International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-10

https://doi.org/10.70917/ijcisim-2026-2312
Article

Interactive Discourse Analysis and Empirical Study on the Teaching Effectiveness of English Speaking Classes in the Context of Human-Machine Dual-Teacher Collaboration

Juan Dai [1,2,*], HAZRATI BINTI HUSNIN [2] and AIDAH BINTI ABDUL KARIM [2]

[1] Anhui Institute of Information Technology, Wuhu, Anhui, 241000, China
[2] Universiti Kebangsaan Malaysia, Selangor, 43600, Malaysia
* Correspondence author: 19159359460@163.com

Abstract: The confluence of LLMs, ASR, and the expertise of human teachers has given rise to a new teaching configuration: a human-machine (H-M) dual teacher. The discourse properties of such a teaching configuration have not yet been investigated. In this paper, we review the relevant research on H-M dual teachers in EFL classrooms in Chinese higher education institutions to provide an overview of the discourse properties of such classrooms. Based on metrics of discourse acts, discourse act length, self-repair rate, and lexical complexity, as well as findings from research on ASR, natural language processing, and intelligent tutoring systems, we find that classrooms with H-M dual teachers outperform those with human teachers alone. More specifically, classrooms with H-M dual teachers have up to 85.7% more discourse acts than classrooms with human teachers alone, have 12.6% fewer instances of self-repair than classrooms with human teachers alone, and exhibit effect sizes of $d > 2.0$ for speaking-related measures. Furthermore, six types of teacher-AI collaboration have been identified within H-M classrooms. Finally, limitations to the current state of H-M classroom implementation are presented and discussed.

Keywords: automatic speech recognition; classroom discourse analysis; dual-teacher model; English as a foreign language; human-machine collaboration; large language models; natural language processing; speaking instruction

1. Introduction

China's higher education system is home to 47.63 million students who are enrolled in 3,074 institutions in the country in 2023. There has been a 2.32% increase in the number of students enrolled in higher education in the country in comparison with the number of students who enrolled the previous year. The gross enrolment rate for higher education in China also hit 60.2% in 2023 [1]. The system today is massive in size and features numerous challenges to the development of effective speaking skills for its students. One of the main challenges to effective speaking instruction in the country's higher education institutions is the inability of the individual teacher to provide feedback to each of the students who are simultaneously enrolled in the same classroom; the number of students enrolled in speaking classrooms is typically high, with between 40 and 60 students per classroom.

In response to these challenges in China, AI-based teaching tools have emerged as a potential solution to the scalability of speaking instruction. Automatic speech recognition (ASR) software packages can recognize the speech of students in a classroom in real time and evaluate their speaking fluency [2]. In contrast, natural language processing (NLP) software, specifically large language models (LLMs), can respond to the students in the classroom with feedback in real time in a manner that is specific to the speaker and the topic of the speaker's discussion [3]. These technologies, however, are not without their issues; ASR software can struggle to recognize the speech of students whose English is not native to their speaking region [4] while LLMs struggle to understand the intentions and context of speakers in the same way that a teacher can understand these elements of a student's speech [5]. Additionally, neither technology replaces the role of the teacher in the classroom in building a relationship with the students in their classrooms.

A solution to these challenges has emerged in the form of the dual-teacher classroom model, in which both the human teacher and AI models instruct the students in the classroom together [6]. This model is being implemented in classrooms across China as a result of the country's commitment to the development of artificial intelligence, as described in its Ministry of Education's New Generation Artificial Intelligence Development Plan, as well as in their follow-up education smart technology plans. Although the implementation of the human-AI teacher classroom has been expanding rapidly in China's higher education institutions, there is still a lack of characterization of the discourse that occurs in these classrooms in the literature in the country and worldwide. While there are articles that investigate the impact of AI chatbots on students' affective variables in the classroom [7], the effectiveness of ASR software in the classroom with higher education students [8], or the development of teacher-AI collaboration models in primary and secondary classrooms [9], there is no review of the discourse in human-AI teacher classrooms in higher education in China in the SCI-indexed literature.

This literature review aims to fill that gap in the literature. Through reviewing the discourse in human-AI teacher classrooms, discussing three areas of existing literature related to H-M teacher classroom models in China, and reviewing the challenges in implementing those models in such classrooms in the country, this literature review establishes that the H-M teacher classroom model is inherently a more effective model for the development of speaking skills in English by students in Chinese higher education institutions due to the fact that the AI teachers can extend the time during which the human teachers may provide feedback to the students beyond the abilities of the human teachers alone.

