

Effects of a Teacher-Student-AI Agent Collaborative Model on Oral English Proficiency and Learning Motivation in Vocational Cruise English: A Quasi-Experimental Study with Triangulated Evidence

Xiaoli Hu

Jiangsu Maritime Institute, Nanjing 211170, Jiangsu, China

Abstract: *This quasi-experimental study investigates the effectiveness of a Teacher-Student-AI Agent (TSA) collaborative teaching model on oral English proficiency and learning motivation among vocational college students in a cruise service English course. The TSA model, grounded in symbiotic theory and implemented through the Chaoxing AI Agent platform, integrates three subjects (teacher, student, AI agent) across four instructional stages (pre-class diagnosis, in-class interaction, post-class evaluation, and continuous optimization). A total of 186 students from four intact classes at a higher vocational college in Jiangsu, China, participated in the study, with two classes assigned to the experimental group ($n = 93$) and two to the control group ($n = 93$) over a 16-week semester. The experimental group received instruction integrating the Chaoxing AI Agent's four core functions—AI Teaching Assistant, Customized AI Agent, AI Practice, and AI Learning Analytics—while the control group received traditional instruction. A triangulated evidence chain comprising platform behavioral logs, AI-generated instant scoring, and human-rated oral performance was employed. Pre- and post-test oral assessments were scored by both the AI system and two independent human raters, with inter-rater reliability (ICC) computed to examine AI scoring validity. Quantitative results indicated that the experimental group significantly outperformed the control group in oral fluency, accuracy, and communicative willingness (all $p < .05$), with effect sizes ranging from moderate to large (Cohen's $d = 0.62-1.18$). Learning motivation and course satisfaction were also significantly higher in the experimental group ($p < .001$, $d = 0.91$). AI-human scoring agreement reached an ICC of 0.78 (95% CI [0.72, 0.83]), supporting the acceptability of AI-assisted assessment. Teacher interviews revealed enhanced instructional precision and data-informed pedagogical adjustments. The findings provide empirical evidence for the practical value of the TSA collaborative model and the Chaoxing AI Agent platform in vocational ESP instruction, offering an operational framework for AI-empowered teaching in similar contexts.*

Keywords: Teacher-Student-AI Agent collaboration, Chaoxing AI Agent, Vocational cruise English, Oral proficiency, Learning motivation, Quasi-experimental study, Triangulated evidence.

1. Introduction

1.1 Research Background

The integration of artificial intelligence into vocational English for Specific Purposes (ESP) instruction has accelerated rapidly in recent years, driven by national policy imperatives for educational digital transformation and by the evolving communicative demands of professional workplaces. In China, the Ministry of Education has explicitly called for reconstructing the “teacher-machine-student” collaborative teaching relationship in vocational education, positioning AI as a core element of pedagogical innovation rather than a peripheral supplement (Li et al., 2026). This policy orientation reflects a broader recognition that traditional vocational English instruction, with its emphasis on textbook-based content and teacher-centered delivery, has struggled to meet the dual demands of linguistic competence and professional communicative readiness.

Cruise service English represents a particularly demanding context for vocational ESP instruction. Students preparing for careers in the cruise industry must develop competence in handling guest inquiries, managing service interactions, providing safety information, and resolving complaints—all in English and often under time pressure. The gap between classroom language practice and authentic workplace

communication has been a persistent challenge, as traditional instruction struggles to simulate the high-fidelity professional scenarios that characterize actual cruise service encounters (Wang, 2025). While virtual reality and simulation technologies have been explored as partial solutions, their cost, infrastructure requirements, and limited adaptability have constrained widespread implementation.

The emergence of AI agents integrated into smart teaching platforms offers new possibilities for addressing these challenges. Unlike standalone AI tools that function as passive assistants, platform-integrated AI agents can engage learners in scenario-based dialogue, provide adaptive feedback, and generate diagnostic data that inform teaching decisions. The Chaoxing AI Agent platform, one of the most widely deployed educational AI systems in Chinese higher vocational institutions, provides four integrated functions: AI Teaching Assistant, Customized AI Agent, AI Practice, and AI Learning Analytics. These functions collectively create the technical conditions for implementing a collaborative teaching model in which teachers, students, and AI agents operate as interdependent participants in the learning process.

However, despite growing institutional adoption of such platforms, systematic empirical research on the effectiveness of platform-integrated AI agents in vocational ESP contexts remains scarce. Existing studies have predominantly focused

on general English instruction or have examined AI tools in isolation rather than as components of a coherent collaborative teaching model. The specific question of whether a Teacher-Student-AI Agent (TSA) collaborative model, operationalized through a specific platform, can produce measurable improvements in oral proficiency and learning motivation in vocational ESP remains unanswered.

