The International Maritime Transport and Logistics (MARLOG 15) - ISSN 2974-3141
http://dx.doi.org/10.21622/MARLOG.2026.15.1.121
Smart & Resilient Logistics Corridors Future Gateway to the Green Trade

DYNAMIC DECISION SUPPORT SYSTEMS: CUSTOMIZING RAG KNOWLEDGE BASES FOR SPECIFIC MARITIME MISSIONS

Kuderna I. Benţa [1], Klára Orbán [1] and Dana C. Deselnicu [2]

(1) Department of Computer Science, Babeş-Bolyai University of Cluj-Napoca, Cluj-Napoca, Romania, email: kuderna.benta@ubbcluj.ro, klara.orban@stud.ubbcluj.ro

(2) Department of Entrepreneurship and Management, Faculty of Entrepreneurship, Business Engineering and Management, National University of Science and Technology Politehnica Bucharest, Bucharest, Romania; email:dana.deselnicu@upb.ro

Keywords: Retrieval-Augmented Generation, Offline LLM, Deep-Sea Navigation, Knowledge Management, Decision Support Systems

ABSTRACT

Maritime operations of long-haul transoceanic voyages frequently face a connectivity desert, where leveraging cloud-based Artificial Intelligence is prohibited due to the absence or high expenses of satellite internet. This paper discusses a new Mission- and Polar-Specific Offline Retrieval-Augmented Generation framework that aims to provide crews with high-fidelity, localized decision support. Unlike generic AI models, the proposed system will be dynamically ingesting a curated corpus comprising vessel-specific technical manuals, cargo-specific Material Safety Data Sheets, and route-specific maritime regulations. The architecture leverages Edge AI by deploying quantized Large Language Models and vector databases on ruggedized onboard hardware, ensuring sub-second inference latency without any external connectivity. To solve the challenge of data freshness during long isolation periods, we have proposed a Hybrid Synchronization Mechanism that enables "delta-vector" updates-transmitting only critical new embeddings through low-bandwidth satellite links to keep local knowledge bases current with minimum data overhead. We show evidence that this framework significantly reduces the cognitive load on officers and engineers and increases operational safety and autonomy. This study concludes that mission-tailored RAG systems are a crucial step toward the digital transformation of resilient and autonomous maritime transport.

1. INTRODUCTION

The maritime industry is today faced with a double transformation: the emergence of autonomous systems and the rise of increased regulation. The modern ships are transportation means. They are complex floating production units that produce also data in extremely large quantities. As the shipping industry is transitioning into the age of Maritime 4.0, there is one important technological issue: the Deep-Sea Connectivity Gap [1]. While other sectors are applying cloud-based Artificial Intelligence (AI) systems for optimization of decision-making, ships sailing across oceans often encounter isolated data silos, where satellite connectivity results in extremely large latency, low bandwidths, and high expenses. These, altogether, make the use of cloud-based AI unfeasible or even impossible for sailing ships. The conventional approach applied today for information management in the maritime industry is based exclusively on static PDF-format user manuals and other documents, along with the collective knowledge of the seafarers' experience during sailing. During emergency or critical situations, searching through thousands of pages is inefficient and error-prone, with a pressing need for a local intelligent decision-support system that can supply accurate context-related information without necessarily requiring internet connectivity.

We show how we can combine the technology of Retrieval-Augmented Generation (RAG) into Edge Computing. RAG is a method of using AI to better utilize the output of a Large Language Model (LLM) by using a well-established, authoritative knowledge base [2]. The unique value of our contribution is that we adopted the principles of RAG and customized them for use on a ship's local server to create Mission-Specific RAG. The innovation of our research can be summarized into three distinct points. The first point is that we demonstrate how to create a standalone contextual entity by using an offline LLM that meets the needs of a single voyage; it is possible to achieve this by resource-effective technologies designed for the challenges faced by the maritime industry. The second point is to outline computational, efficient, architecture to run on ruggedized, low-power hardware that is designed for maritime use. The third point is the exploration of a hybrid synchronization solution that allows for the updating of the ship's knowledge from evolving information that can be received through the minimal satellite transmissions that can be sent on demand as needed.