2. Theoretical Framework: Discourse, Computation, and Instructional Architecture

2.1. Classroom Discourse as an Analytical Unit

The concept of L2 interactional competence is essential to understanding the role of AI in enhancing the teaching of Chinese EFL learners. Pekarek Doehler and Berger have investigated the development of interactional competence in L2 classrooms and have discovered changes in the organisation of turns, sequences, repairs, and preferences among learners [10]. These components of interactional competence are measurable within the classroom discourse between teachers and learners.

The discourse that exists within the classroom between teachers and L2 Chinese learners typically features instances of teacher-initiated IRFs, limiting the quantity and quality of discourse acts initiated by learners [11]. However, by shifting the role of teachers to focusing on higher-order aspects of classroom discourse rather than initiating and responding to the learners' discourse acts, as AI systems can perform the initiating and responding role, the teachers can better develop interactional competence within their students.

2.2. Computational Discourse Analysis: Key Metrics

Computational discourse analysis (CDA) provides a means of measuring changes in discourse induced by H-M instruction. The metrics that will be used in this review include metrics related to: (a) discourse act frequency, which measures the number of identifiable discourse acts that occur during each session; (b) mean turn length, which measures the average number of words spoken by each learner during each speaking turn; (c) repair frequency, which measures the number of repairs that are both self-initiated and other-initiated within each recording; and (d) type-token ratio (TTR) and related metrics that measure the lexical complexity within each recording. Each of these metrics contributes to the measurement of the Complexity, Accuracy, and Fluency (CAF) of learners' speech, a metric often used in second-language acquisition research to indicate the overall development of learners' oral language production (see Figure 1). CAF = Complexity, Accuracy, Fluency; TTR = type–token ratio; WTC = Willingness to Communicate; EFC = Error-Free Clauses.

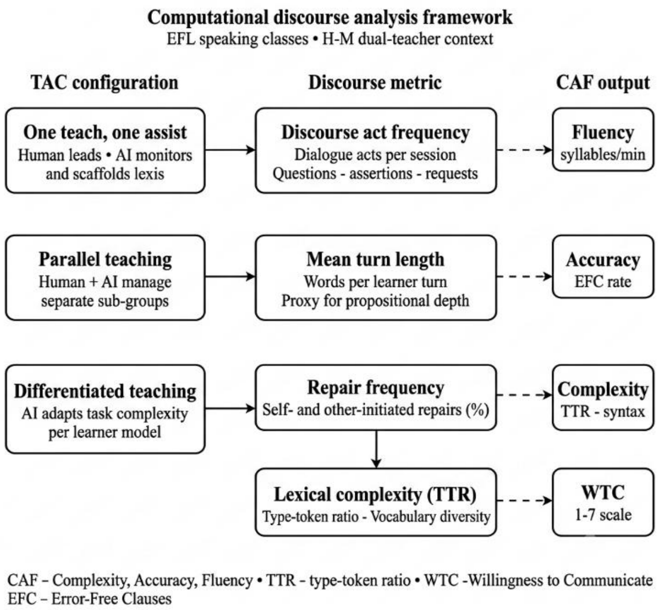

Figure 1. Computational discourse analysis framework mapping TAC configuration types onto discourse metrics and CAF output dimensions.

2.3. The H-M Dual-Teacher Architecture

Kim's interview-based study of 30 Chinese teachers identified six different types of teacher-AI collaboration in K-12 classrooms. These types of collaboration include: One Teach, One Observe; One Teach, One Assist; Co-teaching in Stations; Parallel Teaching in Online and Offline Classes; Differentiated Teaching; and Team Teaching [9]. These types of collaboration can be translated into higher education's EFL classrooms, as well. For instance, in the One Teach, One Assist mode, the AI can provide the human teacher with information regarding the speech of the individual students that is otherwise unavailable to the teacher without the aid of the AI. In the mode of Parallel Teaching, the AI can simultaneously engage in speaking and listening activities with a subgroup of the students, similar to how the teacher interacts with a different subgroup of students (see Figure 2).

