PPSDP International Journal of Education
Volume 5 (1) 2026, 452-463
E-ISSN 2829-5196, P-ISSN 2830-3229

Enhancing EFL Oral Accuracy Through Lightweight AI Feedback in Literature-Based Speaking Tasks: A Quasi-Experimental Study of Chinese University Students

Xue Cao[1], Kun Wang[1]
[1]School of Foreign Languages, Jilin Engineering Normal University, Changchun, China

Corresponding author e-mail: 172903221@qq.com

Article History: Received on 14 May 2026, Revised on 30 June 2026, Published on 25 July 2026

Abstract: This study investigates whether lightweight AI feedback improves Chinese university students' oral accuracy in literature-based EFL speaking tasks compared to conventional teacher feedback. A quasi-experimental pretest-posttest design was employed with 64 intermediate-level English majors over 12 weeks. The experimental group received immediate AI-generated feedback on grammatical errors, while the control group received delayed teacher feedback (48 hours). Oral accuracy was measured using error-free clauses (EFC), error-free T-units (EFT), and grammatical accuracy ratio (GAR). The experimental group demonstrated significantly greater improvements across all metrics than the control group, with large effect sizes ($d > 1.30$). Students receiving AI feedback also reported positive perceptions of its usefulness, ease of use, and intention to continue using such tools. This study uniquely integrates AI feedback with literature-based speaking tasks in Chinese higher education, moving beyond scripted dialogues to examine complex, meaning-oriented oral discourse. AI feedback offers a scalable, immediate, and individualized solution for large EFL classes where teacher feedback is constrained by workload, promoting learner autonomy without disrupting communicative activities. This research provides empirical support for the Noticing and Output Hypotheses, demonstrating that immediate AI feedback accelerates error awareness and grammatical restructuring in oral production, offering actionable insights for EFL curriculum designers and instructors.

Keywords: AI Feedback, EFL Oral Accuracy, Error-Free Clauses, Grammatical Accuracy Ratio, Literature-Based Speaking Tasks

A. Introduction

Oral communication competence has long been regarded as one of the primary objectives of foreign language education. Among the various dimensions of oral proficiency, accuracy plays a significant role because it reflects learners' ability to produce grammatically correct and contextually appropriate language (Ellis, 2008). In Chinese higher education, despite years of English learning, many students continue

to face challenges in producing accurate spoken English, particularly in spontaneous communicative situations (Wen & Johnson, 1997). Common errors include subject-verb agreement violations, tense inconsistencies, preposition misuse, and article omission, all of which can impede intelligibility and undermine communicative confidence.

To address this issue, educators have increasingly adopted communicative and task-based approaches to language teaching. Literature-based speaking tasks, which encourage students to discuss literary texts, analyze characters, and express personal interpretations, have gained popularity in EFL classrooms. Such tasks provide meaningful contexts for language use while promoting critical thinking and learner engagement (Lazar, 1993). Unlike decontextualized grammar drills, literature-based activities immerse learners in authentic language use, requiring them to negotiate meaning, support opinions with textual evidence, and respond to peers' interpretations. However, because these tasks place heavy cognitive demands on content generation and idea organization, learners often prioritize fluency and meaning over formal accuracy, resulting in persistent grammatical errors that may become fossilized without targeted intervention.

At the same time, advances in artificial intelligence have transformed language learning environments. AI-powered tools are capable of generating immediate feedback on learners' linguistic performance. Unlike traditional teacher feedback, which may be delayed due to classroom constraints (e.g., large class sizes, limited office hours, and heavy grading loads), AI feedback can be delivered instantly and repeatedly. This feature makes AI a potentially valuable resource for enhancing speaking development (Holmes et al., 2019). Lightweight AI feedback refers to automated, form-focused feedback that identifies surface-level linguistic errors (e.g., grammar, word choice, syntax) without requiring sophisticated pedagogical reasoning. Such tools are increasingly accessible through speech recognition applications, chatbots, and AI-powered language learning platforms.