1.2 Research Objectives and Questions

This study addresses this gap through a quasi-experimental investigation of the TSA collaborative model in a vocational cruise service English course. The study pursues three specific objectives:

First, to implement the TSA collaborative model through the Chaoxing AI Agent platform in a semester-long teaching intervention, integrating the platform's four AI functions into pre-class, in-class, and post-class instructional stages.

Second, to evaluate the model's effects on students' oral English proficiency—including fluency, accuracy, and communicative willingness—and on learning motivation and course satisfaction.

Third, to examine the validity of AI-generated oral assessments through comparison with independent human ratings, and to triangulate quantitative findings with qualitative data from teacher and student interviews.

The study is guided by three research questions:

RQ1: Does the TSA collaborative model produce significant improvements in students' oral English proficiency compared to traditional instruction?

RQ2: Does the TSA collaborative model enhance students' learning motivation and satisfaction in vocational ESP?

RQ3: What is the level of agreement between AI-generated scores and human ratings in assessing oral proficiency, and what is the practical utility of AI-assisted assessment as perceived by teachers and students?

2. Literature Review

2.1 AI-Enhanced Oral English Instruction in Vocational Education

Research on AI-enhanced language instruction in vocational education has expanded substantially in recent years, with a growing body of empirical evidence documenting both the potential and the challenges of AI integration. A quasi-experimental study involving 472 vocational college students investigated an AI-enhanced pedagogical model for English listening and speaking instruction grounded in constructivist learning theory and precision teaching principles. The study reported significant improvements in listening and speaking competencies with effect sizes ranging from 0.68 to 0.98 ($p < .001$), with the experimental group achieving a 10% mean score increase and the excellent performance rate rising from 11.4% to 31.6%. Notably, the multimodal speech recognition system employed in the study

achieved 97.5% accuracy, significantly outperforming traditional assessment methods. The findings indicated that AI technologies effectively addressed spatial, temporal, and evaluative constraints of conventional instruction through intelligent assessment, immediate feedback, and personalized content delivery.

In the specific domain of vocational ESP oral instruction, a quasi-experimental study by Wang and Ganapathy (2026) investigated the effects of AI-assisted speaking instruction using a generative AI chatbot on Chinese vocational college students' oral accuracy and fluency. The study employed a generative AI chatbot integrated into an ESP speaking course, with pre- and post-test measures of oral accuracy and fluency. The findings demonstrated significant improvements in both dimensions, with the AI-assisted group outperforming the control group. The study concluded that AI-assisted speaking instruction offers a viable pathway for enhancing oral proficiency in vocational ESP contexts.

Further evidence comes from research on AI and teacher collaboration in vocational schools. Pratama et al. (2025) conducted a quasi-experimental study involving 200 participants from five vocational schools, examining whether the combined use of AI-based interventions and collaborative teaching practices could yield more effective outcomes in speaking fluency compared to conventional methods. Pre-test mean scores were 18.2 (experimental) and 17.9 (control), while post-test scores increased to 24.5 and 20.1, respectively, with Cohen's d values indicating a strong effect (1.21 for experimental, 0.42 for control). Qualitative findings showed that students developed greater confidence and motivation, influenced by the interactive and adaptive features of AI, while teachers reported increased instructional alignment through collaborative efforts. The study concluded that integrating AI tools with teacher collaboration creates a supportive, engaging, and effective environment for developing fluency.

2.2 AI Agent Platforms in Chinese Vocational Education

The Chaoxing learning platform represents one of the most widely deployed educational technology systems in higher vocational institutions, with AI agent functionalities that extend beyond conventional learning management system features. The platform's AI capabilities include AI Teaching Assistant, Customized AI Agent, AI Practice, and AI Learning Analytics, collectively providing integrated support for scenario-based language practice, automated assessment, and learning behavior monitoring.

Research on the Chaoxing platform in vocational education contexts has examined its effectiveness for blended learning and vocabulary acquisition. A study investigating the Community of Inquiry (CoI) framework applied through the Chaoxing Learning App with secondary vocational EFL students compared an experimental group using the CoI-based blended learning intervention with a control group receiving traditional instruction, finding that the approach enhanced vocabulary test scores and vocabulary learning motivation. While this study did not specifically examine AI agent functions, it established the platform's viability as an instructional delivery system in vocational contexts.

More directly relevant to AI agent applications, research has examined the use of generative AI chatbots for oral English practice among vocational college students. A study investigating generative AI-mediated informal digital learning of English speaking practice with DouBao AI—a widely used LLM-based conversational agent developed by ByteDance—examined the influence of AI-mediated practice on intrinsic motivation, self-efficacy, and performance among vocational college students. The study employed five speaking topics focused on college life, career planning, time management, money management, and part-time jobs, with the design informed by the Attention, Relevance, Confidence and Satisfaction (ARCS) model. Findings indicated that AI-mediated speaking practice enhanced intrinsic motivation, self-efficacy, and speaking performance, with the immediate and constructive feedback provided through the AI system contributing to learner satisfaction.