Our proposed framework is oriented to advance the safe operation of vessels and to reduce the burden of cognitive task loading placed on seafarers and, nonetheless, to support the establishment of more resilient autonomous shipping systems [3].

2. CONTEXT ANALYSIS AND RELATED TECHNOLOGIES

The digitalization of the maritime sector has created sensor-heavy environments that often overwhelm human cognitive limits. This led to the necessity of using advanced AI decision support [4]. According to Nomikos et al. [5], low bandwidth or even discontinuous internet connection in satellite communications (SATCOM) while roaming the oceans are not a suitable infrastructure for standard cloud-native Large Language Models (LLMs). Additionally, the generalized LLMs tend to produce sometimes inaccurate information that poses significant operational risks in high-stakes settings where precise technical details are essential [6]. Although LLMs' fine-tuning is a method for adapting to specific domains, it is resource-intensive and results in static knowledge. By contrast, an offline Retrieval-Augmented Generation (RAG) approach is more suitable for maritime applications because it provides grounded AI through verifiable references, facilitates easy updates to knowledge via local technical libraries, and operates effectively on low-cost edge hardware without needing a constant internet connection [7].

Edge Computing refers to the act of performing computations closer to the source (in our case, the ship) rather than from the data center (the cloud). Recent breakthroughs in the area of Quantization (compressing AI to make it cloud-like) and Industrial Edge Servers [8], [9], i.e. NVIDIA Jetson or rugged NUC designs, allow the integration of highly complex AI at the edge level. We think that the hardware is robust enough to handle the fluctuating temperatures, humidity, saltiness and electromagnetic interference (EMI) characteristic of shipboard environments, such as engine rooms or the bridge, ensuring the continuous operation of an autonomous, offline RAG system.

3. OFFLINE MARITIME RAG SYSTEM ARCHITECTURE

The proposed system functions as an independent ecosystem within the vessel's local network. It is structured into three separate operational layers-Ingestion, Storage, and Inference-to enhance performance and reliability without relying on the cloud. In the pre-voyage shore-side Ingestion stage, unstructured technical data (such as the IMO Polar Code) is parsed and then processed through semantic chunking and vector embedding using the all-mpnet-base-v2 model, converting the text into high-dimensional mathematical forms [10]. This vectorized knowledge is then moved to the onboard storage layer, where a hybrid search system integrates a ChromaDB vector store for semantic retrieval with a BM25 index for keyword frequency matching [11]. This dual-engine strategy ensures precise capture of essential maritime terminology, while the local data nature guarantees complete sovereignty, keeping sensitive schematics and proprietary procedures secure and private within the ship's physical boundaries [12].

The onboard inference layer, hosted on an edge server, utilizes a quantized Phi-3.5 Mini Instruct model to execute the final generation logic [13]. When a crew member submits a query, the system retrieves the top fifteen candidate chunks through hybrid search and subsequently applies a Cross-Encoder reranker (ms-marco-MiniLM-L-6-v2) to narrow the selection to the three most contextually relevant fragments. These fragments are injected into a strict system prompt that mandates factual adherence, instructing the model to output a "Data is insufficient" warning if the local context does not support the query. This proposed pipeline delivers a low-latency, verifiable response that is anchored by source and page citations, ensuring that critical decision support remains functional and safe from external interference regardless of the vessel's location or satellite connectivity status. Operational advantages of this architecture are further detailed in comparison to traditional cloud-based models in Table 1.

Table 1. Comparative Analysis of Standard Cloud-based LLMs vs. Maritime Offline RAG

Metric	Standard Cloud LLM (e.g., GPT-4)	Maritime Offline RAG (Proposed)
Connectivity	Requires constant high-speed internet	Fully Functional Offline - optional low-bandwidth sync
Latency	Variable - depends on satellite signal/jitter	Low & Consistent - local edge processing
Data Cost	High - expensive VSAT/Starlink data usage	Zero/Negligible - processed on local hardware
Reliability	Fails during "blackouts" or heavy weather	High Resilience - available 24/7 regardless of weather
Data Privacy	Proprietary logs/manuals sent to the cloud	Maximum Security - all data stays on the ship's server
Content Accuracy	Prone to general hallucinations.	Grounded in specific ship manuals and mission data
Initial Setup	Zero (Plug-and-play).	Moderate - requires hardware and pre-mission indexing

