30th
International Maritime Lecturers Association Conference
Split Croatia
www.imla30.org

International Maritime Lecturers Association
FACULTY OF MARITIME STUDIES

AUTOMATED ASSESSMENT OF MARITIME SIMULATOR TRAINING: A DATA-DRIVEN APPROACH USING LARGE LANGUAGE MODELS

Arbresh Ujkani*, Paul Koch

Abstract. This paper presents the technical development and implementation of an automated assessment system within the i-Master EU project for maritime simulator training. Automating performance evaluation in maritime training presents challenges due to the diverse nature of training tasks and the need for flexible assessment methods. To address this, the paper outlines architecture designed to support a wide range of basic tasks by enabling flexible task definition and automatic assignment of relevant metrics and evaluation criteria. This adaptability ensures that assessments remain consistent while allowing customization to fit different training scenarios, instructor styles and student needs.

The system performs evaluation by collecting structured, self-describing metrics that represent individual skills during an exercise. These metrics are aggregated at the end of a training session and assessed against predefined evaluation criteria before being processed by a Large Language Model (LLM). The LLM generates tailored, natural language feedback that mirrors instructor-style debriefings, providing personalized insights into student performance. All recorded metrics, assessment criteria, and generated evaluations are stored in a database with version control, ensuring retrospective re-evaluation as assessment models improve. By providing the history of assessments the system allows for iterative refinement of both evaluation parameters and LLM prompting strategies, ensuring adaptability to evolving maritime training requirements.

The technical implementation provides a foundation for future validation studies with students and instructors at maritime universities. This work presents a novel technical approach to automated maritime assessment, demonstrating the feasibility of LLM integration in maritime training systems.

Keywords: Simulator Training; Assessment; Large Language Model; Simulator Metrics.

Fraunhofer Center for Maritime Logistics and Services, Sea Traffic & Nautical Solutions, Blohmstraße 32 21079 Hamburg, Germany

*** Corresponding author:** arbresh.ujkani@cml.fraunhofer.de

https://doi.org/10.7225/imla30.07

1. INTRODUCTION

1.1. Background and Motivation

Maritime education faces a fundamental scalability challenge in simulator-based training assessment. Traditional instructor-led evaluation creates operational bottlenecks as training demand increases while qualified instructor availability remains limited. Simultaneously, the subjective nature of human assessment introduces variability that conflicts with International Maritime Organization (IMO) Standards of Training, Certification and Watchkeeping (STCW) (International Maritime Organization, 2025) requirements for consistent competency evaluation.

Current maritime simulators generate extensive performance data but lack the analytical capabilities to transform this data into pedagogically valuable feedback. Students receive raw metrics without the interpretive guidance that instructors typically provide, limiting opportunities for self-directed learning and continuous improvement.

1.2. Problem Statement

The maritime training sector faces several critical challenges:

(1) the limited availability of personalized and adaptive training systems,

(2) inconsistent assessment quality due to subjective instructor evaluation,

(3) inability to provide immediate, detailed feedback to students, and

(4) lack of systematic performance tracking across multiple training sessions.

These challenges are particularly evident in the context of STCW requirements, which demand consistent and comprehensive competency assessment.

1.3. Research Objectives

We address these challenges by presenting an Intelligent Learning System (ILS) framework that:

(1) automates collection and analysis of maritime simulator data

(2) provides real-time performance assessment using STCW competency frameworks

(3) generates natural language feedback using Large Language Models

(4) enables systematic performance tracking across training sessions

2. TECHNICAL CONTRIBUTION

This work presents the first integration of Large Language Models with maritime simulator assessment systems, demonstrating the technical feasibility of automated feedback generation in safety-critical training environments.

Primary contributions include:

- **Automated Assesment Architecture:** A modular, portable platform that integrates with existing maritime simulators through standardized protocols (IEEE 1278 DIS, NMEA AIS), enabling institutions to enhance assessment capabilities without infrastructure replacement.
- **LLM-Based Feedback Generation:** Natural language feedback that transforms quantitative simulator metrics into instructor-style debriefings, maintaining technical accuracy while providing personalized learning guidance.
- **STCW Aligned Metrics Frameworks:** Real-time evaluation against maritime competency standards, creating objective performance measurement that complements traditional instructor assessment.

2.1. Scope and Limitations

The current implementation focuses on fundamental navigation competencies in controlled simulator environments. Complex decision-making scenarios requiring maritime law interpretation or emergency coordination that fall outside the immediate scope. This work establishes the technical foundation for future validation studies planned within the i-Master EU project.