Although previous studies have demonstrated the benefits of AI-assisted feedback in writing instruction (Ranalli & Yamashita, 2022), research on AI feedback in oral language learning remains limited. Moreover, little attention has been paid to literature-based speaking activities in Chinese university contexts. Most existing AI speaking studies have focused on scripted dialogues, pronunciation drills, or simulated conversations with virtual agents, rather than on authentic, content-rich discussions of literary texts. This gap is significant because literature-based tasks represent a common pedagogical approach in advanced EFL courses, yet instructors lack empirical guidance on how to integrate AI feedback effectively without disrupting the communicative and interpretive nature of such tasks. Therefore, the present study seeks to explore the following research question: Does lightweight AI feedback improve Chinese university students' oral accuracy in literature-based EFL speaking tasks compared to conventional teacher feedback?

Oral Accuracy in EFL Speaking

Accuracy is generally defined as the extent to which learners produce language that conforms to target-language norms regarding grammar, vocabulary, and pronunciation (Ellis & Barkhuizen, 2005). Together with fluency and complexity, accuracy constitutes a major indicator of second language speaking proficiency (Skehan, 1998). These three dimensions are often in tension: as learners strive to communicate fluently, they may sacrifice accuracy; conversely, excessive attention to form can disrupt fluency. In classroom settings, accuracy is particularly important because it affects both intelligibility and the learner's perceived credibility. Research has shown that EFL learners often prioritize meaning over form during communication, resulting in grammatical errors and inappropriate language use (Housen et al., 2012). This tendency is exacerbated in meaning-focused tasks such as literature discussions, where learners are evaluated primarily on content and participation. Without external feedback, learners may fail to notice the gap between their interlanguage and target norms, leading to error fossilization. Therefore, providing effective feedback remains a crucial component of speaking instruction. Corrective feedback allows learners to identify discrepancies between their output and target language forms. Previous studies have found that feedback contributes positively to learners' grammatical development and speaking performance (Lyster & Saito, 2010). However, the timing, type, and source of feedback influence its effectiveness. Delayed feedback, while allowing for uninterrupted communication, may lose its salience; immediate feedback, by contrast, captures the learner's attention while the error is still accessible in working memory.

Literature-Based Speaking Tasks

Literature-based language teaching emphasizes the use of literary texts as resources for language learning. Literary works expose learners to authentic language while encouraging interpretation and discussion (Carter & Long, 1991). Unlike textbook dialogues, literary texts feature rich vocabulary, complex sentence structures, and culturally embedded meanings that stimulate higher-order thinking. According to Lazar (1993), literature offers meaningful contexts that stimulate communication and learner involvement. Through discussions of themes, characters, and personal responses, students engage in extended oral interaction that supports language development. For example, a discussion of moral dilemmas in a short story requires learners to use conditional sentences ("If I were the protagonist…"), evaluative language ("This suggests that…"), and hypothetical reasoning, all of which are linguistically demanding. Recent studies have suggested that literature-based speaking tasks can enhance motivation, communicative confidence, and critical thinking skills (Paran, 2008). However, because learners focus heavily on content discussion, attention to linguistic accuracy may be reduced. In fact, the cognitive load of literary analysis remembering plot details, interpreting symbolism, formulating original arguments can overwhelm learners' limited attentional resources, leaving

little capacity for self-monitoring of grammatical forms. Consequently, additional support mechanisms such as AI feedback may be necessary to redirect learners' attention to form without interrupting the flow of meaning-making.