The use of a 3.8 billion parameter model may appear modest for complex maritime analytics. However, the development phase prioritized a balance between high-precision accuracy and hardware versatility by utilizing an NVIDIA RTX Ada 5000 for benchmarking. Based on these tests, the architecture was intentionally designed around the Phi-3.5 Mini Instruct model to achieve a lightweight footprint that remains highly performant on varied shipboard hardware. The primary objective was to ensure that the system provides nearly instantaneous offline responses, even when onboard edge servers do not meet the advanced specifications of shore-side data centers [14], [15]. The system is designed to work well even on ships with limited computing power. It helps more ships use advanced AI for important tasks without losing reliability. To keep data up to date on long trips, the system uses a Low-Bandwidth Synchronizer. This helps avoid the high costs and problems of satellite communication [16].

When rules change, the system doesn't download whole manuals again. Instead, it updates only the changed parts. A server on land makes new data packets for the updates and sends them to the ship. These updates are usually very small, say less than 1% of the original file size. This way, the ship's information stays current, and the crew gets the latest safety updates without using too much data. The architecture of this approach is depicted in Figure 1.

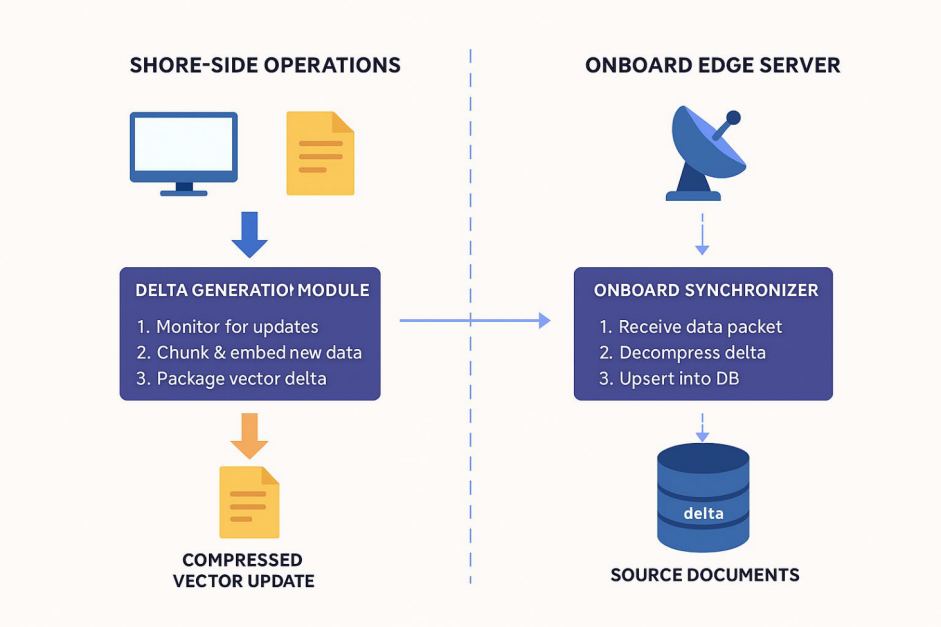

Figure 1: Hybrid Synchronization Mechanism

4. DYNAMIC DATA INGESTION AND MISSION-SPECIFIC CUSTOMIZATION

To establish a dynamic system for this architecture, this proposed framework of work consists of a preparation phase and a shipboard operational phase, which are separated by ship departure, with the former being completed through Shore-Side Operations Center processing, and the latter through Shipboard Operations Center processing.

To ensure prioritization of information by the LLM and prevention of document interference, which could arise due to conflict between general ship operations regulations and cargo handling regulations, this ingestion stage provides a hierarchical taxonomy of information.