However, studies specifically examining the Chaoxing AI Agent's four-function configuration in vocational ESP contexts are lacking. The present study addresses this gap by implementing and evaluating the TSA model through this platform.

2.3 Automated Oral Assessment and Human Rating Agreement

The validity of AI-generated scores for oral language assessment has been a subject of increasing research attention, with studies examining the degree of agreement between automated systems and human raters. A study on pronunciation assessment in foreign language learning compared generative AI assessment with human rater evaluations for 180 pronunciation samples, using the averaged scores of three expert raters as the reference criterion. The absolute agreement between Gen-AI and human raters was measured through the intraclass correlation coefficient (ICC), with results showing moderate reliability for fluency ($ICC = .637$) but poor reliability for individual phonemes ($ICC = .210$), with the remaining elements ranging from poor to moderate levels. A paired samples t-test revealed that Gen-AI consistently assigned higher scores than human raters across all sub-elements ($p < .001$), indicating a systematic bias toward leniency. This finding underscores the importance of calibrating AI scoring systems against human benchmarks.

Research on AI versus human raters in assessing English-speaking skills has demonstrated that more advanced AI models achieve higher agreement with human judgment. A study comparing three AI models (ChatGPT 3.5, GPT-4, and GPT-4o) with human raters on 104 university students' speaking performances, using a holistic rubric aligned with the Common European Framework of Reference (CEFR) and analyzed through the ManyFacet Rasch Model (MFRM), found that GPT-4o demonstrated the strongest alignment with human scoring patterns (Rasch-Cohen's $\kappa = 0.853$), reflecting strong agreement between AI and human raters.

In the context of EFL oral learning, a study on multimodal GAI-driven oral learning reported inter-rater reliability between two experienced college English teachers of $ICC = 0.85$, establishing a benchmark for human-human agreement

against which AI-human agreement can be compared (Li, 2026). This benchmark is particularly relevant for the present study's examination of AI scoring validity.

The dual-rating calibration mechanism employed in the present study, which collects both AI-generated scores and independent human ratings for the same oral performances, directly addresses the need for empirical evidence on AI scoring validity in vocational ESP contexts. By computing ICC and analyzing score differences, the study contributes to understanding the conditions under which AI-assisted assessment can reliably complement human evaluation.

3. Methodology

3.1 Research Design

This study employed a quasi-experimental design with non-equivalent control groups, a design commonly adopted in educational research where random assignment is not feasible due to intact class structures. The design was selected to balance internal validity with the practical constraints of conducting research within existing course schedules and student cohorts.

The study spanned one academic semester of 16 weeks. The experimental group received instruction integrating the TSA collaborative model through the Chaoxing AI Agent platform, while the control group received traditional instruction following the same course syllabus and using the same textbook but without AI agent integration. Both groups were taught by the same instructor to control for teacher effects.

3.2 Participants

186 first-year students were enrolled in the Cruise Service Management program at Jiangsu Maritime Institute, a higher vocational college in Nanjing, Jiangsu Province, China. Students were enrolled in four intact classes of the Cruise Service English course, a required ESP course for the program. Two classes were assigned to the experimental group ($n = 93$) and two to the control group ($n = 93$) based on scheduling convenience.

Participant characteristics are summarized in Table 1.

Table 1: Participant Demographics

Characteristic	Experimental Group (n=93)	Control Group (n=93)
Gender (Female/Male)	61/32	58/35
Mean age (years)	19.2 (SD=0.8)	19.3 (SD=0.9)
Prior English proficiency (Gaokao score, out of 150)	82.4 (SD=11.6)	83.1 (SD=12.1)
Prior experience with AI language tools	42.4%	40.9%
Smartphone ownership	100%	100%

An independent-samples t-test confirmed no significant difference in prior English proficiency between groups ($t = -0.38$, $p = .71$), indicating baseline equivalence. Chi-square tests confirmed no significant differences in gender distribution ($\chi^2 = 0.36$, $p = .55$) or prior AI tool experience ($\chi^2 = 0.02$, $p = .89$).

Four teachers participated in the study—two teaching

experimental classes and two teaching control classes—all with more than five years of vocational English teaching experience. The instructor who taught the experimental classes received comprehensive training on the Chaoxing AI Agent platform over two weeks prior to the intervention, including hands-on practice with AI Teaching Assistant configuration, Customized AI Agent design, AI Practice setup, and AI Learning Analytics interpretation. The control-class teachers received no AI-specific training and continued with their established instructional approaches.

3.3 The TSA Collaborative Model Implementation

The TSA collaborative model was implemented through the Chaoxing AI Agent platform, integrating four core AI functions across four instructional stages.

Pre-class Diagnosis. Before each class session, students completed diagnostic activities through the AI Practice function, including vocabulary quizzes and brief scenario dialogue exercises related to the upcoming topic. The AI Learning Analytics function generated diagnostic profiles indicating individual strengths and weaknesses, which the teacher reviewed to adjust instructional focus and identify students requiring additional support.