AI Feedback in Language Learning

Artificial intelligence has emerged as an increasingly important technology in language education. AI systems can analyze learner performance and generate personalized feedback within seconds (Kohnke et al., 2023). In the context of speaking, AI feedback typically relies on automatic speech recognition (ASR) to transcribe oral output, followed by natural language processing (NLP) to detect grammatical errors, lexical inaccuracies, and pronunciation deviations. Compared with traditional teacher feedback, lightweight AI feedback possesses several advantages: (a) immediacy, as feedback is delivered within seconds of task completion; (b) accessibility, as learners can access feedback anytime without scheduling appointments; (c) consistency, as AI applies the same error detection criteria to all learners; and (d) individualized support, as each learner receives feedback tailored to their specific errors. However, limitations include occasional ASR inaccuracies (particularly with non-native accents), inability to address pragmatic or discourse-level issues, and lack of affective support. Research indicates that AI-assisted feedback can improve language performance by increasing opportunities for self-correction and reflection (Li & Hafner, 2022). In writing contexts, studies have shown that automated feedback leads to significant reductions in grammatical errors over time. In speaking contexts, emerging evidence suggests that AI feedback on pronunciation and grammar can promote accuracy gains, though more research is needed on complex, extended discourse. The present study is informed primarily by Schmidt's Noticing Hypothesis and Swain's Output Hypothesis.

Noticing Hypothesis

Schmidt argues that conscious attention to language forms is necessary for language acquisition (Schmidt, 1990). Learners cannot improve grammatical performance unless they first notice their linguistic errors. In typical classroom interactions, learners may produce non-target forms repeatedly without ever recognizing them as errors, especially if their utterances are understood by interlocutors. Noticing requires either internal self-detection (rare among lower-proficiency learners) or external feedback that draws attention to mismatches between output and target norms. Lightweight AI feedback highlights inaccuracies immediately after speaking activities, thereby increasing learners' awareness of problematic forms. By presenting explicit corrections (e.g., "You said 'She go to school.' The correct form is 'She goes to school.'"), AI feedback makes the error salient and provides a clear target for restructuring.

Output Hypothesis

Swain proposes that language production promotes learning because learners become aware of gaps in their linguistic knowledge while attempting to communicate (Swain, 1985). This "pushing" process having to produce language that is precise, coherent, and appropriate triggers syntactic processing that comprehension alone does not. Moreover, when learners receive feedback on their output, they are prompted to modify and reprocess their interlanguage representations. In literature-based discussions, students are encouraged to express opinions and interpretations. AI feedback enables learners to identify weaknesses in their spoken output and modify future performance. The combination of meaningful output (literature discussion) and focused feedback (AI error correction) creates a cycle of noticing, hypothesis testing, and restructuring that drives grammatical development. Together, these theories suggest that AI feedback may facilitate oral accuracy development through enhanced noticing and self-monitoring. Importantly, both theories emphasize the role of conscious attention; AI feedback is uniquely positioned to trigger such attention immediately after production, when the utterance is still fresh in working memory.

B. Methods

The study employed a quasi-experimental pretest-posttest design involving an experimental group and a control group. The independent variable was feedback type (lightweight AI feedback vs. traditional teacher feedback), and the dependent variable was oral accuracy as measured by three metrics. The intervention lasted twelve weeks, with one 90-minute literature-based speaking session per week.

Table 1. Experiment and Control Group

Group	Treatment
Experimental Group	Literature-based speaking tasks + lightweight AI feedback
Control Group	Literature-based speaking tasks + conventional teacher feedback

The participants consisted of 64 undergraduate students enrolled in a compulsory College English course at a medium-sized university in eastern China. All participants were second-year English majors (the course was an advanced speaking component of the English program). Participants were recruited from two parallel classes to avoid cross-group contamination. One class was randomly assigned to the experimental group, and the other to the control group.

Table 2. Participants

Variable	Information
Total participants	64
Experimental group	32 (26 female, 6 male)
Control group	32 (24 female, 8 male)
Age range	19–21 years (M = 19.8, SD = 0.7)
English proficiency level	Intermediate (CEFR B2), as measured by their university entrance English exam scores (M = 78.6/100, SD = 5.2)
Prior literature-based speaking experience	None had formal experience; all had experience with general topic-based speaking tasks

Prior to the intervention, an independent-samples t-test on pretest oral accuracy scores confirmed that there was no significant difference between the two groups ($t(62) = 0.36$, $p = .72$, Cohen's $d = 0.09$), indicating comparability at baseline.