3. RELATED WORK

3.1. Maritime Training Assessment Landscape

Current maritime training assessment relies on manual instructor observation during simulator exercises, involving live observation, video review, and structured debriefing sessions. Instructors take handwritten notes during multi-student sessions, then provide oral feedback covering competency areas such as collision regulation compliance, navigation techniques, and team coordination. While this approach allows flexible response to unexpected situations, research has identified reliability issues with conventional instructor evaluation. Researchers found that traditional instructor assessment of pilotage operations showed unacceptably low reliability for technical competencies, while computer-assisted assessment demonstrated improved reliability (Jørgen Ernstsen, 2020).

Existing commercial maritime simulator systems provide basic performance logging such as speed-over-ground graphs and positional tracking but lack contextual analysis connecting metrics to specific STCW competency areas. Students receive raw data without interpretive feedback that contextualizes performance within collision regulations and operational procedures. The STCW framework requires comprehensive competency evaluation, but the manual nature of instructor assessment creates operational bottlenecks: limited instructor availability for independent practice sessions, assessment inconsistencies across evaluators, and inability to provide immediate detailed feedback. Maritime training institutions lack systematic performance tracking databases for long-term learning progress monitoring. No commercial systems currently integrate automated feedback generation for consistent, STCW-aligned assessment during independent practice sessions, representing the gap this work addresses.

3.2. AI Integration in Maritime Education

Recent advances in automated assessment systems have demonstrated significant potential for maritime training applications. These systems leverage objective performance data logged during simulation exercises, including positional data, steering commands, and rule violations, to calculate metrics based on pre-modeled criteria with the goal of providing reproducible, consistent feedback without human interpretation bias.

(Øvergård, Nazir, & Solberg, 2017) presented a prototype system that calculated performance indices from defined navigational requirements, achieving strong correlation with expert evaluations. The authors viewed this as a promising step toward "consistent, unbiased" assessment tools. Similarly, (Jørgen Ernstsen, 2020) developed a Computer-Aided Performance Assessment (CAPA) system combining hierarchical performance indicators using Analytical Hierarchy Process (AHP) with Bayesian Networks. Their experimental comparison with 16 pilot trainers demonstrated that automated assessment achieved good reliability metrics for technical competencies, while human assessment proved "unacceptably" unreliable in these areas. However, neither automated nor human assessment achieved reliable ratings for teamwork competencies.

Learning analytics and artificial intelligence approaches represent emerging developments in this field. (Munim & Kim, 2023) investigated learning analytics dashboards for simulator training, emphasizing that actionable key performance indicators (KPIs) can be derived from training log data and visualized to provide valuable insights for both learners and instructors. Advanced machine learning techniques enable pattern recognition in control data for automated feedback generation.

The integration of LLMs represents a particularly novel development in maritime training assessment. While traditional automated systems provide quantitative metrics and basic performance logging, the application of LLMs to generate natural language feedback in maritime training contexts remains largely unexplored. This creates opportunities for AI-enhanced assessment tools that can provide consistent, detailed feedback while reducing instructor workload and addressing the scalability challenges identified in conventional assessment methods.

3.3. Cross-Domain Applications of LLMs in Simulator Training

The application of LLMs in simulator-based training has seen rapid growth across domains such as aviation, medicine, military operations, and industrial training. These developments highlight the potential of LLMs to provide automated assessment, personalized feedback, and scenario generation—capabilities that closely align with the objectives of maritime training.

In aviation, LLMs are integrated into flight simulators for real-time advisory and automated communication assessment. The LeRAAT framework connects an LLM to the X-Plane flight simulator, providing pilots with context-aware decision support during abnormal or emergency procedures (Schlichting, et al., 2025). Similarly, LLM-based air traffic control (ATC) agents have been developed to simulate realistic ATC dialogue and provide immediate feedback on communication protocols (Andriuškevicius & Junzi, 2024).

In medical education, LLMs are used for interactive patient simulations and post-session debriefing. Systems like MedSimAI employ LLM-driven virtual patients to evaluate clinical reasoning and communication, offering structured feedback directly after simulated consultations (Cook, et al., 2025). Controlled studies have shown significant improvements in diagnostic reasoning when medical students receive structured, LLM-generated feedback following virtual patient sessions (Brügge, et al., 2024).

