). Next, it reports a one-semester pilot quasi-experimental questionnaire study comparing experimental and control classes (Section 5) and evaluates the design against five classical test qualities (Section 6), before delimiting its boundary conditions and future research agenda (Section 7). The study thus contributes a replicable design, an honest account of the validity boundaries of prompt-driven LLM scoring, and preliminary in situ evidence that the system operated in routine teaching and was accepted by students.

2. Literature Review

Three strands of prior work bear on such a design: classroom-based and learning-oriented assessment, automated scoring and AI rating, and conversational AI as speaking partner and feedback source. Each is reviewed below for what it contributes and what it leaves open.

2.1 Classroom-Based and Learning-Oriented Speaking Assessment

Classroom-based assessment differs from standardized testing in kind: continuous, embedded in instruction, oriented toward individual growth. Learning-oriented assessment (LOA) offers its most developed theoretical account, organized around assessment tasks as learning tasks, active student involvement, and timely forward-looking feedback [7], extended to second language classrooms as

an ecological model spanning task, learner, and context [1]. For speaking specifically, classroom assessment must elicit ratable performances within ordinary teaching time [9]. In Chinese universities, formative assessment is enacted within a strong examination culture, yet teachers who engage students in assessment-for-learning practices report gains in learner agency [10], and students view self- and peer assessment as supporting self-regulation [8]. Digital tools have also been shown to make formative speaking evaluation more feasible [11]. The remaining gap is operational: these frameworks and small-scale studies rarely specify a complete, repeatable instrument through which one teacher can assess forty to sixty students' speaking each unit—still less one that integrates AI.

2.2 Automated Speaking Assessment and AI Scoring

Research on the automated scoring of spoken and written language offers a complementary perspective. Validation studies of automated speaking scoring report encouraging agreement with human raters and stable reliability across administrations [2], and technology-mediated assessment alters both the construct and the conditions of score interpretation [12]. The advent of LLMs has extended scoring research to generative systems: ChatGPT has graded IELTS writing tasks with acceptable reliability [13], and a GPT-based chatbot scoring EFL speaking examinations against standardized criteria achieved high agreement with expert raters, although most students still preferred a live examiner [14]. Cautions accompany this promise. Validation theory struggles to keep pace with threats to construct representation, fairness, and explainability as AI enters assessment [15], and commercial automatic speech recognition (ASR) exhibits systematic disparities across speaker groups [16]. The remaining gap is one of guidance: an ordinary teacher configuring a general-purpose conversational AI, not a validated engine, as a classroom rater, with safeguards proportionate to that tool's evidentiary limits, has no model to follow.

2.3 Conversational AI as Speaking Partner and Feedback Source in EFL

A rapidly expanding literature examines conversational AI as a speaking partner and feedback source for EFL learners. Systematic reviews indicate that chatbots can reduce speaking anxiety and raise engagement [3], and quasi-experimental work reports improved speaking performance and WTC after AI-mediated interaction [4], with affective benefits recurring among Chinese EFL learners [6] and varying across chatbot designs [17]. As a feedback source, AI speech evaluation programs with automatic feedback can develop speaking skills and self-efficacy [18], yet AI-generated feedback tends to be broader but shallower than teacher feedback [19]. These studies establish conversational AI as a viable interlocutor and a useful, if imperfect, feedback source. The remaining gap is architectural: AI figures as a practice partner or feedback provider in isolation, seldom as one coordinated role within a classroom assessment system in which teacher moderation is designed in from the start.

Taken together, these three gaps converge on a single need: a complete, operable design that embeds a general-purpose conversational AI into classroom-based speaking assessment for large EFL classes, with scoring safeguards proportionate to the tool's limits and teacher moderation built in from the start. The remainder of this paper develops such a design and examines whether it can operate in routine teaching.