Students in both groups completed weekly literature-based speaking tasks based on selected short stories from contemporary and classic English literature. Texts included *The Gift of the Magi* (O. Henry), *The Story of an Hour* (Kate Chopin), *A Clean, Well-Lighted Place* (Ernest Hemingway), and *Lamb to the Slaughter* (Roald Dahl). Texts were selected for their accessibility (lexical difficulty suitable for B2 learners), length (2,000–3,000 words), and thematic richness (topics such as sacrifice, freedom, loneliness, and justice that invite discussion). Each weekly session followed a consistent structure: Pre-task (15 minutes): Students read the assigned text at home before class. In class, the instructor briefly clarified vocabulary and plot points. Task (50 minutes): Students worked in small groups (4–5 members) to complete a speaking task. Tasks rotated weekly and included: (a) character analysis ("Compare the motivations of two characters"), (b) theme discussion ("What does the story suggest about the nature of love?"), (c) literary interpretation ("Explain the symbolism of the window in the story"), and (d) personal response presentation ("Describe a personal experience related to the story's theme"). Feedback phase (15 minutes): After each task, students received feedback based on their group condition. Revision/reflection (10 minutes): Students noted down three errors they had made and corrected them.

The feedback protocols for the two groups were as follows. Experimental Group (AI feedback): After completing each speaking task, students individually uploaded their audio recordings (2–3 minutes of sustained speech per student) to an AI feedback platform (a custom interface using Google's Speech-to-Text API combined with Language Tool for grammar checking). Within 30 seconds, the platform returned a feedback report listing: (a) transcribed text, (b) highlighted grammatical errors with corrections, (c) questionable word choices with alternatives, and (d) sentence structure notes. Students were instructed to review their reports, note three errors, and correct them in writing. No teacher feedback was provided. Control Group (Teacher feedback): After completing each speaking task, students submitted their audio recordings to the course instructor (a PhD candidate with 6 years of EFL teaching experience). The instructor listened to each recording and within 48 hours provided written feedback focusing on grammatical errors, word choice, and sentence structure. Students received individual feedback documents. No AI feedback was provided. To ensure comparability, both groups received feedback on the same linguistic categories. The key differences were immediacy (immediate for AI vs. delayed up to 48 hours for teacher) and delivery mode (automated vs. human).

Pre-test and post-test speaking assessments were administered one week before and one week after the 12-week intervention. Both tests followed an identical format: students individually responded to a literature-based speaking prompt based on an unfamiliar short story (different stories for pre-test and post-test, matched for length

and difficulty). Students had 3 minutes to prepare and 3 minutes to speak. Responses were audio-recorded and transcribed for analysis. Pre-test story: *The Use of Force* (William Carlos Williams), Post-test story: *A Haunted House* (Virginia Woolf). Following previous studies, oral accuracy was measured through three established metrics:

1. Error-Free Clauses (EFC): Percentage of clauses (containing a subject and finite verb) that contained no grammatical errors. Errors counted included subject-verb agreement, tense, aspect, voice, word order, and omission/insertion of obligatory elements.
2. Error-Free T-units (EFT): Percentage of T-units (main clause plus any subordinate clauses) that were error-free. This metric is more stringent than EFC because a single error anywhere in a T-unit disqualifies the entire unit.
3. Grammatical Accuracy Ratio (GAR): Total number of grammatical errors divided by total number of words, multiplied by 100 (errors per 100 words). Lower scores indicate higher accuracy.

Two trained raters (both MA TESOL students) independently coded 20% of the transcripts. Inter-rater reliability was high (ICC = .91 for EFC, .88 for EFT, .93 for GAR). Disagreements were resolved through discussion. A 15-item Likert-scale questionnaire (1 = strongly disagree, 5 = strongly agree) was administered to the experimental group to investigate students' perceptions of AI feedback. The questionnaire comprised three subscales: perceived usefulness (5 items, e.g., "AI feedback helped me notice my grammatical errors"), perceived ease of use (5 items, e.g., "The AI feedback platform was easy to use"), and behavioral intention to continue using AI feedback (5 items, e.g., "I would like to use AI feedback in future English courses"). Cronbach's alpha for the full questionnaire was .89, indicating high internal consistency. Data were analyzed using SPSS version 26.0. Descriptive statistics (means and standard deviations) were calculated for all accuracy measures. Paired-samples t-tests were conducted to examine within-group improvement from pre-test to post-test. Independent-samples t-tests were conducted to compare post-test scores between the experimental and control groups. The significance level was set at $p < .05$. Effect sizes (Cohen's d) were calculated to interpret practical significance (small = 0.20, medium = 0.50, large = 0.80).

