

# Research on Generative AI-Empowered Formative Assessment of Maritime English: Evaluation Indicators and Practical Paths

Xiang Huang

Jiangsu Maritime Institute, Nanjing, Jiangsu, 211170, China

---

**Abstract:** The Manila Amendments to the STCW Convention set compliance requirements for seafarers' maritime communicative competence; however, higher-vocational Maritime English assessment still relies mainly on summative final-term evaluation, with high oral-assessment costs and delayed feedback. Drawing on the Outcomes-Based Education (OBE) philosophy and formative assessment theory, and supported by generative AI and intelligent speech-judgment technology, this study constructs a three-dimensional evaluation indicator system and designs a human-AI collaborative assessment closed loop of "pre-class diagnosis, in-class dynamic monitoring, and post-class iterative feedback". A 16-week quasi-experiment shows that this model significantly improves students' maritime English communicative competence, with the largest gain in the oral dimension (18.7% improvement in the experimental class versus 5.3% in the control class), and enhances students' learning initiative. The study also analyzes the validity defects and academic-integrity risks of AI-driven assessment and proposes human-AI collaborative verification mechanisms, providing practical references for curriculum assessment reform of higher-vocational Maritime English and for STCW-aligned talent training.

**Keywords:** Generative AI; Maritime English; Formative Assessment; STCW Convention; ESP Teaching.

---

## 1. Introduction

The Manila Amendments to the STCW Convention emphasize that the evaluation of seafarers' English competence should focus on communicative performance in real-ship operational scenarios rather than on the mere testing of linguistic knowledge points. However, Maritime English courses in Chinese higher vocational colleges have long been plagued by an assessment practice that "values final examinations over learning processes". Maritime oral communication assessment depends on manual scoring by teachers, which yields limited samples and subjective judgments, making normalized, high-frequency scenario-based assessment difficult to achieve. Feedback is often delayed, so teachers cannot precisely identify students' expressive weaknesses in scenarios such as VHF communication, PSC inspections, and emergency calls, and "promoting learning through assessment" can hardly be realized.

In recent years, generative AI has been rapidly adopted in foreign-language education. Combined with maritime speech-recognition technology, large language models can automatically collect multimodal learning evidence, including students' oral audio and maritime English messages, thereby enabling automatic collection and instant diagnosis of process-oriented data. However, most existing studies focus on AI-enabled assessment for general foreign languages; for Maritime English ESP courses, research on indicator-system construction and classroom-based empirical validation of generative-AI-empowered formative assessment remains scarce. Accordingly, this paper addresses three research questions: (1) What evaluation indicator system should be constructed for generative-AI-empowered formative assessment of Maritime English? (2) How can whole-process assessment practice paths and supporting mechanisms be designed? (3) When applied in real classrooms, can this

assessment mode improve students' maritime-scenario English communicative competence and learning initiative?

## 2. Core Concepts and Literature Review

### 2.1. Formative-Assessment Theory

Black and Wiliam argue that formative assessment is continuous evaluation embedded throughout the learning process. Its core purpose is to diagnose learning deficiencies and deliver improvement-oriented feedback. Unlike summative assessment for sorting and grading, it is essentially assessment-for-learning[1]. Yuan Shuhou et al. note that the Maritime-English evaluation system should cover listening-and-speaking, maritime writing, professional quality and other dimensions, yet conventional pedagogical approaches struggle to sustain the collection of holistic process-oriented learning evidence[2]. Xu Ping proposes that formative assessment for sea-related specialty English should integrate job-competence requirements, adopt multi-agent evaluation, and establish process-learning portfolios[3]. Wu Xuehong explores reform directions for formative assessment in higher-vocational Maritime-English teaching, providing direct references for this research[4].

### 2.2. Generative-AI-Empowered Foreign-Language Assessment

Lu Guoqing et al. hold that generative AI is suitable for learning-data collection, preliminary scoring of multimodal learning outputs and personalized diagnostic feedback. It cannot replace teachers' professional judgement; hence a human-AI collaborative-assessment paradigm should be established[5]. Within ESP education, Chen Xinren discusses the application potential of large-language models in second-language pragmatics instruction[6]. A comparative study by Ren Wei et al. confirms that AI can rapidly detect

linguistic errors in written texts, but it suffers limitations in judging industry-specific pragmatics and job-context communicative appropriateness. Teachers must calibrate evaluation criteria[7].