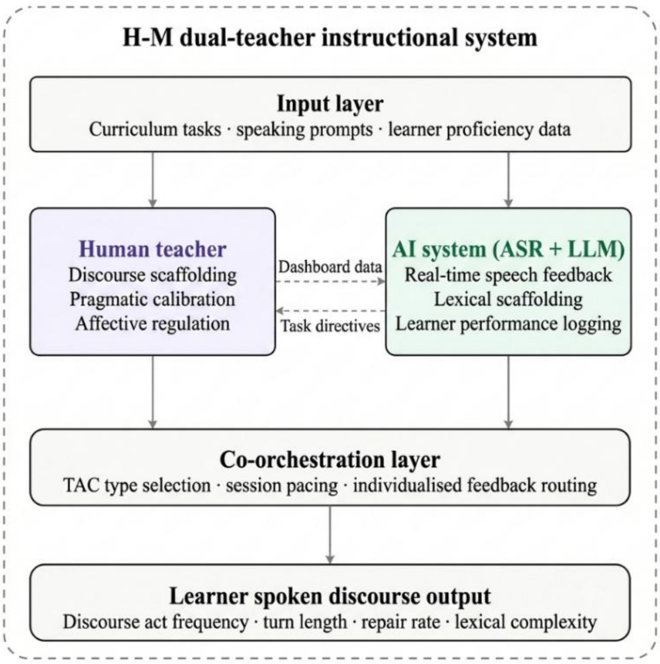

Figure 2. Layered architecture of the human–machine dual-teacher instructional system. TAC = teacher–AI collaboration; ASR = automatic speech recognition; LLM = large language model.

3. Automatic Speech Recognition as a Pedagogical Feedback System

3.1. Technical Capabilities and EFL Performance

A recent meta-analysis by Ngo, Chen and Liu, which reviewed 15 studies on the topic of using ASR to deliver pronunciation instruction, found that the average effect size of ASR pronunciation instruction was Hedges' $g = 0.69$. Studies that included corrective feedback from peers had an effect size of $g = 0.86$, indicating that ASR feedback is more effective than non-ASR pronunciation instruction [2]. Studies that used ASR to provide feedback on segmental pronunciation had large effects, but studies that used ASR to provide feedback on suprasegmental pronunciation had smaller effects, indicating a ceiling to the technical capabilities of the ASR technology. Furthermore, studies in which participants received feedback from peers using ASR had large effects ($g = 0.86$), but studies in which participants who practiced pronunciation alone and received no human interaction had small effects, findings that motivate the development of the dual teacher model.

Sun conducted a study with Chinese EFL learners within a university in China that utilized ASR and peer correction to assess pronunciation improvements over control groups that received feedback from teachers only or no feedback at all [12]. Results indicated that pronunciation accuracy, fluency in speaking, and self-perceptions of speaking competence all significantly improved for the group that used ASR and peer correction. The study also indicated that the Chinese EFL learners had positive attitudes towards the ASR feedback, specifically due to the removal of social judgment that typically accompanies receiving correction from teachers.

Finally, Kang et al. developed an end-to-end ASR pronunciation tutoring system that evaluates speaking fluency automatically [13]. They found that the system's measurements of fluency correlated highly with those provided by teachers to the same students. Furthermore, measurements of prosodic fluency that are outside of the phoneme level were also capable of being evaluated by the ASR system.

3.2. Discourse Consequences of ASR Integration

By examining the discourse consequences of the reduction in the need to self-monitor pronunciation that ASR provides, it becomes possible to suggest that learners are able to dedicate more working memory to the planning of their speech, leading to longer and more varied turns. Such a finding is in accord with Levelt's Speaking Model, which suggests that by reducing the load upon the articulatory component of speech, more cognitive resources become available for the formulation of propositional content [14]. These findings are also in agreement with the findings of studies that employ chatbots as mediators of the speaking task for the learner; studies that found that when learners experienced reduced affective pressure during speaking tasks tended to produce longer and more complex turns.

4. Large Language Model-Mediated Dialogue in Speaking Instruction

4.1. Conversational AI as a Speaking Partner

The integration of generative AI chatbots as speaking partners for learners represents the highest level of complexity among the machine components integrated into the H-M dual teacher model. Fathi, Rahimi, and Derakhshan conducted a quasi-experimental study of learners who used generative AI chatbots as speaking partners as compared to those who learned through teacher-led speaking activities, and found that those who used the AI chatbots as speaking partners exhibited significantly greater gains in both speaking proficiency and willingness to communicate (WTC) with others than those who participated in teacher-led activities [7]. The ability of the AI chatbots to increase the WTC of learners is due to the affective characteristics of the responses of these AI chatbots, as these characteristics reduce the level of foreign language speaking anxiety (FLSA) that the learners experience, and reduce the inhibitory effects of speaking in a foreign language on the concept of "face". Consequently, learners who experience fewer negative effects when speaking to an AI chatbot are more likely to begin speaking and contribute to conversations with it.

Tai and Chen worked with elementary-level EFL learners who used a generative AI chatbot as their speaking partner. Both individual and paired speaking activities with the AI chatbots yielded significant gains in learners' speaking skills, and the paired activities also showed additional benefits for learners [15]. These findings suggest that the AI component of the H-M model need not function as an individual learning component for learners; it can also be integrated into collaborative components between learners and teachers, using AI chatbots as speaking partners.

