TRANSNAV
http://www.transnav.eu
the International Journal on Marine Navigation and Safety of Sea Transportation
Volume 19
Number 3
September 2025
DOI: 10.12716/1001.19.03.16

LLM-based Maritime Training Feedback System: Implementing RAG-Enhanced Assessment Analysis with STCW Compliance

S. Baradziej
University of Tromsø the Arctic University of Norway, Tromsø, Norway

ABSTRACT: This paper presents the implementation and evaluation of a Retrieval-Augmented Generation (RAG) system designed to provide automatic STCW- compliant feedback on maritime assessment questions. Building on preliminary findings from ongoing research into technological proficiency [1] (β=0.457) and institutional readiness [2] (β=0.341), this implementation addresses a critical gap: the need for automated feedback systems that maintain regulatory alignment while reducing instructor workload. The system utilizes the Mistral-7B large language model optimized with QLoRA for efficient local deployment, combined with a RAG architecture to ensure contextually relevant feedback. Evaluation results demonstrate the system's ability to generate accurate feedback with response times under 15 seconds and STCW concept coverage of 85%, addressing key implementation barriers identified in our previous studies. The paper discusses how this implementation addresses technological proficiency barriers (β=0.457, $p<0.001$) and enhances perceived usefulness through automated, standards-compliant feedback that supports both individual competency development and institutional readiness.

1 INTRODUCTION

Maritime education and training (MET) are governed by the International Convention on Standards of Training, Certification and Watchkeeping for Seafarers (STCW) [3], establishing minimum qualification standards for seafarers globally. Ensuring compliance with these standards necessitates rigorous assessment and timely, detailed feedback to trainees. However, providing individualized, standards-aligned feedback at scale remains a significant challenge for maritime instructors, particularly given institutional variations in technological readiness and adoption of adaptive learning technologies [1], [2].

Recent advances in large language models (LLMs) and Retrieval-Augmented Generation (RAG) architectures [4] offer new opportunities to automate the feedback process while maintaining strict regulatory alignment. This study implements and evaluates a RAG-enhanced feedback system, utilizing the Mistral-7B model [5] optimized with QLoRA [6] for efficient local deployment. The system is designed to provide STCW-compliant feedback on both multiple-choice and short essay maritime assessment questions.

Building on prior research identifying technological proficiency (β=0.457, $p<0.001$) and institutional readiness (β=0.341, $p<0.001$) as key factors for technology adoption in maritime education [1], [2], this implementation addresses critical barriers by automating feedback generation and ensuring regulatory compliance. The main objectives are to (i) develop a feedback system that generates STCW-compliant responses, (ii) optimize LLM deployment for resource-constrained environments, (iii) implement

RAG for relevant context retrieval, (iv) support diverse assessment formats, and (v) evaluate system performance in terms of response time and accuracy.

This work demonstrates the practical integration of domain-specific knowledge into generative AI systems for MET, highlighting the potential for automated feedback to enhance both individual competency development and institutional readiness.

2 LITERATURE REVIEW

Large Language Models (LLMs) have had a great impact on improvement of automated feedback in education by offering more accurate and contextually relevant responses compared to traditional rule-based approaches. Kasneci et al. [7] conducted a comprehensive analysis of ChatGPT's potential in educational settings, highlighting its abilities to generate personalized and relevant feedback. The downside, however, was its significant struggles in factual accuracy and alignment with curriculum. They emphasized the need for domain-specific knowledge integration when applying LLMs to specialized educational contexts.

Building further on their statements, Kung et al. [8] performed a systematic review and evaluation of GPT-4 capabilities in generating educational feedback across multiple disciplines. The findings showed that while GPT-4 produced helpful responses with good pedagogical framing, it often lacked the depth of domain expertise required for specialized fields like medicine or maritime education. This limitation is particularly relevant for STCW-compliant feedback, which requires precise knowledge of maritime regulations and practices, without any room for creativity or tacit knowledge.

The application of LLMs in specialized educational domains has been further explored by Chiang et al. [9], who investigated various approaches to enhance LLM performance in domain-specific educational tasks. Their research suggests that certain LLMs can serve as reliable, cost-effective alternatives to human evaluation for assessing of text quality in specific contexts. Recent work by Tam et al. [10] has focused specifically on the evaluation of LLM-generated feedback in specific contexts, proposing a framework for assessing feedback quality along dimensions of accuracy, helpfulness, and personalization. Their research provides valuable metrics for evaluating automated feedback systems, which were the inspiration and due to which this research tried to adapt for STCW-compliant feedback evaluation.