Within maritime education, intelligent evaluation systems for seafarers' English listening-and-speaking have been developed domestically. Speech-based models deliver preliminary automatic scoring for VHF conversations and emergency dialogues. These systems are mostly used for summative competency examinations and have not been fully implemented for routine formative assessment in courses[8].

### **2.3. STCW-Aligned Evaluation Requirements for Maritime-English ESP**

IMO Model Course 3.17 *Maritime English* clarifies that the core of Maritime English assessment lies in the completion of maritime-scenario communicative tasks. Assessment tasks should reproduce real shipboard working contexts such as navigational notices, distress communications, and port-state inspections, and evaluation indicators need to balance linguistic accuracy, the use of Standard Marine Communication Phrases (SMCP), and communicative-task achievement, rather than merely testing grammatical correctness 1).

1) IMO. Model Course 3.17 Maritime English[M]. London: International Maritime Organization, 2015.

### **2.4. Literature Commentary**

Previous research has laid a theoretical foundation for this study, but three gaps remain. First, findings on AI-enabled assessment for general foreign languages cannot be directly transferred to highly industry-contextualized Maritime English courses. Second, existing Maritime English evaluation indicators are mostly designed for summative competency examinations, and indicator frameworks adapted to multimodal process-oriented data that integrate both linguistic competence and maritime professional competence are scarce. Third, classroom-based empirical tests of generative-AI-empowered formative assessment in Maritime English are lacking. These gaps motivate the present study: Research Questions 1 and 2 target indicator-system construction and practical-path design, while Research Question 3 examines the effectiveness of the proposed model through a quasi-experiment.

## **3. Value and Practical Dilemmas of Generative-AI-Empowered Maritime-English Formative Assessment**

### **3.1. Value of Empowerment**

(1) Full-cycle multimodal learning-data collection. Generative AI can automatically record learning evidence, including pre-class preview outputs, in-class audio of maritime scenario dialogues, post-class maritime journals, and simulated VHF tasks, thereby overcoming the small samples and one-off final examinations characteristic of traditional assessment.

(2) Reduced oral-assessment costs and regular scenario-based assessment. Manual oral assessment of Maritime English imposes heavy workloads on teachers, making high-frequency scenario-based assessment for all students infeasible. AI speech models can preliminarily score students'

submitted maritime-dialogue audio and provide instant feedback at any time, enabling regular scenario-based oral training and assessment.

(3) Tiered personalized diagnosis. AI can distinguish among general grammatical errors, misuse of maritime terminology, and deficiencies in scenario-based communication strategies, and can generate personalized improvement suggestions to support differentiated instruction.

(4) Support for multi-agent collaborative evaluation. AI-generated diagnostic reports serve as references for teacher assessment, student self-assessment, and peer assessment, constructing a multi-agent assessment closed loop that is consistent with the evaluation orientation of the higher-vocational "three-education reform" and industry-education integration.

### **3.2. Practical Dilemmas**

(1) Limitations in judging maritime pragmatics. Large language models achieve high accuracy in identifying general-English grammar, yet they tend to misjudge Standard Marine Communication Phrases (SMCP) and the communicative appropriateness of utterances in shipboard contexts. Sole reliance on automatic AI scoring would impair assessment validity.

(2) Academic-integrity risks. Students may use generative AI to produce maritime dialogues and English messages directly, which can distort assessment results and fail to reflect students' actual Maritime English proficiency.

(3) Lack of AI-compatible formative-assessment indicators. Current Maritime English evaluation indicators are oriented toward summative competency examinations for seafarers, and indicator frameworks compatible with multimodal process-oriented data remain insufficient.

(4) Insufficient digital literacy among teachers. Maritime English teachers generally lack competence in prompt engineering, calibration of AI-assessment outputs, and risk control.