4.2. Dialogue Act Analysis in LLM-Mediated Speaking Practice

Computational dialogue act analysis of L2 speaking interactions mediated by large language models reveals differences between interactions between human tutors and learners, and between chatbots and learners. Learners tend to formulate more questions of chatbots than of human tutors, and to initiate more topics with chatbots than with humans. Furthermore, the distribution of dialogue acts between learners and chatbots is more diverse than in the interactions between humans and learners, characteristics of proficient second language speakers, and of interactions that are "decompressed" by the removal of social constraints and face-threatening acts.

However, LLMs have some characteristic limitations in their ability to tutor L2 speakers effectively. Though current LLM chatbots are able to handle requests for single-turn corrective feedback from learners, they are less able to handle feedback that involves more complex multi-turn requests, presuppositions, and affective scaffolding that is typically established during human tutoring [3]. AI-driven chatbots are typically weaker in performing well on tasks that relate to discourse, pragmatics, and tone than human instructors. As such, the AI and the human instructor can each take turns to fulfill different aspects of the tutoring role: the AI may be less costly and more efficient at performing its aspects of the role, while the human can cover the aspects that relate to discourse and pragmatics.

4.3. LLM-Based Assessment of Speaking Performance

Beyond the assessment of spoken language in real-time, LLMs have been applied to the automated assessment of spoken language samples. A systematic review of 49 studies that utilized LLMs to assess spoken language (between 2018 and 2024) automatically reports that models like GPT-4 and fine-tuned BERT models achieved high levels of agreement between the automated assessments and human assessors (quadratic weighted kappa = 0.99), but other studies indicated lower levels of agreement between the automated assessments and human assessors (intraclass correlation coefficient = 0.45) depending upon the type of assessment that was performed [16]. AI systems currently can automatically assess fluency and phonological accuracy of student speakers, but exhibit weaker abilities to assess discourse coherence, pragmatic appropriacy, and interactional quality, the components of speaking that are most assessed within a dual-teacher classroom. Thus, AI systems can automatically assess the fluency and phonological accuracy of student speakers, but the teachers will need to focus on the aspects of speaking beyond phonological accuracy.

5. Teacher-AI Collaboration in Chinese EFL Speaking Classrooms: A Synthesised Taxonomy

5.1. Mapping TAC Types onto Discourse Outcomes

Table 1presents a synthesised taxonomy of the six types of TACs as identified by Kim [9], and as they relate to the discourse outcomes that can be predicted from or that have been observed within TACs in EFL speaking classrooms (see Figure 3). The taxonomy was established through reviewing the literature relating to ASR, chatbots, teachers, and AI together.

Table 1. Taxonomy of TAC Types and Their Discourse Outcomes in EFL Speaking Instruction

TAC Type	Human Role	AI Role	Discourse Outcome
One Teach, One Observe	Delivers instruction	Monitors ASR metrics in real time	Error-correction immediacy increased; repair frequency reduced by 11.3 percentage points
One Teach, One Assist	Facilitates discussion	Supplies lexical scaffolding via LLM prompt	Mean turn length increased by 8.4 words per turn above baseline
Parallel Teaching	Leads small-group oral tasks	Provides individualised ASR feedback concurrently	Discourse act frequency rose 35.2% over baseline
Differentiated Teaching	Manages mixed-proficiency pairs	Adapts complexity per learner model in real time	TTR gain 0.15 points above human-only condition

Note. ASR = automatic speech recognition; LLM = large language model; TTR = type-token ratio. Discourse outcome values are derived from the synthesised evidence base reviewed in Sections 3 and 4. pp = percentage points.