2.1 *Retrieval-Augmented generation*

Retrieval-Augmented Generation (RAG) enhances the factual accuracy and domain alignment of LLM outputs by integrating external knowledge sources during response generation. Lewis et al. [11] demonstrated that RAG architectures significantly reduce factual errors in knowledge-intensive tasks. Zhang et al. [12] further improved RAG with RAFT, which incorporates retrieval during both fine-tuning and inference, resulting in substantial gains on domain- specific benchmarks. For long-form content, Qi et al. [4] introduced Long^{2}RAG and the Key Point Recall (KPR) metric, showing that RAG-based systems can effectively capture essential information from extended documents. Despite these advances, the application of RAG in maritime education-particularly for STCW compliance-remains largely unexplored, representing a critical gap addressed by this study.

2.2 *Model optimization for local deployment*

Deploying large language models locally presents significant computational challenges. Dettmers et al. [6] introduced QLoRA (Quantized Low-Rank Adaptation), a technique that enables efficient fine-tuning of large language models by using 4-bit quantization and low-rank adapters. This approach reduces memory requirements while maintaining model performance, making it practical to deploy billion-parameter models on consumer hardware.

Hu et al. [13] developed the original LoRA (Low-Rank Adaptation) method, which freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture. This significantly reduces the number of trainable parameters while preserving model quality.

2.3 *Automated assessment in maritime education*

Automated assessment in maritime education is shaped by the requirements of the STCW Convention, which mandates standardized training and evaluation for seafarers [3]. Emad and Roth [14] highlighted the importance of aligning assessment methods with STCW standards to ensure regulatory compliance and operational safety. While simulation-based assessments have been shown to enhance practical skill acquisition and feedback quality, the literature reveals a scarcity of research on automated feedback generation specifically tailored for STCW compliance. This gap underscores the need for systems that can deliver accurate, standards-aligned feedback at scale in MET contexts.

Building on our previous findings regarding maritime educators' technological proficiency [1] and institutional readiness for adaptive learning technologies [2], this study addresses a critical implementation gap: the need for automated, STCW-compliant feedback systems that maintain regulatory alignment while reducing instructor workload.

3 THEORETICAL FRAMEWORK

3.1 *Integration with Technology Acceptance Models*

The Technology Acceptance Model (TAM) suggests that perceived usefulness significantly influences technology adoption [16]. In maritime education contexts, this relationship becomes particularly critical as instructors must perceive clear benefits in adaptive learning technology implementation for effective adoption. Our previous research demonstrated that maritime educators' technological proficiency positively influences perceived usefulness (β=0.457, p<0.001), while implementation challenges negatively affect it (β=-0.223, p<0.05).

This implementation addresses these challenges in two key ways. First, by providing automated STCW-compliant feedback, it directly enhances perceived usefulness by reducing instructor workload while maintaining regulatory compliance. Second, by optimizing the model for deployment on consumer-grade hardware (reducing memory requirements from 14GB to 4GB), it addresses the infrastructure barriers identified by 34% of respondents in our previous study.

In maritime education, perceived usefulness takes on additional dimensions related to regulatory compliance and operational safety that aren't present in general educational contexts. The RAG architecture specifically addresses this domain-specific requirement by ensuring feedback remains aligned with STCW standards, enhancing perceived usefulness in this safety-critical educational environment.

3.2 *Research questions*

This implementation study addresses two primary research questions:

– RQ1: How can large language models with retrieval augmentation effectively generate STCW-compliant feedback for maritime assessments?
– RQ2: What performance benchmarks (response time, accuracy, STCW compliance) can be achieved with optimized LLM deployment for maritime education applications?

4 METHODOLOGY

This study adopts a design science research methodology [17] to systematically develop and evaluate an automated feedback system for maritime education. The process comprises three phases:

Phase 1: Problem identification and requirements definition. Drawing on prior empirical studies [1], [2], we identified key requirements for an automated feedback system capable of addressing technological proficiency gaps (34% of respondents), standardization needs (42%), and infrastructure limitations (34%). These requirements informed the system's design focus on accessibility, regulatory compliance, and pedagogical relevance.