C. Results and Discussion

Descriptive Statistics

Table 3 presents the descriptive statistics for oral accuracy scores across the three metrics.

Table 3. Descriptive Statistics for Oral Accuracy Scores (N = 64)

Group	Metric	Pre-test M (SD)	Post-test M (SD)	Mean Change
Experimental (n=32)	EFC (%)	62.34 (8.21)	78.56 (7.43)	+16.22
	EFT (%)	54.17 (9.05)	71.23 (8.89)	+17.06
	GAR (errors/100 words)	8.92 (1.87)	5.14 (1.23)	-3.78
Control (n=32)	EFC (%)	61.89 (7.95)	68.34 (8.12)	+6.45
	EFT (%)	53.42 (8.76)	59.87 (9.34)	+6.45
	GAR(errors/100 words)	9.04 (1.92)	7.23 (1.68)	-1.81

Note. *EFC = Error-Free Clauses; EFT = Error-Free T-units; GAR = Grammatical Accuracy Ratio (lower = better)*

The results indicate that both groups improved from pre-test to post-test. However, the experimental group demonstrated substantially larger gains across all three metrics, with improvements approximately 2.5 times greater than those of the control group.

Within-Group Comparisons

Paired-samples t-tests were conducted to determine whether within-group improvements were statistically significant.

Table 4. Paired-Samples t-test Results for Within-Group Improvement

Group	Metric	t (df=31)	p	Cohen's d
Experimental	EFC	8.42	<.001	2.11
	EFT	7.89	<.001	1.96
	GAR	-9.13	<.001	2.35
Control	EFC	3.45	.002	0.85
	EFT	2.98	.006	0.73
	GAR	-4.01	<.001	0.99

The experimental group showed large, statistically significant improvements on all three metrics (all $p < .001$, Cohen's $d > 1.90$). The control group also showed significant improvements ($p < .01$), but effect sizes were moderate ($d = 0.73–0.99$). This indicates that while both feedback conditions promoted learning, AI feedback yielded substantially larger gains.

Between-Group Comparisons

Independent-samples t-tests compared post-test scores between the experimental and control groups.

Table 5. Independent-Samples t-test Results for Post-test Scores

Metric	t (df=62)	p	Cohen's d	95% CI for Difference
EFC	5.23	<.001	1.31	[6.32, 14.12]
EFT	5.89	<.001	1.48	[7.45, 15.29]
GAR	-5.67	<.001	1.42	[-2.84, -1.34]

All differences were statistically significant with large effect sizes (d > 1.30). Students in the experimental group achieved significantly higher oral accuracy scores than those in the control group. For example, the experimental group's mean EFC was 10.22 percentage points higher than that of the control group post-intervention, representing a substantial practical advantage.

Student Perceptions of AI Feedback (Experimental Group Only)

Questionnaire responses revealed generally positive perceptions. Mean scores (1–5 scale) were: perceived usefulness (M = 4.12, SD = 0.67), perceived ease of use (M = 3.95, SD = 0.74), and behavioral intention (M = 3.88, SD = 0.81). Representative student comments from open-ended responses included: "The AI feedback was very fast, so I could still remember what I wanted to say when I saw the errors," and "Sometimes the AI misheard me because of my accent, but most corrections were helpful."