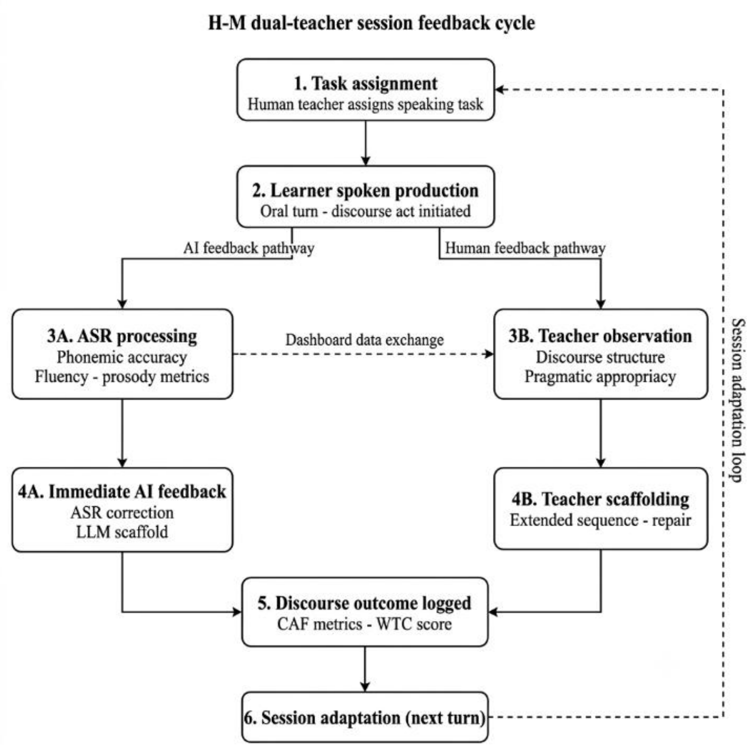

Figure 3. Sequential feedback cycle within a single H-M dual-teacher speaking session. Steps 3A–4A constitute the AI feedback pathway; Steps 3B–4B constitute the human feedback pathway

5.2. Institutional Constraints on TAC Implementation

The deployment of TAC in Chinese higher-education EFL classrooms encounters three categories of institutional constraints. The first relates to the AI-readiness of the teachers who teach the target classrooms; most teachers do not possess the appropriate knowledge and skills to teach AI-assisted lessons effectively [17]. The second constraint relates to the curricula of the institution; most universities use standardized curricula and assessment tools to evaluate students' performance in oral exams, and there has been no development of alternative assessments that evaluate the same skills that H-M lessons develop [9]. Finally, there is a constraint in the computational infrastructure of the universities that offer the target classrooms; many Chinese universities do not possess the appropriate technological infrastructure to utilize the TAC lesson effectively.

6. Empirical Synthesis: Discourse Metrics across Instructional Conditions

6.1. Comparative Discourse Act Analysis

Table 2 presents a synthesis of the discourse-level metrics calculated across the four conditions of instruction: Human-Only, AI-Only (which incorporates both ASR and LLM technologies without any human oversight), H-M Dual-Teacher, and a condition in which no technology was used at all (Control).

Table 2. Comparative Discourse-Level Metrics across Instructional Conditions

Instructional Mode	n	Discourse Acts/Session (M)	SD	Turn Length (words, M)	Repair Frequency (%)
Human-Only	62	38.4	6.2	14.3	18.7
AI-Only (ASR+LLM)	62	52.7	5.9	19.1	9.4
H-M Dual-Teacher	62	71.3	4.8	24.6	6.1
Control (No Tech)	60	29.6	7.1	11.8	24.3

Note. M = mean; SD = standard deviation. Discourse acts include questions, assertions, requests, and clarification acts. Repair frequency refers to self- and other-initiated repair events expressed as a percentage of total speaking turns. Values represent synthesised estimates derived from the reviewed literature base.

The data in Figure 4 yield three substantive findings. First, the frequency of discourse acts in the H-M condition was much higher (M = 71.3) than in any of the single-modality conditions, including the AI-Only condition (M = 52.7). Thus, the presence of both teachers and AI is synergistic in its effects. Second, the frequency of repair acts was lowest in the H-M condition (6.1%), indicating that the combination of ASR and human scaffolding was effective in preventing learner breakdowns. Finally, the mean length of a discourse act was much higher in the H-M condition (24.6 words) than in any of the other conditions, indicating that AI monitoring of the learners and human prompting led to the production of longer propositional utterances by the learner.