3. Context and Test Specifications

The operational gap identified above is addressed first at the level of specification, since test specifications are the "recipe and blueprint" of test construction [20]. The target test takers are first-

year, non-English-major undergraduates aged 18–19 at an application-oriented university in Hubei Province, China, enrolled in compulsory College English classes of 40–60 students. Their first language (L1) is Mandarin Chinese, and they are learners of English as a foreign language (EFL); proficiency at matriculation sits at approximately Levels 4–5 of China's Standards of English Language Ability (CSE), broadly comparable to A2–B1 on the Common European Framework of Reference for Languages (CEFR) [21] and, under the official CSE–IELTS alignment, to IELTS Speaking Bands 5–5.5 [22]. Instruction follows the 48-contact-hour, three-credit compulsory College English course for first-year non-English majors (code GRC213611), which is built on two paired textbooks—*New Horizon College English: Reading and Writing 1* (4th ed. [23]) and *New Horizon College English: Viewing, Listening and Speaking 1* (4th ed. [24])—and every assessment task is anchored in the current unit's theme and target language. In this non-immersion setting, one teacher cannot elicit, observe, and rate individual performances in a class of forty or more—the practical reason for introducing Doubao into the course.

Table 1 Test Specifications Summary for the AI-Mediated Classroom Speaking Assessment

Category	Specification
Test purpose	Unit-based achievement, four-criterion diagnosis, and formative feedback within learning-oriented assessment
Test takers	Non-English-major freshmen (aged 18–19) at an application-oriented university in Hubei Province; Mandarin Chinese as L1; CSE 4–5 / CEFR A2–B1; classes of 40–60
Textbook and materials	Tasks anchored in units of *New Horizon College English: Reading and Writing 1* and *Viewing, Listening and Speaking 1* (4th ed. [23,24]); on-screen digital role card per unit scenario
Task format	One-to-one student–AI voice-call session adapting the IELTS three-part format: warm-up Q&A → scenario role-play (core task) → follow-up discussion
Duration	Approximately 8–10 minutes per student (warm-up ≈2 min; role-play ≈3 min; discussion ≈3–4 min)
Scoring	Four-criterion analytic rubric adapted from IELTS Band 5–7 descriptors; Doubao's prompt-driven simulated scores moderated by a teacher analytic checklist; teacher score decisive
Feedback	Dual-track: immediate AI report (sub-scores, error corrections, suggestions) plus teacher checklist, followed by self-/peer-comments and targeted re-assessment

The assessment serves three interlocking purposes consistent with LOA [1,7]: an *achievement* measure drawing on the unit's lexical, grammatical, and functional targets; a *diagnostic* profile against the four IELTS criteria; and a *formative* loop returning immediate, criterion-referenced, teacher-moderated feedback. Because Doubao has no standardized speaking assessment engine—its scores are prompt-driven, LLM-simulated judgments against the IELTS criteria, with no phoneme-level acoustic pronunciation evaluation—these purposes are bounded accordingly, and teacher–AI dual rating is built in from the start (Section 4). Table 1 summarizes the specifications, adapting McNamara's framework [20] to this context.

The most consequential row in Table 1 concerns scoring: because Doubao's ratings are LLM-

simulated rather than psychometrically validated, they serve a formative, advisory function, and the teacher's judgment is decisive for any reported grade. The remaining rows encode the feasibility logic: an 8–10-minute session per student, impossible for one teacher to run serially across 40–60 students, becomes routine once the interlocutor load is distributed across students' own devices within the blended course structure; unit anchoring keeps the assessment inside the syllabus; and the dual-track row pairs every AI report with human review, so immediacy at scale does not mean unmonitored machine judgment.

4. Design of the AI-Integrated Assessment

4.1 Overall Workflow: A Three-Phase Cycle

Operationalizing these specifications, the assessment runs as a three-phase cycle across each unit. In Phase 1, *in-class trialing*, the scenario task is rehearsed as an ordinary classroom activity—the teacher models it, students try it in pairs, misunderstandings are resolved—before any score is at stake. In Phase 2, the *AI-mediated formal assessment*, each student holds a one-to-one voice-call role-play with the configured agent; because the AI absorbs the interlocutor and first-pass rater load, assessment proceeds in parallel on students' own devices within the blended course's online component. In Phase 3, *dual-track feedback and follow-up*, the teacher reviews each AI report against an analytic checklist, returns both tracks, elicits self- and peer-comments, and assigns targeted practice leading to a second AI assessment. The cycle deliberately mirrors the motivating–enabling–assessing sequence of the production-oriented approach (POA [25]) stipulated by the course syllabus: in-class trialing motivates the productive task, the AI-mediated session elicits the target production, and dual-track feedback operationalizes teacher-mediated assessing.