In military training, LLMs have been applied to wargaming environments to automate scenario planning and outcome analysis. For example, the AFSIM-based wargame framework leverages LLM agents to create dynamic adversary behavior and generate narrative reports, reducing scenario design times from weeks to days (Raj, 2025).

Industrial training has adopted LLMs as conversational assistants in XR/VR maintenance simulators. These systems combine immersive 3D training environments with LLM-driven instructions that respond in real time to trainee actions, referencing equipment manuals and safety protocols (Tomkou, et al., 2025).

4. SYSTEM ARCHITECTURE AND DESIGN

The architecture of the proposed ILS for maritime simulation as shown in Figure 1, is designed as a modular pipeline that supports real-time assessment, adaptive feedback, and personalized learning. Student assessment begins with task delivery, where students receive a simulation-based task defined in a structured training scenario module. These scenarios specify operational parameters and performance metrics that form the basis for evaluation. Once the student performs the task using the maritime simulator, data is continuously collected by a telemetry collector. Telemetry collection focuses on capturing simulation data, such as speed, course as well as sensor and behavioral data from the student if configured. It is to be noted that initial iterations of the ILS concentrate on simulator telemetry alone, excluding additional sensors such as gaze or heart rate to reduce integration-complexity. These data streams are then processed by collector modules, responsible for extracting domain-specific indicators like cross-track deviation, closest point of approach and Convention on the International Regulations for Preventing Collisions at Sea (COLREGs) violations. The extracted data is passed to the metric evaluation module, which interprets the data in accordance with the task-defined metrics. Evaluation outputs are forwarded and prepared for grading, which calculates a weighted performance score based on task criteria. This assessment is then stored and visualized through a learning analytics dashboard, which serves both students and instructors by providing transparent progress tracking. To tailor the feedback in a more natural manner, a comprehensive textual output is provided to the student via a LLM and a modular prompting engine, allowing the system to formulate metric based results in textual descriptions of student performance and useful recommendations in additional lectures or courses to be taken by the student. To support continual improvement and personalization, an adaptive learning module dynamically adjusts task parameters and grading weight using trend analysis and historical performance data. All telemetry, assessment outcomes and learning adjustments are stored, enabling long-term tracking and providing insights at the level of the curriculum, across a student's learning history and a group of students.

Grading results are primarily used to provide user feedback in two ways:

(1) a metric-based display in the learning analytics dashboard, and

(2) full-text feedback generated by the LLM module.

This feedback uses rich context to generate texts suited to different applications, mimicking spoken or written instructor feedback. This written feedback is also displayed in the learning analytics dashboard alongside the metrics, providing students and instructors with comprehensive insight into student performance during simulator runs.