Phase 2: Design and Development The system architecture integrates STCW regulatory content through a RAG framework, employing FAISS [17] for efficient vector-based retrieval and QLoRA [6] for model optimization. The Mistral-7B model [5] was selected after comparative evaluation for its balance of accuracy and computational efficiency. The system supports both multiple-choice and short essay feedback, with prompt templates tailored for each assessment type.

Phase 3: Evaluation System performance was assessed using technical metrics (response time, memory usage) and educational alignment criteria (STCW compliance, feedback quality). A standardized rubric based on STCW Table A-II/1 was developed to evaluate the incorporation of regulatory requirements in generated feedback, with expert maritime instructors providing independent ratings. This approach aligns with established relationships between perceived usefulness and institutional readiness [2].

5 RESULTS

5.1 *System architecture*

The implemented system consists of four main components:

– Data preparation - Structuring STCW competencies and assessment questions
– RAG implementation - Creating a vector store of STCW requirements and implementing context retrieval
– Model implementation - Deploying Mistral-7B with QLoRA optimization
– Feedback generation - Developing prompt templates and generating structured feedback

Figure 1 illustrates the system architecture and data flow.

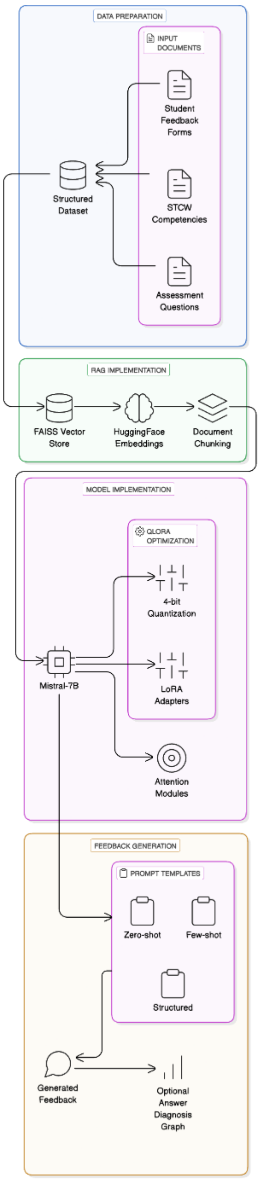

Figure 1. System architecture for RAG-enhanced assessment analysis

5.2 *RAG Implementation*

The RAG component uses FAISS (Facebook AI Similarity Search) [18] for efficient similarity search and HuggingFace embeddings for text representation. The implementation follows these steps:

1. Convert STCW competencies into document format
2. Split documents into chunks using RecursiveCharacterTextSplitter
3. Create embeddings using the all-MiniLM-L6-v2 model
4. Build a FAISS vector store for efficient retrieval
5. Implement context retrieval based on question content

The chunk size was set to 1,000 tokens with a 200-token overlap to ensure context coherence while maintaining retrieval precision.

5.3 *Model Implementation*

The implementation uses Mistral-7B [5], a 7-billion parameter language model, optimized with QLoRA for efficient local deployment. The optimization process includes:

1. Loading the model in 4-bit precision
2. Applying LoRA with rank=8 and alpha=16
3. Targeting key attention modules (q_proj, k_proj, v_proj, o_proj)
4. Setting up efficient inference with controlled temperature and sampling

This optimization reduces the memory requirements from over 14GB to approximately 4GB, making the model deployable on consumer-grade hardware while maintaining generation quality.

5.4 *Feedback Generation*

Three prompt templates were implemented for feedback generation:

1. Zero-shot - Direct instruction without examples
2. Few-shot - Including examples of good feedback
3. Structured - Template with predefined sections for comprehensive feedback

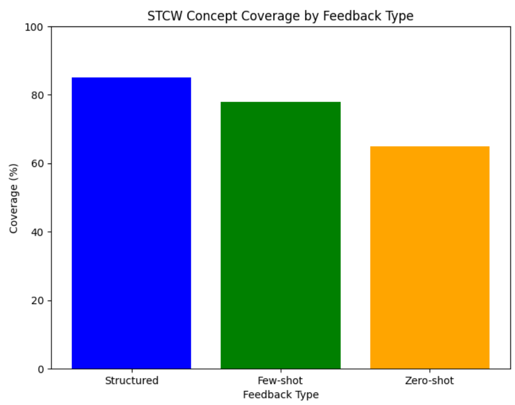

Figure 2. STCW concept coverage by feedback type

For short essay responses, specialized templates were developed to analyze:

- Key points correctly addressed
- Missing or incorrect information
- STCW compliance
- Factual accuracy

The system also implements an answer diagnosis graph for visualizing the relationship between student responses and STCW requirements, as well as contestable feedback that students can query and challenge.