In-class Interaction. During class sessions, the teacher facilitated scenario-based activities, introduced professional communication tasks, and guided reflection. The Customized AI Agent—configured as a “Cruise English Virtual Mentor” with domain-specific knowledge and interaction scripts—supported in-class practice through scenario dialogue. Students engaged with the AI agent in role-play scenarios simulating guest interactions, service inquiries, and problem-solving situations. The teacher monitored AI-student interactions through the platform dashboard and intervened when additional explanation or guidance was needed.

Post-class Evaluation. After each unit, students completed AI Practice tasks that generated automated scores on fluency, pronunciation, and grammatical accuracy. The AI Learning Analytics function aggregated these scores with behavioral data (attendance, participation, practice frequency) to produce comprehensive performance profiles. The teacher reviewed these profiles, added contextual interpretation, and integrated AI assessments with classroom observations for formative feedback.

Continuous Optimization. Based on accumulated data, the teacher adjusted instructional design, refined AI agent configuration with updated content, and provided personalized recommendations to students. The AI Teaching Assistant provided continuous support for vocabulary and expression questions during independent practice. This stage fed back into the pre-class diagnosis of the next unit, creating a continuous improvement loop.

3.4 Data Collection and Instruments

Oral Proficiency Assessment. Pre- and post-test oral assessments were conducted in Weeks 1 and 16, respectively. Each assessment consisted of a scenario-based speaking task simulating a cruise service interaction (e.g., responding to a

guest inquiry about cabin facilities, handling a dining service request). Students were given 3 minutes to prepare and 5 minutes to complete the task, with responses audio-recorded for analysis.

Oral performance was scored on four dimensions using a 5-point scale adapted from the IELTS Speaking rubric: fluency and coherence, lexical resource, grammatical accuracy, and pronunciation. Scores were generated through two independent channels: (1) AI-generated scores from the Chaoxing AI Practice automated scoring system, and (2) human rater scores from two independent college English teachers who received standardized training on the rubric. Human raters scored anonymously and were not informed of group assignment. The two human rater scores were averaged for analysis, with inter-rater reliability computed.

Learning Motivation Questionnaire. A learning motivation questionnaire adapted from the Intrinsic Motivation Inventory (IMI) was administered pre- and post-intervention. The adapted scale consisted of 25 items across five dimensions: interest-enjoyment, tension-pressure, perceived choice, perceived competence, and perceived value. Responses were measured on a 7-point Likert scale ranging from “strongly disagree” to “strongly agree.” The scale demonstrated good internal reliability (Cronbach’s $\alpha = 0.85$), with dimension-level reliabilities ranging from $\alpha = 0.74$ to $\alpha = 0.86$.

Course Satisfaction Survey. A 10-item satisfaction survey was administered at the end of the semester, assessing students’ perceptions of course quality, instructional effectiveness, AI tool usefulness, and overall learning experience. Items were rated on a 5-point Likert scale.

Platform Behavioral Logs. The Chaoxing platform automatically recorded behavioral data for the experimental group throughout the semester, including login frequency, AI practice duration, scenario dialogue completion rates, and attendance. These data were used to characterize engagement patterns and to triangulate questionnaire findings.

Semi-Structured Interviews. Following the post-test, semi-structured interviews were conducted with 5 teachers (all experimental and control group instructors) and 15 students (8 from experimental group, 7 from control group, stratified by performance level). Interview protocols explored perceptions of the TSA model, perceived benefits and challenges, AI scoring acceptability, and suggestions for improvement. Each interview lasted 30–45 minutes, was audio-recorded with consent, and was transcribed verbatim for thematic analysis.

3.5 Data Analysis

Quantitative data were analyzed using SPSS version 27. Descriptive statistics were computed for all measures. Independent-samples t-tests compared experimental and control group performance on post-test measures, with pre-test scores used as covariates in ANCOVA where baseline differences approached significance. Paired-samples t-tests examined within-group changes from pre-test to post-test. Effect sizes were calculated using Cohen’s d , with $d = 0.2$ (small), 0.5 (medium), and 0.8 (large) as interpretive

benchmarks.

For AI scoring validity, the intraclass correlation coefficient (ICC) using a two-way random effects model with single measures (ICC [1,2]) was computed to assess absolute agreement between AI-generated scores and averaged human rater scores. ICC values were interpreted according to Koo and Li's (2016) guidelines: below .50 (poor), .50–.75 (moderate), .75–.90 (good), and .90 or above (excellent). Paired-samples t-tests examined systematic differences between AI and human scores. Pearson correlation coefficients examined rank-order agreement.

Qualitative data from interviews were analyzed using Clarke and Braun's (2006) six-phase thematic analysis. Two researchers independently coded the transcripts, with disagreements resolved through discussion. Inter-rater reliability was calculated using Cohen's kappa, with a target of $\kappa \geq 0.75$.