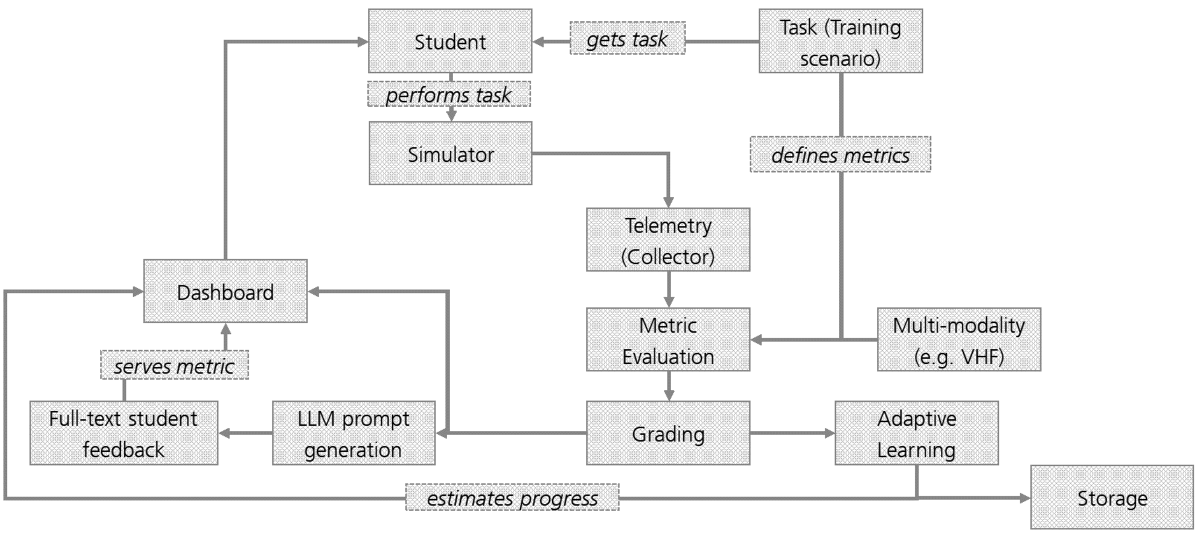

Figure 1. System architecture ILS

4.1. Technical Architecture and ILS integration

The core components of the ILS include the DIS Interface Connector, which captures simulator data in real time using the IEEE 1278 standard and provides fallback support for NMEA AIS protocols. The software components handle analysis and assessment. The Metrics Processing Engine executes maritime-specific evaluation algorithms using inputs such as heading, position and maneuvering data. The LLM feedback generation connects via an Ollama-compatible API, enabling integration with a variety of local or remote language models to generate natural language feedback. Historical assessment data is stored in a MongoDB database with structured collections for sessions, metrics, debriefs, student profiles and scenario parameters. A web-based dashboard serves as the interface for visualizing assessments, performing analytics, and controlling recording sessions. The system is designed for local deployment on dedicated hardware and can operate on a laptop equipped with an RTX 3060 GPU (8 GB VRAM), 16 GB RAM, an Intel i5 CPU and a 256 GB SSD. These hardware and deployment constraints ensure secure offline operation without immediate reliance on cloud-based infrastructure.

5. CORE ILS COMPONENTS

5.1. Task Achitecture and Metrics Processing Engine

The ILS implements a flexible, object-oriented task architecture, as shown in Figure 2 that defines and manages training activities as discrete, assessable units. Each task encapsulates a name and descriptive scenario context, along with a set of evaluation criteria, domain-specific analysis functions, data collectors, and configurable weighting schemes. This design enables a dynamic assessment of user behavior across complex navigational scenarios while simultaneously addressing multiple learning objectives. A core strength of this architecture lies in its end-to-end integration of competencies, which are explicitly embedded from the earliest stage of task design, carried through real-time assessment, and ultimately form the basis for individualized evaluation and feedback. Each task is mapped to one or more competencies within the system's defined framework. such as Course Keeping, Navigation Using Charts, COLREGs Compliance, and Maneuvering. This mapping ensures that the learning objectives are directly tied to recognized performance standards. During execution, the Metrics Processing Engine transforms raw simulator data into normalized performance indicators, applying expert-informed logic such as threshold analysis, phase segmentation, contextual weighting, and event detection. This produces detailed, timestamped assessment data that is explicitly tagged by competency, allowing the system to evaluate not just overall performance but also how well specific skill areas were demonstrated during different phases of the operation (e.g., individual legs of a voyage plan). Critically, this competency tagging persists into the feedback and recommendation stage. Each result is parsed to identify performance across the associated competencies, enabling the system to generate feedback for the student. Competencies that were only partially demonstrated or failed outright are flagged and linked automatically to remedial resources or follow-up tasks designed to strengthen those specific areas. In this way, the system enables a custom learning experience, where each student receives dedicated feedback and tailored learning pathways that address their unique performance profile.

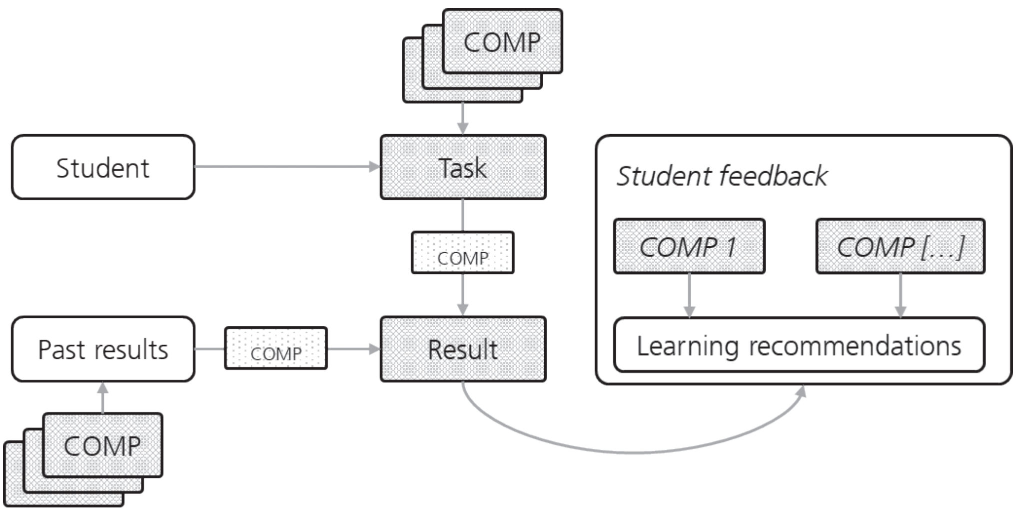

Figure 2. Task architecture with competencies (COMP) linked to tasks and assessment of simulator runs