The taxonomy has three primary levels: a Static Core, which consists of immutable documentation such as P&IDs and engine documentation, and a Variable Mission Layer, which consists of information for each particular mission such as cargo documentation and a Crew-Specific fragment, which reduces documentation complexity according to individual officer training, with options including 'Cadet Quick-Start Guides' and 'Chief Engineer

Schematics.' The 'Knowledge Package' is then transmitted through a fast physical connection from the center to the ship's Edge Server, which communicates through a local RAG Inference Engine with a vector database, allowing this system to function autonomously within this closed network onboard the ship and providing real-time, grounded support to ship personnel for their decisions without external cloud support through utilizing this onboard engine and database setup with vector processing of information.

Designed to handle the varied characteristics of maritime data, the system leverages an automatic multimodal processing pipeline aimed at consolidating varied data sources into a coherent vector format. The pipeline utilizes specific data preprocessing steps, such as Optical Character Recognition (OCR) for scanned blueprints and markdown serialization for structured tabular data, to ensure that even data traditionally considered 'silent,' such as cargo listings or parts inventories, are translated into semantically dense, machine-readable text. As a result, the system allows the LLM to reason about the entire maritime information ecosystem, not just the text, through a streamlined retrieval interface based on its computer-programmed translation of varied data sources into a coherent whole.

Beyond its backend processing, the system operates as a versatile assistant to the shipboard user on any authorized terminal on the vessel's local network. The architecture deploys the interface across bridge consoles, engine room terminals, and durable handheld tablets, ensuring that essential information is readily accessible wherever needed - whether for navigating icy waters or conducting maintenance in machinery areas. This decentralized access transforms the static knowledge base into a mobile decision-support tool that stays synchronized across all departments.

Operating within a secure, strictly offline environment, the system allows data access only to personnel physically present on board, protecting sensitive technical manuals and proprietary vessel schematics from cyber threats. This air-gapped setup guarantees that the digital repository remains secure and fully operational, regardless of the ship's location or the status of external satellite connections. By eliminating reliance on cloud-based authentication, the system offers a robust fail-safe that upholds data sovereignty and ensures uninterrupted availability even in high-latitude blackout zones.

5. REGULATORY VALIDATION OF CURRENT DEVELOPMENT

The current stage of development emphasizes safety and adherence to environmental standards in polar regions, where operational needs differ greatly from typical international shipping rules. The system's knowledge base was built using essential regulatory frameworks, such as the Guidelines for the Development of a Polar Water Operational Manual (PWOM), which outlines the basic requirements for operating vessels in ice-covered waters. This was further enhanced with technical details from the International Association of Classification Societies (IACS), particularly UR I2, which specifies the structural needs for polar-class vessels [17], and UR I3, which addresses the design and machinery aspects for ships in extremely cold conditions [18]. By incorporating these specific documents, the localized RAG architecture enables the crew to access complex ice-navigation protocols and structural constraints in real-time (see Figure 2), ensuring that decision-making is based on authoritative standards even when the vessel is in complete isolation of high-latitude blackout zones.

This localized approach is further validated by recent open-source simulation frameworks designed for polar navigation [19]. Figure 2 illustrates the flow of the RAG engine in processing regulatory message prompts for the purpose of real-time responses for queries, thus filling the gap between raw data of IACS/IMO and decision-making. Contrary to traditional LLMs, its engine is designed to analyze the user's inquiry in retrieving technical

constraints based on the vessel's specific Polar Class. As evidenced in its output example, its function is more than presenting generic suggestions; it already introduces a specific regulatory anchor like hull stresses in justification for its claim in its domain.

This comprehensive approach is crucial when dealing with uncertainties that arise from the interaction of operational guidelines and technical parameters. While PWOM offers broad guidelines for operational maneuvers, the RAG engine utilizes the data to alert the navigator if a proposed action exceeds the specific stress limits of the vessel's assigned polar ice class. By integrating these typically separate information streams, it becomes clear how the system evolves from a functional search engine to a consultative safety platform that aligns regulatory restrictions with immediate operational needs. Additionally, the localized database prevents inaccuracies by anchoring model responses firmly to the information provided by the IACS [20] and IMO documents [21].