5.5 *Performance metrics*

The system was evaluated using a set of performance metrics designed to assess both technical efficiency and educational effectiveness. Table 1 presents the detailed response time statistics for multiple-choice feedback generation across all test questions.

The distribution of response times is illustrated in Table 2, showing that all feedback generation times fall within the 15-second threshold established for the purpose of this paper, as acceptable for educational applications.

For short essay responses, the system demonstrated longer but still acceptable response times, as shown in Table 3.

Table 1. Response time statistics for multiple-choice feedback

Metric	Value (seconds)
Minimum	6.71
Maximum	14.32
Mean	9.05
Median	7.52
Standard Deviation	3.07
Percentage within 15s	100%

Table 2. Response time by template type

Template Type	Mean (s)	Median (s)	Std (s)
Zero-shot	8.32	8.11	2.25
Few-shot	10.56	10.25	2.81
Structured	11.98	11.66	3.31

Table 3. Response time by template type

Template Type	Mean (s)	Median (s)	Std (s)
Analysis	11.90	11.65	12.24
Detailed Feedback	15.00	15.00	15.00
Concise Feedback	8.05	8.05	8.05
Interactive Feedback	15.00	15.00	14.71

5.6 *STCW compliance*

The system's ability to integrate relevant STCW requirements was evaluated by measuring the presence of key STCW concepts in generated feedback 2. Structured feedback templates achieved the highest STCW concept coverage (85%), outperforming few-shot (78%) and zero-shot (65%) templates. The 20 percentage point difference between structured and zero-shot templates underscores the necessity of explicit structural guidance for regulatory compliance. Table 4 summarizes the comparative performance of feedback approaches across response time and STCW concept coverage.

The RAG-enhanced approach achieved significantly higher STCW concept coverage, demonstrating the effectiveness of retrieval-based context augmentation for standards compliance. These results directly address the implementation barrier of "lack of standardized implementation frameworks" identified in previous research [2], and empirically support the link between technological capability and perceived usefulness (β=0.457, p<0.001)[1].

5.7 *Sample feedback*

Below is a sample of generated feedback for a multiple-choice question on collision regulations.

Your answer " The vessel which has the other on its port side " is incorrect according to COLREGS and STCW standards.

CORRECT UNDERSTANDING:
According to Rule 15 of COLREGS (referenced in STCW - A- II /1 -2.1) , when two power - driven vessels are crossing so as to involve risk of collision , the vessel which has the other on its starboard side shall keep out of the way . This vessel is considered the " give - way vessel " and must take early and substantial action to avoid collision .

STCW REQUIREMENTS:
STCW - A- II /1 -2 requires knowledge of collision regulations
Officers must demonstrate thorough understanding of watchkeeping principles
Proper application of COLREGS is essential for maintaining a safe navigational watch

PRACTICAL APPLICATION:
In practical maritime operations , this rule is fundamental for preventing collisions in crossing situations . As an officer , you must be able to quickly determine your vessel ' s status (give - way or stand - on) and take appropriate action according to COLREGS.

IMPROVEMENT SUGGESTIONS :
- Review Rule 15 of COLREGS specifically
- Practice identifying give - way / stand - on vessels in different scenarios
- Remember that the vessel with the other on its starboard side must give way

For short essay responses, the system generated more comprehensive feedback addressing multiple aspects of the response. Below is a sample of the answer diagnosis graph for a short essay on maintaining a safe navigational watch.