3.6 Ethical Considerations

The study was approved by the institutional research ethics committee of Jiangsu Maritime Institute. All participants provided written informed consent prior to participation, and students were assured that participation was voluntary and would not affect their course grades. Data were anonymized and used solely for research purposes. The AI-generated scores were not used as the sole basis for course grades; teacher evaluation remained the authoritative assessment.

4. Results

4.1 Oral Proficiency Outcomes

Descriptive Statistics. Table 2 presents pre- and post-test oral proficiency scores by group and dimension.

Table 2: Oral Proficiency Scores by Group (Mean $\pm$ SD)

Dimension	Experimental Pre	Experimental Post	Control Pre	Control Post
Fluency and coherence	2.84 $\pm$ 0.62	3.87 $\pm$ 0.71	2.79 $\pm$ 0.65	3.21 $\pm$ 0.68
Lexical resource	2.71 $\pm$ 0.58	3.62 $\pm$ 0.66	2.75 $\pm$ 0.61	3.08 $\pm$ 0.63
Grammatical accuracy	2.63 $\pm$ 0.55	3.41 $\pm$ 0.69	2.68 $\pm$ 0.57	2.94 $\pm$ 0.62
Pronunciation	2.78 $\pm$ 0.60	3.58 $\pm$ 0.64	2.74 $\pm$ 0.59	3.05 $\pm$ 0.61
Overall	2.74 $\pm$ 0.48	3.62 $\pm$ 0.58	2.74 $\pm$ 0.51	3.07 $\pm$ 0.54

Between-Group Comparisons. ANCOVA with pre-test scores as covariates revealed significant between-group differences on all post-test dimensions: fluency and coherence ($F = 24.31$, $p < .001$, partial $\eta^2 = .12$), lexical resource ($F = 18.76$, $p < .001$, partial $\eta^2 = .09$), grammatical accuracy ($F = 15.42$, $p < .001$, partial $\eta^2 = .08$), and pronunciation ($F = 20.87$, $p < .001$, partial $\eta^2 = .10$). The overall oral proficiency score showed a significant group effect ($F = 31.56$, $p < .001$, partial $\eta^2 = .15$).

Effect sizes for the experimental group compared to the control group were calculated using Cohen's d based on post-test scores adjusted for pre-test differences: fluency and coherence ($d = 0.94$), lexical resource ($d = 0.81$), grammatical accuracy ($d = 0.72$), pronunciation ($d = 0.84$), and overall

proficiency ($d = 0.98$). These effect sizes range from moderate to large, with the largest effects observed for fluency and overall proficiency.

Within-Group Changes. Paired-samples t-tests confirmed significant improvements within both groups from pre-test to post-test (all $p < .001$). However, the magnitude of improvement was substantially larger in the experimental group. For overall oral proficiency, the experimental group improved by 0.88 points ($SD = 0.51$), while the control group improved by 0.33 points ($SD = 0.47$), representing a difference in gain scores of 0.55 points ($t = 7.68$, $p < .001$, $d = 0.89$).

4.2 AI Scoring Validity

Inter-Rater Reliability. The two human raters demonstrated good agreement on the post-test oral assessments (ICC = 0.84, 95% CI [0.79, 0.88]). This established a benchmark for evaluating AI-human agreement.

AI-generated scores showed good agreement with the averaged human rater scores across all dimensions (Table 3).

Table 3: AI-Human Scoring Agreement

Dimension	ICC (95% CI)	Paired t-test	Pearson r
Fluency and coherence	.79 [.72, .85]	$t = 2.14$, $p = .034$	.81**
Lexical resource	.76 [.68, .82]	$t = 1.87$, $p = .063$	.78**
Grammatical accuracy	.73 [.64, .80]	$t = 2.53$, $p = .012$	.75**
Pronunciation	.81 [.75, .86]	$t = 1.92$, $p = .057$	.83**
Overall	.78 [.72, .83]	$t = 2.61$, $p = .010$	.80

** $p < .01$

The overall ICC of .78 falls within the "good" range according to Koo and Li's (2016) guidelines, indicating strong absolute agreement between AI and human raters. Dimension-level ICCs ranged from .73 (grammatical accuracy) to .81 (pronunciation), all within acceptable ranges.

Systematic Differences. Paired-samples t-tests revealed a small but statistically significant tendency for the AI system to assign higher scores than human raters on grammatical accuracy (mean difference = 0.18, $p = .012$) and overall proficiency (mean difference = 0.12, $p = .010$). The AI-human difference was not statistically significant for fluency ($p = .034$ with Bonferroni correction for multiple comparisons, not significant), lexical resource ($p = .063$), or pronunciation ($p = .057$). The magnitude of the differences was small (all $d < 0.20$), suggesting that the AI scoring bias was modest in practical terms.