## **4. Construction of the Generative-AI-Empowered Maritime-English Formative-Assessment Indicator System**

Following STCW compliance requirements and the Outcomes-Based Education (OBE) philosophy, and adhering to the principles of "human-AI division of labor, job orientation, and process measurability", this paper constructs a three-dimensional indicator system that distinguishes quantitative indicators for automatic AI scoring from qualitative indicators for manual teacher evaluation (see Table 1).

## **5. Practical Paths for Generative-AI-Empowered Maritime-English Formative Assessment**

A three-phase closed loop of "pre-class diagnosis, in-class dynamic monitoring, and post-class iterative feedback" is established throughout the Maritime English curriculum, supported by AI-assessment verification and academic-integrity-control mechanisms.

**Table 1.** Generative-AI-Empowered Maritime-English Formative-Assessment Indicator System

| Primary Indicator (Weight)                                                    | Secondary Indicator                             | Tertiary Observation Point                                                                                                                                         | Evaluator                                                                       |
|-------------------------------------------------------------------------------|-------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------|
| Dimension 1: Foundational maritime-English linguistic competence (25%)        | Mastery of maritime vocabulary and terminology  | Recognition of maritime-specialty vocabulary; accuracy rate of SMCP phrases; spelling of specialized terminology                                                   | AI-led, teacher review                                                          |
|                                                                               | Maritime grammar and sentence patterns          | Standardization of tenses and sentence structures in maritime messages and oral utterances; statistics of grammatical-error types                                  | AI-led, teacher review                                                          |
| Dimension 2: Maritime-scenario communicative competence (55%, core dimension) | Maritime listening comprehension                | Accuracy of extracting key maritime information from VHF communications, navigational notices, and shipboard instructions                                          | AI speech recognition + teacher calibration                                     |
|                                                                               | Scenario-based oral communication               | Fluency, appropriateness of professional expressions, and achievement of task objectives in simulated emergency, berthing/unberthing, and PSC-inspection dialogues | AI preliminary scoring + teacher qualitative scoring                            |
|                                                                               | Maritime written communication                  | Text logic, stylistic standardization, and information completeness of maritime reports, maritime journals, and communication messages                             | AI preliminary scoring; teachers focus on the soundness of professional content |
| Dimension 3: Professional quality and sustainable-learning competence (20%)   | Learning engagement                             | Preview completion rate; frequency of task submission; number of revisions based on AI feedback; learning journals                                                 | Automatic AI collection of process data                                         |
|                                                                               | Maritime cross-cultural communication awareness | Foreign-related communication etiquette; maritime-safety discourse awareness; risk-expression strategies                                                           | Teacher evaluation + peer assessment; AI assists in evidence collection         |
|                                                                               | Self-reflection and metacognitive competence    | Student self-assessment reports; improvement plans based on AI diagnostic reports                                                                                  | Teacher evaluation; AI assists text analysis                                    |

Note on indicator weights: maritime-scenario communicative competence is the core dimension (weight: 55%). AI undertakes the collection and preliminary scoring of objective quantitative indicators; for qualitative items such as the appropriateness of maritime-scenario communication and professional quality, teachers make the final judgment and AI serves only as an auxiliary reference, to mitigate assessment bias caused by AI misjudgment of maritime-professional contexts.

### 5.1. Pre-class: AI-Based Diagnostic Assessment for Precise Learning-Situation Analysis

Lightweight preview tasks are designed with generative AI before class, including maritime-vocabulary quizzes, listening comprehension of short maritime audio materials, and short maritime-sentence translation. AI auto-grades the assignments and generates class-level learning-situation dashboards to identify common class-wide weaknesses (e.g., distress-communication terminology and meteorological English expressions) as well as individual student deficiencies. Based on the AI diagnostic reports, teachers adjust scenario-based teaching tasks and implement differentiated instruction, realizing “assessment prior to instruction”.