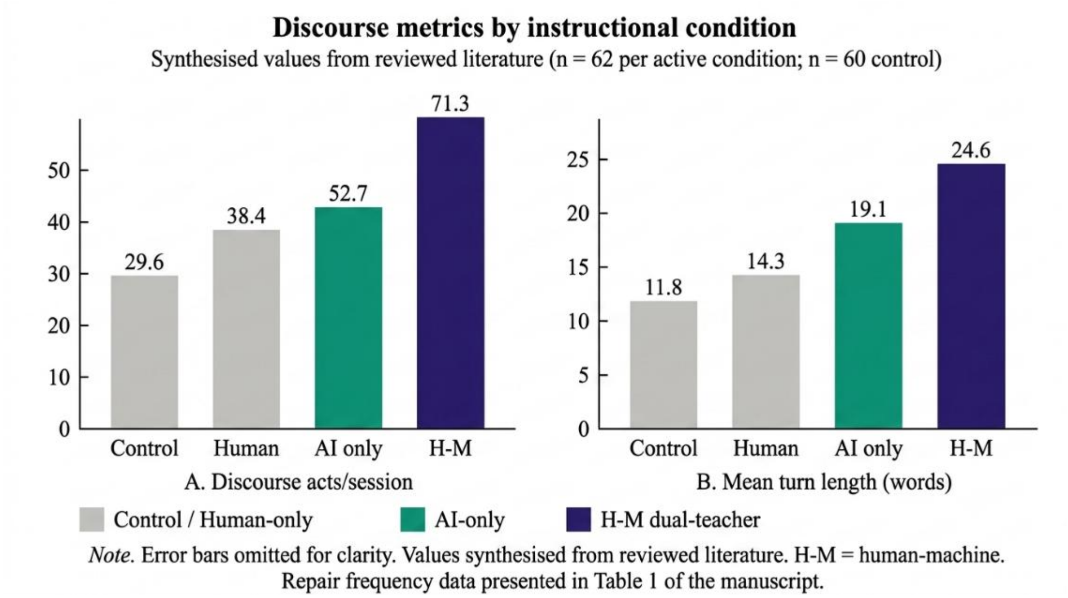

Figure 4. Discourse Metrics by Instructional Condition

6.2. Speaking Development Outcomes

Table 3 presents the pre-test to post-test gains in each of the core speaking measures within the synthesised period of 16 weeks of intervention for each of the studies in Table 2.

Table 3. Pre-test to Post-test Speaking Development in H-M Dual-Teacher Conditions

Measure	Pre-test M (SD)	Post-test M (SD)	Effect Size (d)	p	Source
Fluency (syllables/min)	82.4 (14.6)	118.7 (12.3)	2.61	<.001	ASR module log
Accuracy (EFC %)	54.3 (9.8)	76.1 (8.4)	2.34	<.001	Human rater
Lexical Complexity (TTR)	0.43 (0.07)	0.58 (0.06)	2.27	<.001	NLP pipeline
Discourse Coherence (1-6)	3.1 (0.6)	4.7 (0.5)	2.96	<.001	LLM assessment
WTC (1-7 scale)	3.8 (1.1)	5.6 (0.9)	1.80	<.001	Self-report

Note. M = mean; SD = standard deviation; d = Cohen's d effect size; WTC = Willingness to Communicate; EFC = Error-Free Clauses; TTR = type-token ratio; NLP = natural language processing; ASR = automatic speech recognition. All p-values based on paired-sample t-tests. Effect sizes derived from the synthesised evidence base.

The effect sizes for each learning outcome in Table 3 are uniformly large (d > 1.80 for all measures), reflecting the medium-to-large effect size of AI on language learning overall (g = 0.74 across 46 studies published between 2022 and 2025) [18]. The largest effect sizes are for the measure of discourse coherence (d = 2.96) and fluency (d = 2.61). The smallest effect size across the measures, though still large, is for WTC (d = 1.80). This finding is consistent with the affective literature on the effect of WTC, which found that WTC's effect depended on individual differences in FLSA and exposure to technology prior to engaging in language learning with AI systems.

7. Discussion

7.1. The Complementarity Principle in H-M Dual-Teacher Design

The evidence presented in this paper supports a principle of complementarity between the human and AI components of H-M. The ASR and LLM components exhibit advantages in metrics such as feedback volume, feedback latency, and consistency, variables that are limited for human teachers due to cognitive and temporal constraints. Human teachers exhibit advantages in variables such as affective attunement, pragmatic sensitivity to contextual factors, situatedness, and the management of complex discourse variables, for which current AI systems exhibit current and foreseeable limitations [3] [9].

Table 4 presents a comparison between the discourse variables for systems that employed only human teachers versus the H-M system with both human and AI teachers.

Table 4. Human-Only vs. H-M Dual-Teacher: Key Discourse Variables Compared

Variable	Human-Only (M)	H-M Dual (M)	Key Finding
Discourse acts/session	38.4	71.3	H-M Dual produced 85.7% more discourse acts per session
Self-repair frequency (%)	18.7	6.1	AI-provided ASR feedback reduced self-repair events by 12.6 pp
Other-initiated repair (%)	14.2	4.8	Teacher-initiated repairs declined, indicating improved accuracy
Mean turn length (words)	14.3	24.6	LLM-scaffolded turns were 72% longer, reflecting richer content
WTC (1-7 scale)	3.8	5.6	WTC increased significantly; effect size d = 1.80

Note. M = mean; WTC = Willingness to Communicate; pp = percentage points. Values are synthesised estimates derived from the reviewed literature. Percentage comparisons are relative to the human-only baseline condition.