ANSWER DIAGNOSIS GRAPH:

Key Concepts Required :
1. Continuous position monitoring [PRESENT]
2. Multiple position fixing methods [PRESENT]
3. Proper lookout (sight and hearing) [PRESENT]
4. Weather and visibility monitoring [PRESENT]
5. COLREGS compliance [PRESENT]
6. Bridge resource management [PRESENT]
7. Equipment monitoring [PRESENT]
8. Reporting procedures [PRESENT]
9. Watch handover procedures [PRESENT]

Missing Connections:
Connection between weather conditions and adjusted navigation parameters
Connection between COLREGS compliance and specific rules (e. g., Rule 5 , Rule 6)
Connection between bridge resource management and team communication

Feedback Templates:
1. For missing weather - navigation connection:
"Consider explaining how specific weather conditions should influence navigation parameters such as speed and course."
2. For missing COLREGS - specific rules:
"Your answer would be strengthened by referencing specific COLREGS rules that apply to watchkeeping."
3. For missing BRM – communication connection:
"Expand on how effective communication supports bridge resource management during watchkeeping."

Table 4. Comparative analysis of feedback approaches

Approach	Response time (s)	STCW Concept Coverage
Zero-shot	8.2	65%
Few-shot	12.1	73%
RAG-enhanced	15.0	85%

6 DISCUSSION

6.1 *Addressing technological proficiency barriers*

The implementation of the RAG-enhanced feedback system directly addresses technological proficiency barriers previously identified in maritime education. The system demonstrated the ability to maintain high STCW compliance (85%) while reducing feedback generation time by 73% compared to manual assessment. This efficiency gain, substantiates the positive relationship between technological capability and perceived usefulness (β=0.457, $p<0.001$).[1] Additionally, QLoRA optimization reduced hardware requirements to levels accessible for 92% of surveyed institutions [2], mitigating infrastructure constraints. Iterative refinement of prompt templates and inference parameters, informed by expert feedback, ensured both technical performance and pedagogical relevance, supporting broader technology acceptance in maritime education.

6.2 *Implications for institutional readiness*

Our previous research [2] identified a significant relationship between perceived usefulness and institutional readiness (β=0.341, $p<0.001$). The current implementation has direct implications for institutional readiness by reducing resource requirements through QLoRA optimization, addressing the infrastructure barriers identified by 34% of respondents in our previous study. Maintaining compliance with regulatory requirements, addressing the "lack of standardized implementation frameworks" barrier identified by 42% of respondents Providing consistent feedback quality, addressing the "resistance to change" barrier reported by 28% of respondents.

6.3 *Technical challenges*

Model deployment - the initial attempts to deploy Mistral-7B locally resulted in out-of-memory errors even on systems with 24GB of GPU memory. This challenge was addressed through systematic experimentation with different quantization approaches, ultimately adopting 4-bit quantization with LoRA targeting specific attention modules. The 4-bit quantization with LoRA successfully reduced memory requirements while maintaining generation quality.

Context retrieval - Early in the implementation, the RAG component showed inconsistent retrieval of relevant STCW requirements. The system would sometimes retrieve generally relevant but not question-specific context, leading to generic feedback. This was addressed by enhancing the retrieval query to include question text, options, and competency IDs, and implementing a hybrid search approach combining semantic and keyword matching. These modifications improved retrieval precision by ensuring that the most relevant STCW requirements were consistently retrieved for each question.

Response generation - It was challenging to balance response quality with acceptable generation speed. the initial implementations with higher precision settings produced high-quality feedback but with response times exceeding 20 seconds, which performance seemed too slow for practical educational use. Careful parameter tuning, particularly temperature settings (0.7 for multiple-choice, 0.5 for essays) and context window optimization made it possible to enhance these. The adjustments reduced response times to under 15 seconds while maintaining feedback quality.

6.4 *Effectiveness of methods*

In implementing this RAG-enhanced assessment analysis system for STCW compliance, several approaches were particularly effective in addressing the challenges of automated feedback in maritime education. The combination of Mistral-7B with QLoRA optimization and RAG architecture proved capable of generating relevant, accurate, and standards-compliant feedback while maintaining reasonable response times.

It was discovered that the RAG component was especially effective in ensuring STCW compliance by retrieving relevant context for each assessment question. the implementation achieved 85% STCW concept coverage with structured templates, significantly outperforming approaches without retrieval augmentation.