4.2 Task Design: Scenario Role-Play with Doubao

Each task is anchored in the current unit's theme and addresses a recognizable real-world domain—ordering food and drink, campus-service exchanges, a job or club interview—so the test task corresponds to features of a target language use (TLU) situation [26]. Doubao plays the scenario interlocutor—a role in which voice-based generative agents have proved workable in EFL classes [27]—and each student receives a digital role card listing communicative goals and unit language items, roughly ten per cent of which are optional "stretch" items that keep the assessment an occasion for learning. The tasks follow the paired textbooks unit by unit. In Unit 1 ("Fresh start" [23]), scenarios rehearse campus-service exchanges for new students; in Unit 4 ("Social media matters"), whose companion Listening and Speaking section "Relax and explore" trains dining and food-ordering language [24], the agent plays the barista of a campus café the student orders for a visiting friend, asks prices, handles a sold-out item, and recommends an alternative with reasons, deploying unit forms such as *Would you like…?*. Task demands are calibrated to the syllabus's speaking objectives—describing wishes and plans, narrating chronologically, retelling, and stating views on everyday topics with prepared evidence.

The agent is teacher-authored, not off-the-shelf. Its system prompt specifies the role and persona; the scenario, agenda, and scripted complications; interaction management rules (short turns, never answering for the student, recasts rather than lectures); and the post-interaction scoring instruction—0–9 sub-scores on the four IELTS criteria with comments and corrections (the full prompt is available from the corresponding author upon request). Design decisions are thus encoded in text before any student meets the agent and held constant across the class.

4.3 Scoring Model: IELTS-Based Criteria Adapted for the Classroom

The scoring model adapts the four official IELTS speaking criteria—Fluency and Coherence (FC), Lexical Resource (LR), Grammatical Range and Accuracy (GRA), and Pronunciation (P)—into classroom descriptors derived from the public Band 5–7 wording [28]. Band 6, aligned with CSE Level 6 [22], is the typical first-semester target, Band 5 the acceptable minimum, Band 7 the stretch zone; each descriptor is reworded in observable, unit-specific terms (for LR, "uses the unit's food and price vocabulary accurately").

Table 2 The Four IELTS Criteria: Classroom-Adapted Focus and the Doubao/Teacher Division of Labour

Criterion	Classroom-adapted focus (from Band 5–7)	What Doubao scores well	What requires teacher compensation
Fluency and Coherence	Maintains flow with manageable hesitation; uses basic connectives; extends answers beyond single words	Detectable from transcript: speech length, hesitation/repetition markers, use of connectives and discourse markers	Judging whether hesitation reflects language gaps or task/cognitive load; tolerance thresholds for L1-typical pausing
Lexical Resource	Unit vocabulary used accurately; paraphrase attempts; ≥1 stretch item deployed	Vocabulary range and accuracy from transcript; flags inappropriate word choice; suggests upgrades	Whether the unit's target items were genuinely deployed as the syllabus requires; credit for strategic paraphrase under pressure
Grammatical Range and Accuracy	Mix of simple and complex forms; errors rarely block meaning; unit structures attempted	Error detection and correction from transcript; range of structures; generates polished reformulations	Distinguishing developmental from careless errors; deciding which unit structures were required versus optional
Pronunciation	Generally intelligible; individual word/sound mispronunciation noted; stress and intonation support meaning	Only coarse, transcript-plus-prosody inference; can annotate stress and phonics rules but cannot "hear" phoneme-level errors (no acoustic engine)	Direct auditory judgement of segmental accuracy, intelligibility, and intonation via observation/recording—the teacher's primary scoring responsibility

Precision is essential here: Doubao has no official, standardized speaking assessment engine. Its scores are *prompt-driven LLM-simulated ratings*: the model, prompted to act as an IELTS examiner, renders the speech into an internal ASR transcript and generates a 0–9 sub-score and comment per criterion, with error corrections and suggestions. Equally important, Doubao has no phoneme-level acoustic pronunciation evaluation: its Pronunciation rating rests on the transcript plus coarse prosodic inference, making P the weakest scoring channel—a constraint the teacher checklist compensates for. Table 2 summarizes the classroom-adapted focus and the Doubao/teacher division

of labour per criterion.