### 5.2. In-class: Real-Time AI Collection of Scenario-Task Data for Human-AI Collaborative Dynamic Assessment

Classroom instruction adopts authentic maritime job-oriented tasks, such as simulated VHF calls, English PSC-inspection dialogues, and oral ship-incident reports, as ESP communicative activities. Students record their oral task performances; AI speech models conduct real-time transcription, extract linguistic indicators, and output preliminary quantitative scores. Teachers observe students’ on-site communicative performance and mainly evaluate whether the maritime tasks are fulfilled and whether professional expressions conform to maritime norms. AI stores all audio and textual records in course-learning portfolios. Student self-assessment and group peer assessment are conducted in class, and evaluation data from all agents are incorporated into the learning portfolios.

### 5.3. Post-class: AI-Generated Diagnostic Reports to Drive Iterative Improvement

After class, students complete written assignments such as maritime journals and brief maritime-incident reports. Generative AI rapidly provides feedback, marks terminology and sentence-pattern errors, and offers revision suggestions. Students revise their drafts against the AI diagnostics and submit iterative versions. Teachers check AI feedback for professional misjudgments and add qualitative comments on maritime-scenario pragmatics. AI preserves multiple revised drafts, traces students’ competence-growth trajectories, and forms visualized learning-growth portfolios as the basis for formative course grades.

### 5.4. Supporting Mechanisms

(1) Human-AI review mechanism. A manual sampling-review system for AI scoring is established: teachers review no less than 30% of samples in each task batch to correct AI misjudgments in maritime-professional contexts and to continuously optimize AI prompts and scoring rubrics.

(2) Task traceability and anti-cheating mechanism. Scenario-based tasks are assigned with time limits and segmented real-time recording requirements to prevent students from using AI to generate complete answers directly. Clear boundaries for AI-tool use in courses are defined to foster academic-integrity awareness.

(3) Improvement of teachers’ digital competence. Prompt engineering for AI assessment and interpretation of AI-assessment outputs are incorporated into training programs for Maritime English teachers under the “three-education reform”, enhancing teachers’ capability to deploy generative-AI assessment tools.

## 6. Empirical Study

### 6.1. Research Design

A nonequivalent pretest-posttest quasi-experimental design was adopted. Participants were two parallel classes (45 students per class, 90 in total) of the 2024 cohort of the Maritime Technology program at a higher vocational college. The experimental class applied the human-AI collaborative formative-assessment model proposed in this study, while the control class received conventional assessment (a final-term written examination plus a one-off final oral interview, supplemented by a small number of after-class written assignments). To control for teacher effects, both classes were taught by the same instructor using identical textbooks and teaching schedules. The experiment lasted 16 weeks, covering a complete semester of Maritime English instruction.

The research instruments included: (1) a pretest and posttest using parallel forms of a standardized competency-examination paper for seafarers' Maritime English, covering listening comprehension, maritime reading, scenario-based oral communication, and maritime writing (Cronbach's  $\alpha = 0.842$ ); (2) a post-course student-satisfaction questionnaire (87 valid responses; response rate 96.67%); and (3) semi-structured interviews with eight randomly selected students from each class.

### 6.2. Experimental Results

(1) Pretest-posttest scores. An independent-samples t-test showed no statistically significant difference in total scores between the two classes at pretest ( $P > 0.05$ ), indicating equivalent baseline proficiency. After the 16-week intervention, the posttest total score of the experimental class was significantly higher than that of the control class ( $t = 2.376$ ,  $df = 88$ ,  $P < 0.05$ , Cohen's  $d = 0.502$ ). The largest gain occurred in scenario-based oral communication: the mean oral-score improvement reached 18.7% in the experimental class versus 5.3% in the control class. A paired-samples t-test revealed that the posttest scores of the experimental class were significantly higher than its pretest scores ( $t = 4.125$ ,  $df = 44$ ,  $P < 0.001$ ), whereas no significant pre-post difference emerged for the control class ( $t = 1.631$ ,  $df = 44$ ,  $P > 0.05$ ).

(2) Questionnaire results. In the experimental class, 86.7% of the students reported that instant AI feedback helped them identify weaknesses in maritime-English expression; 82.2% approved of the "multi-round revision" formative tasks and believed that these tasks helped them improve job-oriented English communicative competence; and 77.8% stated that the assessment model enhanced their initiative in learning Maritime English.