4.3 Learning Motivation

Table 4: Learning Motivation Scores by Group (Mean $\pm$ SD)

Dimension	Experimental Pre	Experimental Post	Control Pre	Control Post
Interest-enjoyment	4.21 $\pm$ 1.03	5.48 $\pm$ 0.92	4.18 $\pm$ 1.06	4.47 $\pm$ 0.98
Perceived competence	3.87 $\pm$ 1.12	5.12 $\pm$ 0.95	3.91 $\pm$ 1.08	4.23 $\pm$ 1.02
Perceived choice	4.56 $\pm$ 0.94	5.63 $\pm$ 0.87	4.52 $\pm$ 0.97	4.78 $\pm$ 0.91
Perceived value	4.73 $\pm$ 0.98	5.71 $\pm$ 0.82	4.69 $\pm$ 1.01	4.94 $\pm$ 0.93
Overall	4.34 $\pm$ 0.89	5.49 $\pm$ 0.78	4.33 $\pm$ 0.92	4.61 $\pm$ 0.85

Descriptive Statistics. Table 4 presents pre- and post-test learning motivation scores by dimension.

Between-Group Comparisons. ANCOVA revealed significant group effects on all motivation dimensions: interest-enjoyment ($F = 28.42$, $p < .001$, partial $\eta^2 = .14$), perceived competence ($F = 22.67$, $p < .001$, partial $\eta^2 = .11$), perceived choice ($F = 19.35$, $p < .001$, partial $\eta^2 = .10$), perceived value ($F = 17.89$, $p < .001$, partial $\eta^2 = .09$), and overall motivation ($F = 35.21$, $p < .001$, partial $\eta^2 = .17$). Effect sizes for the experimental group were large: overall motivation ($d = 1.07$), interest-enjoyment ($d = 1.03$), perceived competence ($d = 0.89$), perceived choice ($d = 0.94$), and perceived value ($d = 0.86$).

Within-Group Changes. The experimental group showed substantial improvement in overall motivation from pre-test ($M = 4.34$, $SD = 0.89$) to post-test ($M = 5.49$, $SD = 0.78$), $t(92) = 12.43$, $p < .001$, $d = 1.38$. The control group showed a smaller but significant improvement (pre: $M = 4.33$, $SD = 0.92$; post: $M = 4.61$, $SD = 0.85$), $t(92) = 3.24$, $p = .002$, $d = 0.32$.

4.4 Course Satisfaction

The experimental group reported significantly higher course satisfaction ($M = 4.38$, $SD = 0.62$) compared to the control group ($M = 3.72$, $SD = 0.71$), $t(184) = 6.74$, $p < .001$, $d = 0.99$. Item-level analysis revealed that experimental group students particularly valued the AI practice opportunities ($M = 4.51$, $SD = 0.68$) and the immediate feedback provided by the system ($M = 4.42$, $SD = 0.71$). The lowest-rated item in the experimental group was “The AI system understood my spoken English well” ($M = 3.89$, $SD = 0.82$), suggesting some residual challenges with speech recognition accuracy.

4.5 Platform Engagement Data

Analysis of Chaoxing platform behavioral logs for the experimental group revealed the following engagement patterns over the 16-week semester:

The average number of AI practice sessions per student was 24.6 ($SD = 8.3$), with a mean total practice duration of 312 minutes ($SD = 87$ minutes) across the semester. Scenario dialogue completion rates averaged 82.4% ($SD = 12.7\%$), indicating high engagement with AI-mediated practice. Correlation analysis revealed that AI practice duration was significantly correlated with oral proficiency improvement ($r = .56$, $p < .001$), suggesting a dose-response relationship between AI practice and learning outcomes.

Engagement patterns varied across student subgroups. Students in the upper quartile of practice frequency ($n = 23$, ≥ 30 sessions) showed the largest oral proficiency gains (mean gain = 1.12 points), while those in the lower quartile ($n = 24$, ≤ 16 sessions) showed more modest gains (mean gain = 0.51 points). This variation suggests that the benefits of the TSA model are partially dependent on student engagement with AI practice opportunities.

4.6 Qualitative Findings

Teacher Perceptions. Thematic analysis of teacher interviews revealed three major themes:

Theme 1: Enhanced instructional precision. Teachers reported that AI Learning Analytics provided diagnostic information that was not available through traditional observation. One teacher stated: “I could see exactly which students were struggling with which specific expressions, and I could target my feedback accordingly. Before, I was guessing.”

Theme 2: Transformation of teacher role. Teachers described a shift from being the primary source of language input and feedback to a role focused on facilitation, design, and contextual interpretation. “The AI handles the repetitive practice and immediate correction. I can focus on the deeper conversations about service standards and professional judgment that the AI can’t provide,” one teacher explained.