Read column-wise, Table 2 shows where the design places its trust. Doubao's strengths cluster where the construct is recoverable from text—fluency markers, vocabulary, grammar—while Pronunciation is almost entirely human-rated: the AI receives the dimensions it can plausibly evidence, and the rest remain with the teacher. The ordering follows the construct-driven principle that technology implements, not defines, the construct [12]; this division of labour, not any claim about Doubao's accuracy, makes the scoring model defensible.

4.4 Dual-Track Feedback: Teacher Checklist and AI Report

The first feedback track is a one-page teacher analytic checklist in the tradition of direct speaking assessment [9], adding items an AI interlocutor cannot judge—interaction strategies, pragmatic appropriateness, and pronunciation intelligibility. The second track is Doubao's immediate post-interaction report, whose chief value is immediacy at scale: all forty to sixty students receive individualized written feedback minutes after speaking, and such automatic feedback can support speaking development and self-efficacy [18]. Because AI-generated feedback tends toward breadth over depth [19,29] and learners perceive limits in its personalization [30], the report is a *first* layer that the teacher reviews and contextualizes, not the final judgment. Students then write a short self-comment and exchange peer-comments, building evaluative judgement [8].

4.5 Feedback Loop and Follow-Up Learning

Each unit's assessment feeds the next: students complete a short targeted-practice assignment—re-recording a weak turn or drilling corrected forms—and later re-run the scenario for a second scored attempt. Comparing the two reports gives students evidence of progress and the teacher diagnostic information for reteaching, turning isolated scoring into learning-oriented assessment [7].

5. Pilot Empirical Evidence

To obtain preliminary in situ evidence of the design just described, a pilot quasi-experimental questionnaire survey followed one semester of implementation, using the author's teaching data from the 2025–2026 academic year.

5.1 Method

Participants and intervention. Participants were non-English-major freshmen (Chinese EFL learners with Mandarin as L1) from several intact College English classes (each of 40–60 students) at the same application-oriented university in Hubei Province, all following the same paired textbooks [23,24]. The intact classes were assigned, without randomization, to two conditions. The experimental group (n = 127; 113 male, 88.98%; 14 female, 11.02%) used Doubao as a required weekly speaking-practice tool for one semester: each week, every student completed at least one voice-call speaking practice session with the configured Doubao agent—the scenario dialogues, simulated IELTS-criterion scoring, and immediate AI feedback described in Section 4—and the teacher regularly reviewed the AI-generated learning reports and addressed common difficulties in class. The control group (n = 128; 86 male, 67.19%; 42 female, 32.81%) followed conventional speaking instruction over the same period—teacher-led classroom speaking activities such as pair work and group discussion with teacher feedback, without any AI tool—and had no exposure to Doubao. All participants completed the questionnaire at semester's end.

Instrument. The self-report questionnaire used five-point Likert scales (1 = completely

untrue/lowest; 5 = very true/highest) and covered four blocks: usage behavior (X1–X4: frequency, duration, functional and scenario diversity), process dimensions (M1–M4: material fit, dialogue extension, feedback immediacy and revision uptake, output complexity), technology adaptability (W1), and seven dimensions of perceived speaking-ability gains (Y1–Y7: fluency, grammar accuracy, communicative effectiveness, pronunciation clarity, vocabulary richness, speaking confidence, situational coping). Each process construct and the technology-adaptability block were each measured by three indicator items; as the survey platform export provides item-level frequency distributions, all group comparisons are conducted at the item level. Experimental-class respondents (n = 125) also answered qualitative multiple-choice items on mechanisms of use. One point must be stated plainly: all outcomes are *perceived* gains—self-assessed improvement, not objective speaking test scores.

Analysis. Group differences were tested with independent-samples Welch t-tests (dfs ≈ 243–253 across items), with Cohen's d as the effect size. Because the design is quasi-experimental and the measures self-reported, strong causal attribution is not warranted. Likert responses are treated as interval-scaled, following common practice in second-language assessment research; given the number of item-level comparisons, unadjusted alpha is acknowledged below where relevant.