5.2. LLM Integration for Prompt Engineering

Figure 3 illustrates the architecture used to generate contextualized prompts for feedback generation using a LLM in the ILS environment. The processing layer combines all parts for student feedback into coherent, purposeful instructions for the LLM. To aid in the generation of such instructions, role introduction (framing the model's communication style), performance data (quantitative and qualitative outputs from the assessment system) and a feedback request (providing explicit guidance on what the model should focus on), are integrated into the template. This ensures that the generated prompt accurately reflects both the learner's activity and the instructional goals. The output layer represents the structured feedback tailored to the student's performance and framed by the specified role and context. For example, in a route-following task, the system might start the prompt with:

> *"(Prompt Role) You are an experienced maritime instructor at a maritime academy. Your task is to assess a student's performance in a transit and waypoint-following exercise, where the student must maintain a safe and accurate route using electronic navigation tools."*

This is followed by a structured summary of the student's performance data. An example might look like:

> *"(Prompt Context) The student's overall performance score is 82.5%, with the following breakdown:*
>
> *Course Keeping (Score: 90.0%, Weight: 45.0%) → Impact on Total Score: 45.0%*
>
> *COLREGs Compliance (Score: 60.0%, Weight: 25.0%) → Impact on Total Score: 15.0%"*

The template is concluded by a feedback request section, in which a structured text can be constructed by the instructors in advance, to adjust the feedback to their style and additional formatting requirements. It specifies expectations for the model, such as evaluating strengths and weaknesses, suggesting targeted improvements, and recognizing performance trends. For example, the prompt may request:

"(Feedback Request) Provide detailed feedback including:

1. *Overall evaluation of performance [...]*
2. *Specific strengths and weaknesses [...]*
3. *Actionable recommendations for improvement [...]*
4. *Notable performance patterns [...]*
5. *Suggested focus areas for future training [...]"*

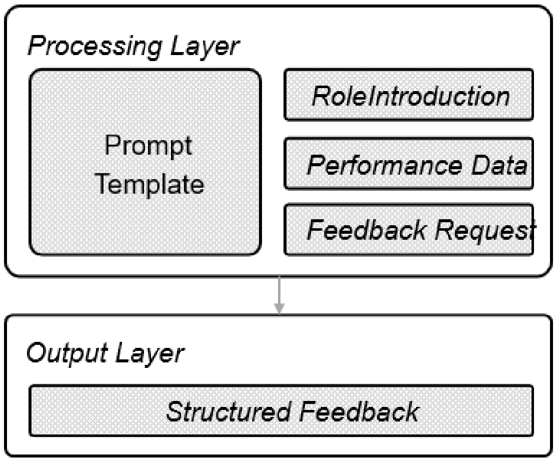

Figure 3. LLM structured feedback generation flow

Together, these sections are dynamically composed into a complete, context-aware prompt. By segmenting prompt in this way, the system allows for customization, extensibility, and instructional control, enabling educators to adapt feedback generation to their pedagogical preferences. This structured composition ultimately aims at producing, role-aligned, and learner-specific feedback that reflects both task performance and instructional priorities.

6. DISCUSSION

6.1. Current Limitations and Development Scope

The current implementation focuses on basic and intermediate maritime training scenarios. Complex decision-making situations such as collision avoidance in high-traffic conditions fall outside the immediate system scope. Several technical constraints should be acknowledged:

- **Validation Status**: While the technical architecture has been implemented and tested for functionality, comprehensive validation with actual students and instructors in real training environments is planned as the next phase of the i-Master project.
- **Computational Dependencies**: Real-time LLM processing requires dedicated GPU resources, potentially affecting system responsiveness in resource-constrained environments or when processing multiple simultaneous assessments.
- **Simulator Fidelity Dependency**: Assessment accuracy is inherently limited by the fidelity and accuracy of the underlying maritime simulator, as the system can only evaluate data provided by the simulation environment.

These limitations define the current boundaries of the system while establishing clear directions for future development and validation activities.