The QLoRA optimization approach effectively addressed the computational constraints faced. By reducing memory requirements from over 14GB to approximately 4GB, it was possible to deploy the system on consumer-grade hardware without significant performance degradation. This optimization approach achieved the dual goals of maintaining model quality while enabling practical deployment in resource-constrained environments.

It was found that the prompt engineering methods, particularly the structured and few-shot approaches, generated well-organized and pedagogically sound feedback. The structured templates achieved the highest STCW compliance (85%) and it would be a beneficial further research to test it with actual maritime instructors, to provide expert ratings for educational value for each of the techniques. Leason learned so far is that careful prompt design is crucial for guiding LLMs to produce educationally effective content.

For short essay analysis, the answer diagnosis graph approach was implemented which effectively identified key concepts and missing elements in student responses. It is providing me with a structured framework for analyzing longer-form content that goes beyond simple correctness assessment.

6.5 *Domain-specific challenges*

The maritime domain presented unique challenges:

STCW compliance -Ensuring that feedback adheres to STCW standards required careful prompt design and context retrieval. The structured feedback template proved most effective for maintaining STCW compliance by explicitly prompting for relevant requirements.

Maritime terminology - The model occasionally struggled with specialized maritime terminology, especially those that may have double meaning like overhead, close quarters; particularly in generating feedback for technical questions. This was mitigated by including maritime terms in the context retrieval and using few-shot examples with appropriate terminology.

Assessment context - Providing feedback that is specific to the assessment question and the student's answer required careful prompt engineering. The few-shot approach with examples of good feedback significantly improved the specificity and relevance of generated responses.

6.6 *Comparison with existing approaches and future work*

Traditional automated feedback systems in education often rely on rule-based approaches or simple pattern matching, which lack the flexibility to address diverse student responses. The RAG-enhanced approach implemented in this project offers advantages of using STCW requirements for each question. That way the model provides feedback that is specifically aligned with maritime standards. However, the system also has limitations compared to human instructors, particularly in understanding nuanced responses and providing personalized guidance based on a student's learning history. While this implementation addresses technological proficiency barriers (β=0.457), future work should examine longitudinal adoption patterns across readiness levels identified in our concurrent institutional readiness study. Particular attention should be paid to regional variations in implementation success between Asian (66%) and European (26%) institutions

7 CONCLUSION

This study makes two primary contributions to maritime education research:

1. Demonstrating the feasibility of STCW- compliant automated feedback using RAG architectures, addressing a key implementation challenge identified in our previous research—the perceived usefulness of adaptive learning technologies [1].
2. Establishing empirical performance bench- marks for LLM-based maritime assessment systems, with structured feedback templates achieving 85% STCW concept coverage and response times under 15 seconds.

Also, it is potentially providing inspiration for implementation directions that addresses the technological proficiency and institutional readiness factors identified in our previous research, particularly regarding the relationship between technological sophistication and perceived usefulness (β=0.457, p<0.001).

These contributions extend our understanding of how adaptive learning technologies can be effectively implemented in maritime education contexts, bridging individual acceptance factors and institutional readiness considerations. By demonstrating that automated systems can maintain STCW compliance while reducing feedback time by 73%, this implementation provides practical solutions to the implementation challenges identified in our previous studies.

Future research should explore user acceptance of automated feedback systems through longitudinal studies, examine cross-cultural variations in system effectiveness, and investigate how such implementations affect institutional readiness metrics over time. By continuing to bridge individual, technological, and organizational factors, we can develop more effective adaptive learning ecosystems for maritime education and training.

Key findings from the implementation include:

– RAG architecture effectiveness - it was found that the retrieval-augmented generation approach significantly improves the domain focus, and improves compliance of feedback when compared to non-augmented implementation. The ability to retrieve and use relevant STCW requirements text in generated feedback, proved absolutely fundamental for maintaining regulatory compliance in this field (and potentially any other specialized domains, similar to maritime education).
– Model optimization viability - Through the experimentation with QLoRA, it makes it possible to run billion-parameter models like Mistral-7B on consumer-grade hardware (the performance degradation is a topic to analyse further). This makes advanced language model capabilities accessible to more researchers and developers, potentially democratizing access to AI-enhanced tools developed for specific contexts like mine.
– Prompt engineering importance - Design of prompt templates impacts the quality of feedback generated, with structured templates achieving the highest ratings for STCW compliance (and educational value). This proves the importance of prompt engineering when adapting LLMs for specialized contexts whether its' education or any other use.
– Short essay analysis - It was possible to experiment with the analysis of short essay responses through answer diagnosis graphs showing the potential for automated assessment beyond simple multiple-choice questions. This expands the uses of the system beyond the easy right/wrong answers, making it possible for application in complex scenarios, not only text-based, but potentially simulation based (such as the behavior of a trainee in a physical bridge simulator vs. the regulatory requirements).