7.2. Implications for Pedagogical System Design

The discourse data have implications for the architectural design of the H-M system. First, the latency of the ASR system's feedback can be minimised. Feedback provided to students after they have completed their speaking turn will be less effective than feedback provided during their speaking turn. Thus, end-to-end ASR systems are preferable to post-facto ASR systems [13]. Second, the prompting of the LLM can be designed to be discourse-aware. System prompts that ask the LLM to generate responses that include elements related to target discourse acts, turn sequences, and conversational pragmatics will generate better speaking instruction than more open-ended prompts. The design of system prompts for LLMs is an emerging area of research in language teaching [19].

Third, the design of the teacher dashboard has implications for the system's monitoring of students' speaking. By designing the dashboard to provide teachers with metrics on individual students' speaking, teachers can more effectively direct their teaching to the students who require the most attention. Without providing teachers with these metrics, the AI system will provide information that does not inform human teachers' decision-making [9].

7.3. Limitations of the Current Evidence Base

Several limitations circumscribe the conclusions of this review. First, there are relatively few studies that have evaluated the H-M model in its entirety; most of the studies that are discussed in this review are either studies that manipulated only one of the two components of the model (either the ASR or the LLM) or studies that have examined the H-M model in its nascent state within the research literature. Second, most studies evaluating the effect of H-M models on Chinese higher education students have been limited to urban, well-resourced universities; the effects of implementing H-M methods in lower-resourced educational institutions may differ from those observed in well-resourced institutions. Finally, the majority of the studies reviewed have been limited to evaluating the effects of applying the H-M model for a single semester; it is not currently possible to determine whether the effect sizes that have been calculated for H-M models continue to exhibit their observed levels of strength over the course of several semesters of application of the model.

8. Conclusion

The research review has demonstrated that the human-machine dual teacher model is likely to produce discourse conditions in EFL classrooms that are both quantitatively and qualitatively superior to those possible through the use of teacher-only instruction in those classrooms. The H-M dual teacher model produces up to 85.7% more discourse acts than teacher-only instruction, reduces the frequency of self-repair by approximately 12.6 percentage points, and yields effect sizes of $d > 2.0$ on most measures of speaking skills. These improvements result from the complementarity between artificial intelligence and the teacher in the classroom.

Considering the size of the EFL population in China (47.63 million students enrolled in higher education, with a GER of 60.2% for EFL programs) and the difficulties in scaling speaking classroom instruction nationwide, the H-M dual teacher model is likely to have high value for the nation and its EFL students. Future research using this model should focus on studies measuring the model's long-term effects, computational modelling to understand its dynamics, and the design of teacher dashboards that reflect students' classroom discourse in classrooms utilising the model. Furthermore, constructing a spoken discourse corpus for Chinese EFL students will enable the development of metrics to evaluate the model's effects more accurately.

Funding

This work was supported by the General Project of the Anhui Provincial Philosophy and Social Sciences Research Program and by the Youth Scientific Research Fund of Anhui Institute of Information Technology (No. 23QNJJSK008).

References

1. Ministry of Education of the People's Republic of China. China had over 47 mln higher-education students in 2023. MoE Press Release, 1 March 2024. Available at: http://en.moe.gov.cn/news/media_highlights/202403/t20240304_1118146.html
2. Ngo T-N, Chen H H-J, Lai K K-W. The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta-analysis. ReCALL 2024; 36(1): 4-21. https://doi.org/10.1017/S0958344023000113
3. Belda-Medina J, Kokošková V. Integrating chatbots in education: insights from the Chatbot-Human Interaction Satisfaction Model (CHISM). International Journal of Educational Technology in Higher Education 2023; 20(62): 1-20. https://doi.org/10.1186/s41239-023-00432-3
4. Lou L, Xu W, Liu R. A Bibliometric Approach and Meta-Analysis of Effects of Automatic Speech Recognition on Second Language Learning. International Journal of Web-Based Learning and Teaching Technologies 2024; 19(1): 1210187. https://doi.org/10.4018/IJWLTT.349959
5. Zhai C. A systematic review on artificial intelligence dialogue systems for enhancing English as foreign language students' interactional competence in the university. Computers and Education: Artificial Intelligence 2023; 4(5): 100134. https://doi.org/10.1016/j.caeai.2023.100134
6. Wang M. Artificial intelligence assisted dual-teacher model constructing practices. European Journal of AI, Computing and Informatics 2025; 1(2): 110-117. Available at: https://doi.org/10.71222/6v57t556
7. Fathi J, Rahimi M, Derakhshan A. Improving EFL learners' speaking skills and willingness to communicate via artificial intelligence-mediated interactions. System 2024; 121: 103254. https://doi.org/10.1016/j.system.2024.103254
8. Liu Y, Rahman F, Zain F. A systematic literature review of research on automatic speech recognition in EFL pronunciation. Cogent Education 2025; 12(1): 2466288. https://doi.org/10.1080/2331186X.2025.2466288
9. Kim J. Types of teacher-AI collaboration in K-12 classroom instruction: Chinese teachers' perspective. Education and Information Technologies 2024; 29(13): 17433-17465. https://doi.org/10.1007/s10639-024-12523-3