Discussion

The findings provide a clear answer to the research question: lightweight AI feedback significantly improves Chinese university students' oral accuracy in literature-based EFL speaking tasks compared to conventional teacher feedback. Students who received immediate AI feedback demonstrated substantially greater gains across all three-accuracy metrics, with effect sizes in the large range (d > 1.30). These results align with and extend previous research on AI-assisted language learning (Kohnke et al., 2023; Li & Hafner, 2022) while offering novel insights specific to literature-based oral tasks.

Theoretical implications: first, the results support Schmidt's Noticing Hypothesis (Schmidt, 1990). Immediate AI feedback enabled learners to identify and notice grammatical errors that might otherwise remain unnoticed. In the teacher feedback condition, the 48-hour delay meant that by the time students received feedback, their memory of the original utterance had faded, reducing the salience of the error. This temporal proximity effect likely explains much of the difference between groups. The AI group's feedback arrived while the utterance was still active in working memory, allowing for direct comparison between erroneous output and target form. Second, the findings align with Swain's Output Hypothesis (Swain, 1985). Through repeated speaking practice and feedback cycles over 12 weeks, students in the experimental group became more aware of linguistic gaps and adjusted their language production accordingly. The cyclical process produce output, receive feedback, notice gap, modify internal representation was accelerated in the AI condition due to the immediacy of feedback. Students reported that seeing errors highlighted in their own transcribed speech made the gap between interlanguage and target norms highly concrete. Third, literature-based speaking tasks provided meaningful communicative contexts that encouraged sustained oral interaction. Unlike decontextualized grammar exercises, literature discussions required learners to express genuine

opinions, defend interpretations, and respond to peers—all cognitively demanding activities. The integration of AI feedback helped students maintain attention to linguistic form while discussing literary content, effectively addressing the "fluency vs. accuracy" trade-off. This finding suggests that form-focused feedback does not necessarily disrupt meaning-making when delivered after task completion, contrary to concerns that corrective feedback interrupts communicative flow.

Pedagogical implications: The study offers several practical recommendations for EFL instructors. First, lightweight AI feedback can be integrated into existing literature-based curricula with minimal disruption. The 15-minute feedback phase fit naturally at the end of each 90-minute session, allowing students to reflect on their linguistic performance without interrupting collaborative discussion. Second, AI feedback is particularly valuable in large classes where individualized teacher feedback is impractical. In this study, the instructor spent approximately 20 hours per week providing feedback to the control group, whereas the AI feedback required no instructor time after initial setup. Third, AI feedback appears to promote learner autonomy. Students in the experimental group reported reviewing their feedback reports independently and keeping error logs, behaviors that were less common in the control group.

Comparison with teacher feedback: The superiority of AI feedback in this study should be interpreted cautiously. Teacher feedback was delayed by 48 hours due to realistic workload constraints; under ideal conditions (e.g., one-on-one tutoring with immediate feedback), teacher feedback might be equally or more effective. However, in authentic classroom settings with 30+ students per class, immediate individualized feedback from teachers is rarely feasible. AI feedback thus represents a scalable solution that does not require additional human resources. Moreover, teacher feedback in this study was provided in written form (to ensure consistency), whereas AI feedback was delivered through an interactive interface that allowed students to click on errors for explanations. The multimodal presentation may have enhanced learning.

Unexpected findings: The control group also showed significant improvement, with moderate effect sizes ($d = 0.73–0.99$). This suggests that literature-based speaking tasks themselves promote oral accuracy development, even without immediate or technology-enhanced feedback. Simply engaging in repeated, meaningful oral production over 12 weeks led to noticeable gains. The experimental group's larger gains indicate that AI feedback builds upon and amplifies the natural learning effects of task repetition.

Several limitations should be acknowledged. First, the sample size was relatively limited ($n = 64$) and drawn from a single university, limiting generalizability to other contexts. Second, the intervention period was restricted to one academic semester (12 weeks); long-term retention of accuracy gains was not measured. Third, the AI

feedback focused exclusively on grammatical, lexical, and syntactic accuracy; pronunciation feedback was limited due to ASR constraints. Fourth, the study did not measure fluency or complexity; it is possible that increased attention to accuracy negatively affected fluency, though anecdotal evidence from transcript analysis did not suggest dramatic trade-offs. Fifth, the teacher feedback condition may have been disadvantaged by the 48-hour delay; future studies should compare AI feedback with immediate teacher feedback to isolate the effect of feedback source from timing.