(3) Interview findings. Students generally reported that one-off final oral examinations imposed heavy pressure and failed to expose long-standing deficiencies in terminology and scenario-based expression. Regular AI-supported scenario-based assessment lowered the psychological pressure of oral practice. Meanwhile, students noted that AI may occasionally misjudge maritime-professional utterances, making manual teacher review indispensable.

### 6.3. Empirical Discussion

The results demonstrate that generative-AI-empowered human-AI collaborative formative assessment can effectively resolve the pain points of conventional Maritime English oral assessment, such as low assessment frequency and delayed feedback, and can significantly improve students' maritime-

scenario English communicative competence. This finding is consistent with the theoretical expectation of formative assessment theory that assessment should promote learning, and it validates the rationale for designating maritime-scenario communicative competence as the core dimension (weight 55%). The experiment also shows that AI cannot replace teachers' qualitative judgment of maritime pragmatics and task accomplishment in job-oriented communication; human-AI collaboration constitutes the core of this assessment model. In addition, some students tended to over-rely on AI for ready-made outputs, indicating that anti-cheating mechanisms should be embedded in task design and assessment rules to constrain such behaviors.

## 7. Conclusion

Generative AI creates new opportunities for higher-vocational Maritime English to break through conventional assessment limitations and build a whole-process formative-assessment system. As an ESP course oriented toward seafarer job requirements, Maritime English assessment reform cannot simply copy AI-assessment solutions developed for general foreign languages. It must be anchored in the maritime-communicative-competence requirements stipulated by the STCW Convention and adhere to the principle of "human-AI collaboration, AI as auxiliary support, and teacher-led decision-making". This study constructs a three-dimensional, tiered indicator system and a "pre-class, in-class, post-class" closed-loop practical framework, and verifies its effectiveness through a quasi-experiment, offering operable solutions to long-standing problems in Maritime English formative assessment, including the difficulty of collecting process-oriented evidence, high oral-assessment costs, and delayed feedback.

This study has two limitations. First, the empirical data were collected from only two classes, which limits generalizability; future research should expand the sample across multiple maritime vocational colleges to further test the reliability and validity of the indicator system. Second, frontline industry opinions have not yet been incorporated into the assessment criteria; future revisions should invite senior seafarers from shipping enterprises to participate in indicator refinement so that the assessment standards better match real-world demands aboard ocean-going vessels.

## References

- [1] Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5–31.
- [2] Yuan, S. H., & Shu, D. F. (2017). Review and reflection on formative assessment research in foreign-language teaching in China (2002–2016). *Foreign Language Teaching and Practice*, (4), 51–56, 21. (in Chinese)
- [3] Xu, P. (2023). Design and practice of formative-assessment model for sea-related specialty English practical courses. *Journal of Jimei University (Education Science Edition)*, 24(3), 72–78. (in Chinese)
- [4] Wu, X. H. (2018). Current situation and countermeasures of Maritime-English teaching research in China. *Journal of Maritime Education Research*, 35(1), 69–75. (in Chinese)
- [5] Lu, G. Q., Yang, Q., & He, X. M. (2024). Value, challenges and paths of generative-AI empowering formative assessment in higher education. *E-education Research*, 45(11), 84–91. <https://doi.org/10.13811/j.cnki.eer.2024.11.012> (in Chinese)

- [6] Chen, X. R. (2024). ChatGPT and L2 pragmatics teaching. *Foreign Language Teaching and Practice*, (4), 16–29. (in Chinese)
- [7] Ren, W., Liu, Y. B., & Xie, Y. (2024). A comparative study of ChatGPT feedback versus teacher feedback in EFL writing teaching. *Foreign Language Teaching and Practice*, (4), 30–38, 60. (in Chinese)
- [8] Xue, H. L., & Qiao, H. Y. (2019). Curriculum reconstruction of marine-engineering English based on the localization of IMO Model Course. *Journal of Maritime Education Research*, 36(4), 106–109. (in Chinese)