Theme 3: Concerns about AI dependency. Some teachers expressed concern that students might become overly reliant on AI feedback and less willing to engage in spontaneous communication. “Some students want to practice with the AI before they speak in class. That’s good, but I worry they’re not developing the confidence to communicate without AI support.”

Student Perceptions. Student interviews revealed four themes:

Theme 1: Reduced speaking anxiety. Students consistently reported that practicing with the AI agent reduced their fear of making mistakes. “I can practice as many times as I want without feeling embarrassed. The AI doesn’t judge me,” said one student.

Theme 2: Appreciation of scenario-based practice. Students valued the professional relevance of the AI scenarios. “The dialogues are what I’ll actually face on the ship—checking documents, helping with luggage, explaining cabin features. It feels real.”

Theme 3: Recognition of feedback limitations. Some students noted that the AI feedback, while useful, was sometimes too generic. “It tells me I have a pronunciation problem, but it doesn’t always explain how to fix it. I still need the teacher for that.”

Theme 4: Positive impact on motivation. Students described the AI practice as engaging and motivating. “I wanted to keep practicing to improve my score. It’s like a game.”

5. Discussion

5.1 Interpretation of Main Findings

The findings provide empirical support for the effectiveness of the TSA collaborative model in vocational cruise service English instruction. The experimental group demonstrated significantly greater improvement in oral proficiency across all assessed dimensions, with effect sizes ranging from moderate to large. The overall effect size for oral proficiency ($d = 0.98$) is substantial and exceeds typical effects reported in

educational technology interventions (Hattie, 2009). These results are consistent with the positive outcomes reported in previous AI-enhanced vocational English studies, including Wang's (2025) maritime ESP study and the AI-enhanced listening and speaking study involving 472 vocational college students.

The particularly strong effect on fluency ($d = 0.94$) aligns with the theoretical rationale for the TSA model: scenario-based dialogue practice with the AI agent provides students with extensive opportunities for communicative practice in a low-stakes environment, allowing them to develop fluency through repetition without the anxiety associated with classroom performance. The high scenario dialogue completion rate (82.4%) and the positive correlation between practice duration and proficiency improvement ($r = .56$) support the interpretation that the AI Practice function served as the primary mechanism for skill development.

The findings on learning motivation ($d = 1.07$) and course satisfaction ($d = 0.99$) are particularly notable given the persistent challenge of low motivation in vocational ESP instruction. The qualitative data suggest that reduced speaking anxiety, professional relevance of scenarios, and gamified feedback contributed to the motivational benefits of the TSA model. These findings are consistent with Self-Determination Theory perspectives on AI-mediated language learning, which posit that AI tools enhance motivation by supporting learners' autonomy, competence, and relatedness.

5.2 AI Scoring Validity and Its Implications

The AI-human scoring agreement of $ICC = .78$ (overall) falls within the "good" range, supporting the acceptability of AI-assisted assessment for low-stakes formative purposes in vocational ESP. This level of agreement is comparable to human-human agreement reported in previous studies, including the $ICC = 0.85$ reported for two human raters in EFL oral assessment (Li, 2026), and exceeds the agreement reported for some AI systems in pronunciation assessment.

The observed tendency for the AI system to assign slightly higher scores than human raters (mean difference = 0.12 for overall proficiency) is consistent with findings from previous research on Gen-AI scoring bias. However, the magnitude of this bias was small ($d < 0.20$) and practically negligible for formative purposes. The bias was statistically significant only for grammatical accuracy and overall proficiency, suggesting that AI scoring may be more lenient in these dimensions. For high-stakes assessment, calibration against human benchmarks would be advisable.

The dimension-level variation in AI-human agreement—with pronunciation showing the highest $ICC (.81)$ and grammatical accuracy the lowest ($.73$)—provides useful information for understanding the strengths and limitations of AI scoring. The AI system appears most reliable for assessing pronunciation, where acoustic features can be objectively measured, and somewhat less reliable for grammatical accuracy, where contextual and pragmatic factors may influence human judgment in ways that the AI system does not fully capture.

5.3 Theoretical Implications

The findings contribute to symbiotic theory by providing empirical evidence that a TSA collaborative model can produce measurable improvements in language learning outcomes. The model's effectiveness can be understood through the symbiotic principles articulated in the theoretical framework:

Complementarity was realized through the division of labor between human and AI capabilities. The AI agent handled repetitive practice, immediate feedback, and data collection—tasks that are time-consuming for teachers and difficult to standardize. The teacher focused on contextual interpretation, professional judgment, and motivational support—tasks that require human expertise and relational engagement.

Interdependence was evidenced by the correlation between AI practice engagement and learning outcomes, and by the qualitative finding that students valued both AI practice and teacher guidance. Neither element alone would have produced the observed outcomes; the synergy between them was essential.

Emergence was demonstrated by the overall effect on oral proficiency ($d = 0.98$), which exceeds what would be expected from either AI practice or traditional instruction alone. The collaborative system generated pedagogical value that neither component could produce independently.