5.2 Results

Table 3 Group Comparison of Perceived Speaking Gains (Y1–Y7): Welch t-Tests and Effect Sizes

Item	Exp M (SD)	Ctrl M (SD)	t	p	Cohen's d
Y1 Fluency	3.32 (1.00)	2.66 (1.11)	5.04	<.001	0.63
Y2 Grammar accuracy	3.24 (0.96)	2.66 (1.12)	4.51	<.001	0.56
Y3 Communicative effectiveness	3.21 (1.02)	2.48 (1.16)	5.39	<.001	0.67
Y4 Pronunciation clarity	3.07 (0.99)	2.56 (1.05)	3.98	<.001	0.50
Y5 Vocabulary richness	3.07 (0.99)	2.47 (1.06)	4.71	<.001	0.59
Y6 Speaking confidence	3.07 (0.97)	2.49 (1.07)	4.52	<.001	0.57
Y7 Situational coping	2.93 (1.01)	2.39 (1.13)	4.01	<.001	0.50

Note. Exp = experimental class (n = 127); Ctrl = control class (n = 128). Welch dfs range from 247.7 to 252.5. Rates of rating 3 or above—the scale midpoint anchors are mild affirmatives ('basically conforms', 'some improvement') rather than neutral—ranged from 69.3% to 85.8% in the experimental class versus 43.0% to 53.9% in the control class; strict top-two-box rates (rating ≥ 4) were 21.3%–36.2% versus 13.3%–21.1%.

The primary analyses compared the two groups on the seven dimensions of perceived speaking gains. Descriptively, experimental-class means cluster around or above the scale midpoint across all seven dimensions (2.93–3.32), whereas control-class means fall consistently below it (2.39–2.66); the descriptive gap is widest for communicative effectiveness and fluency and narrowest for situational coping and pronunciation clarity. The inferential results confirm this pattern. Table 3 presents the group comparison on the seven perceived speaking-gain dimensions, the main outcome variables.

All seven comparisons reached significance at $p < .001$, with medium effect sizes ($d = 0.50$–0.67). Communicative effectiveness showed the largest difference ($d = 0.67$), followed by fluency

(d = 0.63) and vocabulary richness (d = 0.59); the smallest differences appeared for situational coping (d = 0.50) and pronunciation clarity (d = 0.50). Read against the design's purpose, these uniform positive responses indicate that students accepted the assessment process across every dimension the questionnaire covered; they are perceptions of students who used the system, not measured proficiency gains (Section 7).

Usage and process variables index whether students actually used the system and whether its process operated as designed. All four usage items favored the experimental class (frequency, d = 0.58; duration, d = 0.49; functional diversity, d = 0.30; scenario diversity, d = 0.59), and every indicator of the four process constructs differed in the expected direction (material fit, d = 0.37–0.53; dialogue extension, d = 0.44–0.58; feedback immediacy and revision uptake, d = 0.44–0.57; output complexity, d = 0.30–0.43). Technology adaptability likewise favored the experimental class (d = 0.44–0.58), indicating that students could operate the tool. Because these item-level tests are unadjusted for multiple comparisons, the weakest effects (functional diversity; the lower output-complexity indicators) should be read as exploratory; the seven primary outcome effects in Table 3 remain significant under any conventional correction.

Qualitative multiple-choice items (experimental class, n = 125) corroborated that the process operated as designed and was positively received. First, 83.2% reported that Doubao's follow-up questions increased their number of dialogue turns, and 78.4% reported frequently revising their expressions after its corrections. Second, dual-track fidelity was visible: 79.2% reported that the teacher regularly checked their learning reports and explained difficult points. Most strikingly, every respondent—100%—reported an accumulating sense of achievement and no fear of mistakes.

5.3 Interpretation

The pilot answers the question this paper set out to test: whether the assessment system can operate in a real classroom. It ran in routine teaching for a full semester—weekly AI-mediated sessions were completed, the feedback loop was used, and dual-track teacher moderation was enacted—and students accepted the process, reporting positive perceptions on all seven dimensions and, unanimously, an accumulating sense of achievement. Whether these perceptions correspond to measured proficiency gains is a separate research question requiring objective testing and lies beyond this paper's scope; the evidence presented here supports operability and acceptance only (Section 7).