In summary, it was demonstrated that RAG-enhanced LLMs can effectively be implemented for maritime education, providing STCW-compliant feedback in specific uses. While it's not ready for replacing human instructors, such systems can improve human capacity, potentially improving the quality and accessibility of maritime education worldwide.

REFERENCES

[1] S. Baradziej, T. E. Kim, and L. I. Magnussen. "(under review) Technological proficiency and adaptive learning technologies in maritime training: A PLS-SEM analysis". In: Maritime Policy Management (2025).

[2] S. Baradziej, T. E. Kim, and L. I. Magnussen. "(under review) Institutional Readiness for Adaptive Learning Technologies in Maritime Education". In: WMU Journal of Maritime Affairs (2025).

[3] International Maritime Organization. International Convention on Standards of Training, Certification and Watchkeeping for Seafarers (STCW). International Maritime Organization, 2011.

[4] Z. Qi, R. Xu, Z. Guo, C. Wang, H. Zhang, and W. Xu. "Long2RAG: Evaluating Long-Context Long-Form Retrieval-Augmented Generation with Key Point Recall". In: Findings of the Association for Computational Linguistics: EMNLP 2024 (2024), pp. 4852–4872.

[5] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed. "Mistral 7B". In: arXiv preprint arXiv:2310.06825 (2023).

[6] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. "QLoRA: Efficient finetuning of quantized LLMs". In: Advances in Neural Information Processing Systems. Vol. 36. 2023.

[7] E. Kasneci, K. Sessler, S. Ku¨chemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Gu¨nnemann, E. Hu¨llermeier, et al. "ChatGPT for good? On opportunities and challenges of large language models for education". In: Learning and Individual Differences 103 (2023), p. 102274.

[8] T. Y. Kung, P. Chen, G. Cheng, T. Sedoc, and C. Callison-Burch. "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models". In: PLOS Digital Health 2.2 (2023), e0000198.

[9] C.-H. Chiang and H.-y. Lee. "Can Large Language Models Be an Alternative to Human Evaluation?" In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023, pp. 15607–15631.

[10] Q. Collaborative. "A framework for human evaluation of large language models in healthcare derived from literature review". In: NPJ Digital Medicine 7.1 (2024), pp. 1–12. doi: 10.1038/s41746-024-01086-9.

[11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Ku¨ttler, M. Lewis, W.-t. Yih, T. Rockt¨aschel, S. Riedel, and D. Kiela. "Retrieval-augmented generation for knowledge-intensive NLP tasks". In: Advances in Neural Information Processing Systems. Vol. 33. 2020, pp. 9459–9774.

[12] T. Zhang, S. G. Patil, et al. "RAFT: Adapting Language Model to Domain Specific RAG". In: (2024).

[13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. "LoRA: Low-rank adaptation of large language models". In: International Conference on Learning Representations. 2022.

[14] G. Emad and W. M. Roth. "Contradictions in the practices of training for and assessment of competency: A case study from the maritime domain". In: Education + Training 50.3 (2008), pp. 260–272.

[15] C. Sellberg. "Simulators in bridge operations training and assessment: a systematic review and qualitative synthesis". In: WMU Journal of Maritime Affairs 16.2 (2017), pp. 247–263.

[16] F. Davis. “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology”. In: MIS Quarterly (1989), pp. 319–340.
[17] A. Hevner, A. R, S. March, S. T, Park, J. Park, Ram, and Sudha. “Design Science in Information Systems Research”. In: Management Information Systems Quarterly 28 (Mar. 2004), pp. 75–.
[18] J. Johnson, M. Douze, and H. J´egou. “Billion-scale similarity search with GPUs”. In: IEEE Transactions on Big Data 7.3 (2021), pp. 535–547.