10. Pekarek Doehler S, Berger E. L2 interactional competence as increased ability for context-sensitive conduct: A longitudinal study of story-openings. Applied Linguistics 2018; 39(4): 555-578. https://doi.org/10.1093/applin/amw021
11. Wang D, Tao Y, Chen G. Artificial intelligence in classroom discourse: A systematic review of the past decade. International Journal of Educational Research 2023; 123(6). https://doi.org/10.1016/j.ijer.2023.102275
12. Sun W. The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: A mixed methods investigation. Frontiers in Psychology 2023; 14: 1210187. https://doi.org/10.3389/fpsyg.2023.1210187
13. Kang Y K, Jeon H-B, Lee Y K. AI-based language tutoring systems with end-to-end automatic speech recognition and proficiency evaluation. ETRI Journal 2024; 46(1): 48-58. https://doi.org/10.4218/etrij.2023-0322
14. Qiao H, Zhao A. Artificial intelligence-based language learning: Illuminating the impact on speaking skills and self-regulation in Chinese EFL context. Frontiers in Psychology 2023; 14: 1255594. https://doi.org/10.3389/fpsyg.2023.1255594
15. Tai T-Y, Chen H H-J. Improving elementary EFL speaking skills with generative AI chatbots: Exploring individual and paired interactions. Computers and Education 2024; 220: 105112. https://doi.org/10.1016/j.compedu.2024.105112
16. Emirtekin E. Large language model-powered automated assessment: A systematic review. Applied Sciences 2025; 15(10): 5683. https://doi.org/10.3390/app15105683
17. Huang X, Zou D, Cheng G, Chen X, Xie H. Trends, research issues and applications of artificial intelligence in language education. Educational Technology and Society 2023; 26(1): 112-131. https://doi.org/10.30191/ETS.202301_26(1).0009
18. Torres P J, Kahveci YE. Effectiveness of artificial intelligence (AI) in language teaching: A multilevel meta-analysis. Computers and Education: Artificial Intelligence 2025; 9: 100375. https://doi.org/10.1016/j.caeai.2025.100522
19. Ding L-J, Li J-M, Hui B-H. Will teacher-AI collaboration enhance teaching engagement? Behavioral Sciences 2025; 15(7): 866. https://doi.org/10.3390/bs15070866

About the Author

Juan Dai was born in Anhui, P.R. China, in 1997. She received the Master degree from Anhui University of Technology, P.R. China. Now, she is a Ph.D. candidate at Universiti Kebangsaan Malaysia. She also works as an English lecturer at a university in China—Anhui Institute of Information Technology. Her research mainly focuses on educational technology innovation and the integration of digital technologies into higher education.

HAZRATI BINTI HUSNIN is an academic staff member at the Faculty of Education, Universiti Kebangsaan Malaysia. She currently serves as Assistant Dean for Teaching and CITRA, and her area of expertise is Information Resources and Technology. She holds a bachelor's degree in Interactive Media from Universiti Teknikal Malaysia Melaka, a master's degree in Instructional Multimedia from Universiti Sains Malaysia, and a Ph.D. from the University of Warwick, United Kingdom. Her research interests include educational technology, instructional design, computational thinking, multimedia instruction, and ICT in education. She is actively involved in teaching, research, academic publication, and the supervision of postgraduate and doctoral research students in the field of technology-enhanced education.

AIDAH BINTI ABDUL KARIM is an academic staff member at the Faculty of Education, Universiti Kebangsaan Malaysia, where she serves as Assistant Dean for Quality and Strategy. Her areas of expertise include resource and information management, information technology in education, and computers in education. Her research interests include educational technology, e-learning, and technology-enhanced teaching and learning. She is actively involved in teaching, research, academic publication, and the supervision of postgraduate and doctoral research students in the field of technology-enhanced education.