6. Test Principles

With the design specified and the pilot evidence in hand, the design is evaluated below against five classical test qualities, bounded by what Doubao is: a prompt-driven simulated rater, not a validated engine.

Validity. As an achievement measure, the assessment's content domain is the textbook unit: scenarios are anchored in unit themes, role cards sample unit target language, and students perform target speech acts directly [9]. Two construct threats qualify this claim. First, underrepresentation: Doubao's ratings derive from its ASR transcript, so segmental pronunciation, prosody, and pragmatic stance are invisible to it [15]; the design answers by assigning such facets to the teacher track. Second, construct-irrelevant variance: ASR errors, which fall disproportionately on underrepresented speaker groups [16], propagate into every transcript-derived score. Validity claims are therefore bounded: Doubao's simulated scores support only formative, diagnostic interpretations, with teacher judgment decisive for any reported grade.

Reliability. An automated marker applies the same criteria identically across students and occasions [2], and generative-AI scoring has reached acceptable to high agreement with human

judgment [13,14]. An AI first pass is thus likely more consistent than an unaided teacher rating sixty students, though published figures transfer only cautiously to prompt-driven simulation on an untested platform. Reliability is therefore treated as a moderation problem: the teacher reviews each AI report, re-listens where the tracks diverge, and remains decisive.

Practicality. In classes of 40–60, teacher-only individual oral assessment is infeasible. The AI absorbs the two loads that do not scale—the interlocutor role and first-pass ratings with written feedback—so assessment runs in parallel on students' own devices, with no specialist equipment or licensing. Pilot usage data suggest engagement followed: the experimental class reported significantly higher frequency, duration, and scenario diversity of practice ($d = 0.49–0.59$).

Authenticity. Authenticity is "the degree of correspondence of the characteristics of a given language test task to the features of a TLU task" [26]. The scenarios preserve the information gaps, complications, and extended-turn demands of recognizable service and interview encounters. A student–AI exchange is admittedly less authentic in stakes and unpredictability; the honest response is to specify the new task characteristics rather than deny them [12].

Washback. Here the pilot supplies evidence where only principled reasoning previously stood. Individually, 78.4% of experimental-class respondents reported regularly revising their output after AI corrections, and positive perceptions exceeded controls on all seven dimensions (all $p < .001$, $d = 0.50–0.67$); every respondent reported an accumulating sense of achievement—together indicating that the feedback loop produced engagement with the assessment process. Negative washback remains possible: students may learn to "perform for the algorithm," and over-reliance on AI feedback may displace deeper teacher guidance [31,32]. Low stakes, teacher moderation, and scripted complications mitigate these risks.

7. Limitations and Future Directions

Four limitations qualify the design and the pilot evidence. First, all outcome data are self-reported *perceived gains*, not objective speaking test scores, and self-perception can diverge from measured proficiency. The survey was administered once, with no pre-test, so baseline equivalence cannot be established. The control class answered the same questionnaire containing Doubao-specific wording, interpreted against their conventional speaking-learning experience—a compromise of uncertain measurement equivalence. The gender composition was imbalanced (88.98% male in the experimental class versus 67.19% in the control class), and the sample comes from a single semester at a single institution; the group differences should therefore be read as encouraging but strictly preliminary.

Second, technical limits persist. Doubao has no phoneme-level acoustic pronunciation evaluation, so segmental errors that leave the ASR transcript intact can pass as correct, and every transcript-based judgement inherits ASR error [16,33]. Third, prompt-driven scoring adds consistency and explainability risks: outputs may vary across sessions, prompt phrasings, and silent model updates, and validation theory has yet to catch up [15]. Fourth, pedagogical and ethical concerns remain: uploading students' voice recordings to a commercial platform raises consent and storage questions [34], and most learners still prefer human interlocutors [14].

Future work should pursue two strands: an objective pre-test/post-test study of whether perceived gains match measured gains, and a human–machine agreement study benchmarking Doubao's ratings against teacher ratings criterion by criterion [15,35].