7. CONCLUSION AND FUTURE WORK

This work demonstrates the technical implementation of an automated assessment system for maritime simulator training, successfully integrating real-time data collection, competency-based metrics processing, and LLM-based feedback generation within a modular architecture. The system achieves end-to-end integration from raw simulator telemetry to structured natural language feedback, with explicit competency mapping that enables personalized learning pathways and targeted recommendations.

Technical achievements include the development of a flexible, object-oriented task architecture that embeds STCW competencies throughout the assessment pipeline, from initial task design through final feedback generation. The Metrics Processing Engine successfully transforms simulator data into normalized performance indicators using expert-informed logic, producing timestamped assessment data tagged by specific competencies. The LLM integration demonstrates effective prompt engineering through structured templates that combine role definition, performance context, and customizable feedback requests, enabling instructors to adapt automated feedback to their pedagogical preferences while maintaining consistency across assessments.

Implementation insights reveal that real-time LLM processing requires careful resource management to maintain system responsiveness, particularly when handling multiple simultaneous assessments. The competency-driven architecture proves essential for generating actionable feedback, as explicit mapping between simulator metrics and learning objectives enables fine-grained performance analysis across different operational phases. The modular design successfully supports iterative improvement of individual components while maintaining overall system integrity, and local deployment on standard hardware (RTX 3060, 16GB RAM) demonstrates practical deployment viability for maritime training institutions.

Critical future work will focus on comprehensive validation of educational impact and operational performance in real training environments through the i-Master project's planned validation phase. Essential next steps include structured comparative analysis of system-generated versus instructor assessments using identical simulator runs, user acceptance testing with maritime education professionals, validation of LLM feedback quality and pedagogical effectiveness, and collection of empirical learning outcome data across student cohorts. The system architecture supports expansion of metrics repositories and assessment models, integration of additional performance indicators, and adaptation based on observed user behavior and performance metrics.

The successful technical implementation provides a foundation for future validation research in automated maritime education, establishing the architectural and technical prerequisites necessary to empirically investigate whether intelligent systems can enhance training quality while reducing instructor workload and institutional resource constraints.

REFERENCES

Andriuškevicius, J., & Junzi, S., 2024. Automatic Control With Human-Like Reasoning.

Brügge, E., Ricchizzi, S., Arenbeck, M., Keller, M. N., Schur, L., Stummer, W., . . . Darici, D., 2024. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *National Library of Medicine*.

Cook, D. A., Overgaard, J., Pankratz, S., Del Fiol, G., Aakre, C. A., & ., 2025. Virtual Patients Using Large Language Models: Scalable, Contextualized Simulation of Clinician-Patient Dialogue With Feedback. *Jounal of Medical Internet Research*.

International Maritime Organization. (16. 06 2025). *International Convention on Standards of Training, Certification and Watchkeeping for Seafarers (STCW)*. Von imo.org: https://www.imo.org/en/ourwork/humanelement/pages/stcw-conv-link.aspx abgerufen

Jørgen Ernstsen, S. N., 2020. Performance assessment in full-scale simulators – A case of maritime pilotage operations. *Safety Science*.

Karahil, M., Lützhöft, M., & Scanlan, J., 2023. Formative assessment in maritime simulator-based higher education. *WMU Journal of Maritime Affairs*, 181-207.

Munim, Z. H., & Kim, T.-E., 2023. A Review of Learning Analytics Dashboard and a Novel Application in Maritime Simulator Training. *Training, Education, and Learning Sciences Vol.109*.

Øvergård, K., Nazir, S., & Solberg, A., 2017. Towards Automated Performance Assessment for Maritime Navigation. *TransNav Journal Vol. 11 No. 2*.

Raj, A. (23. 07 2025). *Generative AI Wargaming Promises to Accelerate Mission Analysis*. Von Johns Hopkins Applied Physics Laboratory: https://www.jhuapl.edu/news/news-releases/250303-generative-wargaming abgerufen

Schlichting, M. R., Rasmussen, V., Alazzeh, H., Liu, H., Jafari, K., Hardy, A. F., . . . Kochenderfer, M. J., 2025. LeRAAT: LLM-Enabled Real-Time Aviation Advisory Tool.

Tomkou, D., Fatouros, G., Andreou, A., Makridis, G., Liarokapis, F., Dardanis, D., . . . Kyriazis, D., 2025. Bridging Industrial Expertise and XR with LLM-Powered Conversational Agents.