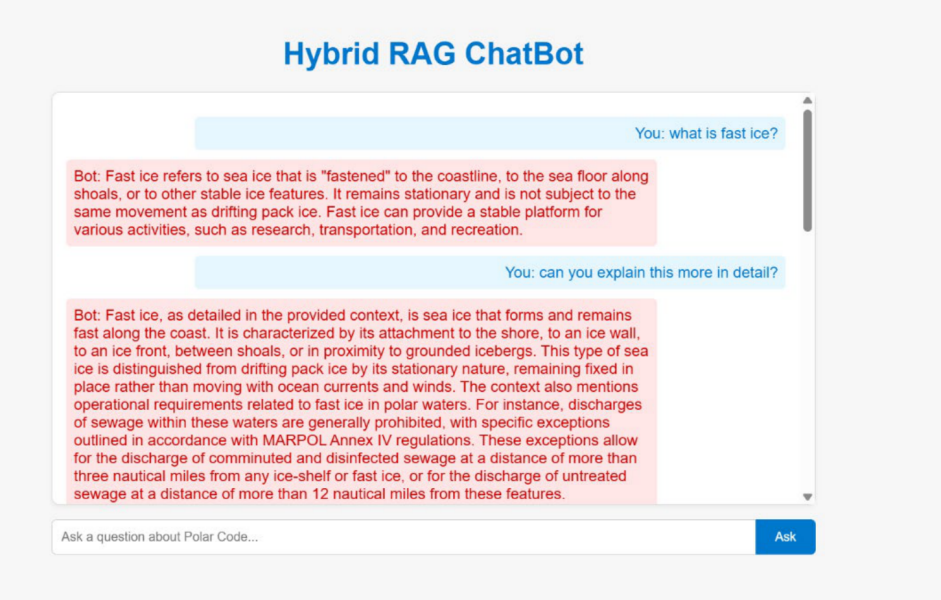

Figure 2: Hybrid RAG Chatbot sample

A comprehensive approach of this kind is essential, especially when resolving uncertainties in terms of operational directives and technical constraints. While PWOM promotes general directives for operational maneuvers, the RAG engine enables it to utilize the dataset for notification of the navigator regarding any action plan that violates specific stresses of operational constraints of the assigned Polar Ice Class of the concerned ship. However, due to the consolidation of both distinctly disparate information flows, it is very easy to understand how it shifts from being a search engine tool for operational activities to becoming a consultative support tool for safety, aligning operational constraints of immediate requirements against restrictions of regulations. Further, because of the localized database, it inhibits hallucination, very closely keying results of the model on operational directives of information from IACS/IMO documents.

6. CONCLUSIONS

This integration of Mission-Specific Offline RAG systems is an imperative leap towards the digitalization of the maritime industry. Through the integration of Large Language Models that are no longer cloud-dependent, as previously required, it is evident that the concept of "intelligence at the edge" is not just possible but imperative to deep-sea missions. Having moved from a repository of information stored on paper and presented in a hardcopy format towards an investigative AI assistant is an incredible leap towards sailing through the complexities of the sea at incredible speed and accuracy.

The proposed solution caters to the important barriers affecting the inclusion of AI technology in the shipping sector, which include intermittent connectivity, worries about data security, and the potential for hallucinations. The examples are highlighted. The importance of this solution is more than a luxury, as it is a vital tool that improves the preparedness of the crew in the event of technological, regulatory, or medical distress.

As the industry moves toward increased autonomy and Smart Shipping, the role of offline RAG systems will likely expand. The integration of real-time data from IoT sensors within the RAG framework directly into the AI system pipeline will be an important area of research for the future, enabling the AI system not only to "read" the manual but also "sense" its surroundings. The use of mission-aware AI localization will thus initiate a whole new era of maritime self-reliance, assuring that even when physically lost at sea, a ship will never be intellectually lost.

7. ACKNOWLEDGMENT

This work was partially supported by the ERASMUS-EDU-2023-EUR-UNIV, Project 101124676 - EELISA, funded by the European Union, https://eelisa.eu/.