8. Conclusion

This study designed an AI-assisted, student-centered speaking assessment for Chinese College English. Its core purpose was practical: in large classes of forty to sixty students staffed by a single

teacher, routine speaking assessment is arithmetically infeasible, and the design answers that infeasibility by assigning the interlocutor and first-pass rater roles to Doubao, a freely available conversational AI, while keeping teacher judgment decisive.

Three contributions follow. First, the paper provides a complete, replicable design architecture: test specifications derived from the instructional context, unit-anchored scenario role-play tasks, classroom-adapted IELTS Band 5–7 rubrics, a teacher-authored agent prompt, and a dual-track feedback loop pairing immediate AI reports with a teacher analytic checklist. Second, it states Doubao's technical boundaries explicitly—prompt-driven simulated scoring, with neither a standardized assessment engine nor phoneme-level acoustic pronunciation evaluation—together with the corresponding compensating measures, notably teacher–AI dual rating and bounded, formative-only validity claims. Third, a one-semester pilot (experimental n = 127, control n = 128) found that students responded positively to the process across all seven perceived dimensions (all $p < .001$, $d = 0.50–0.67$) and in behavioral engagement—usage, dialogue extension, and revision uptake—indicating that the assessment operated at scale and was accepted.

For practitioners, the design transfers directly: begin with one unit scenario as ungraded formative practice, add the teacher checklist once the workflow is stable, and introduce dual rating before any score is reported. For administrators and policy makers in application-oriented universities, the design shows that routine speaking assessment can be embedded into existing curricula at near-zero marginal cost, without new equipment or licensing. For researchers, the next steps are human–machine rating-agreement studies and objective pre-/post-test validation of learning gains, the subject of a follow-up study.

Acknowledgement

This research was funded by the Guiding Project of the 2024 Philosophical and Social Sciences Research Project of the Department of Education of Hubei Province, China, with the project entitled "Research and Practice on Blended Teaching of College English Empowered by Digital Intelligence in Applied Universities of Hubei Province" (Project No. 24G161).

References

[1] Turner, C. E., & Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Banerjee (Eds.), Handbook of second language assessment (pp. 255–272). De Gruyter Mouton.

[2] Xu, J., Jones, E., Laxton, V., & Galaczi, E. (2021). Assessing L2 English speaking using automated scoring technology: Examining automarker reliability. Assessment in Education: Principles, Policy & Practice, 28(4), 411–436. https://doi.org/10.1080/0969594X.2021.1979467

[3] Du, J., & Daniel, B. K. (2024). Transforming language education: A systematic review of AI-powered chatbots for English as a foreign language speaking practice. Computers and Education: Artificial Intelligence, 6, Article 100230. https://doi.org/10.1016/j.caeai.2024.100230

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

[5] 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

[6] Zhang, C., Meng, Y., & Ma, X. (2024). Artificial intelligence in EFL speaking: Impact on enjoyment, anxiety, and willingness to communicate. System, 121, Article 103259. https://doi.org/10.1016/j.system.2024.103259

[7] Carless, D. (2007). Learning-oriented assessment: Conceptual bases and practical implications. Innovations in Education and Teaching International, 44(1), 57–66. https://doi.org/10.1080/14703290601081332

[8] Yan, Q. (2024). Exploring Chinese university EFL students' perceptions of formative assessment: A qualitative study. System, 125, Article 103391. https://doi.org/10.1016/j.system.2024.103391

[9] Luoma, S. (2004). Assessing speaking. Cambridge University Press.

[10] Wu, X., Zhang, L. J., & Liu, Q. (2021). Using assessment for learning: Multi-case studies of three Chinese