D. Conclusion

This study investigated the effectiveness of lightweight AI feedback in improving oral accuracy during literature-based EFL speaking tasks among Chinese university students. Employing a quasi-experimental design with 64 intermediate-level participants over 12 weeks, the study found that students receiving immediate AI feedback demonstrated significantly greater improvements in error-free clauses, error-free T-units, and grammatical accuracy ratio compared to students receiving delayed conventional teacher feedback, with large effect sizes ($d > 1.30$). Questionnaire data revealed positive student perceptions of AI feedback usefulness, ease of use, and behavioral intention to continue using such tools. The findings make several theoretical contributions. First, they provide empirical support for Schmidt's Noticing Hypothesis by demonstrating that immediate feedback enhances error awareness and facilitates the noticing of linguistic gaps. Second, they extend Swain's Output Hypothesis by showing how technology-mediated feedback accelerates the noticing-restructuring cycle in oral production. Third, they contribute to the emerging body of literature on AI-assisted language learning by demonstrating AI feedback effectiveness in complex, meaning-oriented tasks, moving beyond scripted dialogues and pronunciation drills. Pedagogically, the study offers practical guidance for EFL instructors. Lightweight AI feedback can be integrated into literature-based curricula with minimal disruption, providing immediate, individualized support in large classes where teacher feedback is constrained by workload. The immediacy and consistency of AI feedback appear to be key advantages over delayed teacher feedback in authentic classroom contexts. Moreover, AI feedback appears to promote learner autonomy by encouraging self-monitoring and error correction behaviors. The study acknowledges several limitations, including the single institutional context, relatively small sample, absence of delayed post-tests, lack of fluency/complexity measures, and potential confounding of feedback source with timing. Future research should address these limitations through larger, multi-site studies, longer interventions with delayed post-tests, comparisons with immediate teacher feedback, integration of multiple proficiency measures, and investigation of learner characteristics as moderating variables. Qualitative research exploring how learners interact with AI feedback over time would complement the quantitative findings. Overall, lightweight AI feedback appears to be a promising, scalable tool for enhancing oral accuracy in EFL classrooms, particularly in literature-based courses where meaningful communication is prioritized but formal accuracy remains an important learning

objective. By providing immediate, consistent, and accessible form-focused feedback, AI technologies can support the development of accurate oral production while preserving the communicative and interpretive benefits of literature discussion.

E. Acknowledgement

We thank all friends who helped us in this meaningful paper.

References

Carter, R., & Long, M. (1991). *Teaching literature*. Longman.

Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). Oxford University Press.

Ellis, R., & Barkhuizen, G. (2005). *Analyzing learner language*. Oxford University Press.

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education*. Center for Curriculum Redesign.

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency* (pp. 1–2e0). John Benjamins.

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

Lazar, G. (1993). *Literature and language teaching*. Cambridge University Press.

Li, Y., & Hafner, C. A. (2022). Mobile-assisted language learning and AI-enhanced feedback. *Computer Assisted Language Learning*, *35*(7), 1508–1532. https://doi.org/10.1017/S0958344021000146

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

Paran, A. (2008). The role of literature in instructed foreign language learning and teaching. *Language Teaching*, *41*(4), 465–496. https://doi.org/10.1017/S026144480800520X

Ranalli, J., & Yamashita, T. (2022). Automated written corrective feedback: Error-correction performance and timing of delivery. *Language Learning & Technology*, 26(1), 1–25. https://doi.org/10.64152/10125/73465

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Newbury House.

Wen, Q. F., & Johnson, R. K. (1997). L2 learner variables and English achievement: A study of tertiary-level English majors in China. *Applied Linguistics*, 18(1), 27–48. https://doi.org/10.1093/applin/18.1.27