8. DISCLAIMER

Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

9. REFERENCES

[1] I. Sanchez-Gonzalez, L. Diaz-Gallego, and J. Canning, "Digital Transformation in the Shipping Industry: A Network-Based Bibliometric Analysis," Journal of Marine Science and Engineering, vol. 13, no. 5, p. 894, 2025.

[2] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 9459-9474.

[3] M. Lind, M. Michaelides, R. Ward, and R. T. Watson, Eds., Maritime Informatics, 1st ed. Heidelberg: Springer, 2020.

[4] A. S. Abdou, A. El-Hamawy, and M. Mostafa, "Artificial Intelligence in Maritime Transportation: A Comprehensive Review of Safety and Risk Management Applications," Applied Sciences, vol. 14, no. 18, p. 8420, 2024.

[5] N. Nomikos, P. K. Gkonis, P. Trakadas, and D. Vouyioukas, "Sailing into the Future: Technologies, Challenges, and Opportunities for Maritime Communication Networks in the 6G Era," Frontiers in Communications and Networks, vol. 5, p. 1439529, 2024.

[6] A. Reitz, "Optimizing Satellite Communication for Maritime Autonomous Surface Ships (MASS) Monitoring," Journal of Physics: Conference Series, vol. 3123, no. 1, p. 012039, 2025. doi: 10.1088/1742-6596/3123/1/012039.

[7] A. Vaswani et al., "Attention is All You Need," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 5998-6008.

[8] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," arXiv preprint arXiv:2305.14314, 2023.

[9] J. Lin et al., "AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration," arXiv preprint arXiv:2306.00978, 2023.

[10] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6769-6781.

[11] Y. A. Malkov and D. A. Yashunin, "Efficient and Robust Approximate Nearest Neighbor Search using Hierarchical Navigable Small World Graphs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 4, pp. 824-836, 2020.

[12] Qdrant Team, "Vector Database for the Edge: High-Performance Similarity Search on Constrained Devices," 2025. [Online]. Available: https://qdrant.tech/documentation/. [Accessed: Dec. 26, 2025].

[13] G. Gerganov, "llama.cpp: Inference of LLaMA model in pure C/C++," 2024. [Online]. Available: https://github.com/ggerganov/llama.cpp. [Accessed: Dec. 26, 2025].

[14] H. Wang, Z. Liu, W. Liu, and X. Zhang, "AI-powered Radar and Satellite Imaging for Maritime Threat Detection," Marine Policy, vol. 118, p. 104015, 2020.

[15] Y. Zhang and S. Liu, "Predictive Analysis of Vessel Movements using Machine Learning Models," Journal of Navigation, vol. 72, no. 4, pp. 892-910, 2019.

[16] P. Johnson and K. Smith, "The Cost of Isolation: Economic Impacts of Satellite Latency on Global Fleet Management," Maritime Economics & Logistics, vol. 25, no. 2, pp. 215-234, 2023.

[17] International Association of Classification Societies (IACS), "UR I2: Structural Requirements for Polar Class Ships," Rev. 4, Dec. 2019. [Online]. Available: https://iacs.s3.af-south-1.amazonaws.com/wp-content/uploads/2022/04/12082952/ur-i2-rev4-dec-2019-ul.pdf

[18] International Association of Classification Societies (IACS), "UR I3: Machinery Requirements for Polar Class Ships," Rev. 2, Dec. 2024. [Online]. Available: https://www.classnk.or.jp/hp/pdf/info_service/iacs_ur_and_ui/ur-i3-rev.2-corr.1-dec-2024-ul.pdf

[19] Klára Orbán, Stefan Lüdtke, and Kuderna-Iulian Benţa. "Navigating Polar Routes: An Open-Source Simulation Framework for Autonomous Shipping." Zenodo, 2025, https://doi.org/10.5281/zenodo.17120185.

[20] International Association of Classification Societies (IACS), "UR I1: Polar Class Descriptions and Application," Rev. 2, 2016. [Online]. Available: https://iacs.org.uk/publications/unified-requirements/ur-i/

[21] International Maritime Organization (IMO), International Convention for the Safety of Life at Sea (SOLAS), Consolidated ed. London: IMO Publishing, 2024.