university English as a foreign language (EFL) teachers engaging students in learning and assessment. Frontiers in Psychology, 12, Article 725132. https://doi.org/10.3389/fpsyg.2021.725132
[11] Cetin Koroglu, Z. (2021). Using digital formative assessment to evaluate EFL learners' English speaking skills. GIST Education and Learning Research Journal, 22, 103–123. https://files.eric.ed.gov/fulltext/EJ1308042.pdf
[12] Chapelle, C. A., & Douglas, D. (2006). Assessing language through computer technology. Cambridge University Press.
[13] Koraishi, O. (2024). The intersection of AI and language assessment: A study on the reliability of ChatGPT in grading IELTS writing task 2. Language Teaching Research Quarterly, 43, 22–42. https://doi.org/10.32038/ltrq.2024.43.02
[14] Susoy, Z. (2026). Reducing anxiety and enhancing performance: The impact of AI chatbots versus human facilitation on EFL speaking assessment outcomes. Frontiers in Psychology, 16, Article 1745942. https://doi.org/10.3389/fpsyg.2025.1745942
[15] Xi, X. (2023). Advancing language assessment with AI and ML—Leaning into AI is inevitable, but can theory keep up? Language Assessment Quarterly, 20(4–5), 357–376. https://doi.org/10.1080/15434303.2023.2291488
[16] Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117
[17] 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, Article 103533. https://doi.org/10.1016/j.system.2024.103533
[18] Zou, B., Du, Y., Wang, Z., Chen, J., & Zhang, W. (2023). An investigation into artificial intelligence speech evaluation programs with automatic feedback for developing EFL learners' speaking skills. SAGE Open, 13(3). https://doi.org/10.1177/21582440231193818
[19] Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. Education and Information Technologies, 29(7), 8435–8463. https://doi.org/10.1007/s10639-023-12146-0
[20] McNamara, T. (2000). Language testing. Oxford University Press.
[21] Modern Languages Division. (2001). Common European framework of reference for languages: Learning, teaching, assessment. Cambridge University Press.
[22] National Education Examinations Authority, & British Council. (2019, January 15). Success in linking IELTS and Aptis to China's Standards of English Language Ability. https://cse.neea.edu.cn/html1/report/19011/9306-1.htm
[23] Zheng, S. (Ed.). (2023). New horizon college English: Reading and writing 1 (4th ed.). Foreign Language Teaching and Research Press.
[24] Zheng, S. (Ed.). (2024). New horizon college English: Viewing, listening and speaking 1 (4th ed.). Foreign Language Teaching and Research Press.
[25] Wen, Q. (2018). The production-oriented approach to teaching university students English in China. Language Teaching, 51(4), 526–540. https://doi.org/10.1017/S026144481600001X
[26] Bachman, L. F., & Palmer, A. S. (1996). Language testing in practice: Designing and developing useful language tests. Oxford University Press.
[27] Wan, Y., & Moorhouse, B. L. (2024). Using Call Annie as a generative artificial intelligence speaking partner for language learners. RELC Journal, 56(2), 489–498. https://doi.org/10.1177/00336882231224813
[28] IELTS. (2007). IELTS speaking band descriptors (public version). https://ielts.org/cdn/ielts-guides/ielts-speaking-band-descriptors.pdf
[29] Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. International Journal of Educational Technology in Higher Education, 20, Article 57. https://doi.org/10.1186/s41239-023-00425-2
[30] Zou, B., Liviero, S., Ma, Q., Zhang, W., Du, Y., & Xing, P. (2024). Exploring EFL learners' perceived promise and limitations of using an artificial intelligence speech evaluation system for speaking practice. System, 126, Article 103497. https://doi.org/10.1016/j.system.2024.103497
[31] Choi, S. (2025). Ethical issues of using AI in language education: A critical perspective. English Teaching, 80(5), 229–248. https://files.eric.ed.gov/fulltext/EJ1494237.pdf
[32] Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148
[33] Ngo, T.-N., Chen, H.-J., & Lai, K.-W. (2024). The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta-analysis. ReCALL, 36(1), 4–21. https://doi.org/10.1017/S0958344023000113
[34] Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Buckingham Shum, S., Santos, O. C., Rodrigo, M. M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a

community-wide framework. International Journal of Artificial Intelligence in Education, 32(3), 504–526. https://doi.org/10.1007/s40593-021-00239-1
[35] Ma, M., Noordin, N., & Razali, A. B. (2025). Improving EFL speaking performance among undergraduate students with an AI-powered mobile app in after-class assignments: An empirical investigation. Humanities and Social Sciences Communications, 12, Article 468. https://doi.org/10.1057/s41599-025-04688-0