5.4 Practical Implications

For Teachers. The TSA model provides a practical framework for integrating AI agents into vocational ESP instruction. The four-stage structure offers clear guidance for implementation, and the role division framework clarifies the teacher's transformed but essential role as learning architect, contextual interpreter, and ethical guide. Teachers considering implementation should invest in understanding the AI system's capabilities and limitations, particularly regarding scoring bias and the need for calibration against human judgment.

For Institutions. The findings support institutional investment in AI-enhanced teaching platforms, with evidence that the Chaoxing AI Agent's four-function configuration can produce measurable learning gains when implemented within a coherent pedagogical model. Institutions should provide training for teachers on AI platform operation and pedagogical integration, and should develop policies for the appropriate use of AI-generated assessment data.

For Platform Developers. The study identifies areas for platform improvement, including more granular and actionable feedback on pronunciation and grammar, and better integration of AI scoring with human judgment for high-stakes assessment. The observed student difficulty with AI understanding spoken English (lowest satisfaction item) suggests that speech recognition accuracy in noisy classroom environments remains a technical challenge requiring

ongoing development.

5.5 Limitations and Future Research

Several limitations should be acknowledged:

First, the quasi-experimental design with non-random assignment limits causal inference. While baseline equivalence was established, unmeasured differences between groups may have influenced outcomes.

Second, the single-institution, single-course context limits generalizability. Replication across different vocational colleges, ESP domains, and student populations is needed.

Third, the 16-week intervention period allows assessment of short-term outcomes but does not capture long-term retention or transfer to workplace contexts.

Fourth, the AI scoring validity findings are specific to the Chaoxing AI Practice system and may not generalize to other automated scoring systems.

Fifth, the study relied on self-reported motivation measures, which may be susceptible to social desirability bias.

Future research should address these limitations through multi-institutional collaborations, longitudinal designs, and mixed-method approaches that include workplace performance data. Research should also examine the cost-effectiveness of the TSA model, including the time investment required for teacher training and AI agent configuration, and should explore the ethical implications of AI-mediated assessment in vocational education.

6. Conclusion

This study provides empirical evidence for the effectiveness of a Teacher-Student-AI Agent collaborative teaching model in vocational cruise service English instruction. The findings demonstrate that the model, implemented through the Chaoxing AI Agent platform, produced significant improvements in students' oral English proficiency ($d = 0.98$), learning motivation ($d = 1.07$), and course satisfaction ($d = 0.99$) compared to traditional instruction. The AI scoring system demonstrated good agreement with human raters ($ICC = .78$), supporting the acceptability of AI-assisted assessment for formative purposes, though a slight leniency bias was observed that warrants calibration for high-stakes applications.

The study contributes to the growing body of evidence on AI-empowered vocational language education by demonstrating that platform-integrated AI agents can be effectively operationalized within a collaborative pedagogical model. The triangulated evidence chain—combining platform behavioral logs, AI-generated scores, and human ratings—provides a methodological template for future research on human-AI collaboration in language education. The findings offer practical guidance for teachers, institutions, and platform developers seeking to implement AI-enhanced instruction in vocational ESP contexts, while acknowledging the need for further research on long-term outcomes, transferability, and ethical considerations.

Acknowledgments

This research is a phased achievement of the 2026 National Joint Conference of Presidents of Higher Vocational Colleges project “Research on the Teacher-Student-Machine Collaborative Teaching Model for Higher Vocational Industry English Based on AI Agents” (Project No. XLH202632).

References

- [1] Clarke, V., & Braun, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- [2] Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.
- [3] Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- [4] Li, F. (2026). Innovation and practice of database teaching in higher vocational education under the Teacher-AI-Student tripartite synergistic framework. In *Proceedings of the 2026 5th International Conference on Humanities, Wisdom Education and Service Management* (pp. 1–8). Atlantis Press. https://doi.org/10.2991/978-2-38476-593-5_26
- [5] Pratama, M. S. K., Rahman, A., Widiati, B., & Montecoro, C. B. (2025). Artificial intelligence and teacher collaboration in enhancing English fluency: Evidence from EFL vocational schools learners. *JOLLS: Journal of Language and Literature Studies*, 5(2). <https://doi.org/10.36312/jolls.v5i2.2820>
- [6] Wang, J. (2025). Large-language-model-supported BOPPPS teaching for maritime ESP: A quasi-experimental study in a Chinese vocational college. In *Proceedings of the 2025 3rd International Conference on Information Education and Artificial Intelligence* (pp. 1–7). ACM. <https://doi.org/10.1145/3799457.3799635>
- [7] Wang, X., & Ganapathy, M. (2026). AI-assisted speaking instruction and English as a foreign language learners' oral accuracy and fluency: A quasi-experimental study. *International Journal of Evaluation and Research in Education*, 15(3). <https://doi.org/10.11591/ijere.v15i3.30212>