ICIC Express Letters
Volume **20**, Number **5**, May **2026**

ICIC International ©2026 ISSN 1881-803X
pp. **543–552**

LURKER: LANGUAGE MODELS FOR UNVEILING MARITIME RISKS VIA KNOWLEDGE-DRIVEN EVENT FORECASTING AND REASONING

Seongmoon Hong[1], Taekhyun Park[1], Dohee Kim[2], Sangmin Jo[3] and Hyerim Bae[1,*]

[1]Department of Data Science
Graduate School of Data Science
[3]Major in Industrial Data Science and Engineering, Department of Industrial Engineering
Pusan National University
2, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan 46241, Korea
{ hongsco0718; pthpark1; 27461a }@pusan.ac.kr; *Corresponding author: hrbae@pusan.ac.kr

[2]Department of Artificial Intelligence Engineering
Changwon National University
20, Changwondaehak-ro, Uichang-gu, Changwon-si, Gyeongsangnam-do 51140, Korea
kimdohee@changwon.ac.kr

Received July 2025; accepted October 2025

Abstract. *Maritime accident risk prediction has highlighted the limitations of purely statistical classifiers and numerical scoring schemes, driving demand for scenario-driven analytical methods that can elucidate both the context and progression of incidents. In response, we propose LURKER (Language models for Unveiling maritime Risks via Knowledge-driven Event forecasting and Reasoning), an explainable framework that leverages the reasoning capabilities of Large Language Models (LLMs) to analyze maritime accident severity. LURKER quantifies accident risk by integrating a tree-based classification model with isotonic regression for probability calibration, and then generates dynamic scenarios through a prompt-engineering module that summarizes tribunal reports and incorporates real-time data. By employing Retrieval-Augmented-Generation and chain-of-thought prompting, LURKER produces context-rich descriptions of causal factors, incident progression, and potential impacts. This approach transcends purely numerical analysis, offering comprehensive explanations of accident causes and consequences, and functions as an intelligent decision-support tool. Case-based experiments integrating tribunal reports, AIS vessel-track data, and meteorological observations demonstrate LURKER's feasibility and practical utility as a severity-prediction tool for operational maritime safety management.*

Keywords: Maritime accident, Large language model, Prompt engineering, Severity prediction, Probability calibration

1. **Introduction.** Maritime accidents inflict severe societal and economic losses through casualties, property damage, and environmental contamination, and both their frequency and impact have escalated with the advent of larger vessels and increased traffic density [1,2]. In the Republic of Korea, an average of roughly 2,800 maritime accidents per year have been reported over the past five years, underscoring the urgent need for enhanced safety management and preventive strategies [3]. Unlike land-based collisions, maritime incidents lack well-defined physical pathways and arise from complex interactions among vessel condition, meteorological factors, and traffic congestion, making root-cause attribution particularly challenging [4]. Moreover, the relative rarity of severe accidents gives rise to data sparsity, further complicating robust risk modeling [4,5]. These characteristics highlight the necessity of an integrated, systematic approach that can both elucidate

DOI: 10.24507/icicel.20.05.543

accident causes and quantitatively forecast risk in advance [6]. Recent research in maritime risk prediction has embraced diverse machine learning and statistical techniques to improve quantitative risk assessment.

Munim et al. [4] applied Automated Machine Learning (AutoML) to forty years of Norwegian coastal accident records – covering groundings, collisions, and fires – and demonstrated performance gains by incorporating meteorological variables. Brandt et al. [5] fused fifty-one weather features with Norwegian accident data and showed that a Light Gradient Boosted Trees model achieved markedly higher accuracy, with wind speed, sea-level pressure, and visibility emerging as dominant predictors. Spatial risk modeling has also garnered attention. Yang et al. [6] leveraged GIS-based kernel density estimation and spatial autocorrelation to identify high-density accident zones in China's Fujian waters and employed Random Forest and Gradient Boosting Decision Trees to predict grid-level risk. Efforts to forecast incident specifics have likewise advanced. Shin and Yang [1] benchmarked XGBoost, Random Forest, and neural networks on Busan Port data, finding Random Forest superior and highlighting the importance of VTS operator expertise and zone-specific characteristics. Kim et al. [7] proposed an integrated framework combining tribunal verdicts and accident statistics, using KeyBERT to extract causal keywords and visualizing them on spatial grids to underpin quantitative severity assessment. Shintani et al. [2] developed a LightGBM model for small-vessel accidents in Japan's Seto Inland Sea and applied SHAP analysis to pinpointing key risk factors, informing safety management and rescue operations.

Although these studies have enhanced predictive fidelity and provided actionable insights, existing studies have primarily relied on binary classification or statistical analyses and therefore lack causal explanatory power regarding why accident is likely to occur. Moreover, research integrating diverse data sources to explain the causes and progression of maritime accidents remains limited. In particular, the application of Large Language Models (LLMs) to generate reasoning-driven, narrative-style scenarios remains in its infancy, with few examples harnessing LLMs' capabilities to contextualize accident causes and trajectories.

To address these gaps, we present LURKER, an explainable framework that synergizes quantitative risk estimation with dynamic scenario generation, where risk scenarios are adaptively derived according to varying meteorological, oceanographic, and spatiotemporal conditions. LURKER comprises two principal modules: 1) A risk estimation module that integrates a tree-based classifier with isotonic regression to yield well-calibrated severity scores; 2) A scenario generation module that employs prompt engineering to merge maritime tribunal summaries with real-time AIS and meteorological data, leveraging Retrieval-Augmented-Generation (RAG) and Chain-of-Thought (CoT) prompting to produce interpretable narrative scenarios of causal factors, and severity assessments. By transcending mere numerical outputs, LURKER offers comprehensive, context-rich explanations that enhance situational awareness and decision support in maritime.

The contributions of this study are as follows: 1) The development of a prompt engineering-based LLM summarization system for tribunal verdicts, augmented with reasoning-based refinement to improve informational accuracy and trustworthiness; 2) The design of a dynamic scenario generation architecture that combines calibrated risk scores from a tree classifier-isotonic regression pipeline with real-time data and RAG/CoTinputs to produce explainable severity analyses; 3) A case-based evaluation using real tribunal reports, AIS vessel-track data, and weather observations, demonstrating LURKER's practical applicability as a dynamic severity analysis tool in operational maritime environments.

The remainder of this paper is organized as follows. Section 2 presents a description of the proposed method. Section 3 provides the results derived from our experiments and presents results on interpretability and applicability. Finally, Section 4 presents the conclusions along with future research ideas.

2. **Proposed Method.** This section describes our methodology. An overall framework of the proposed method is shown in Figure 1.

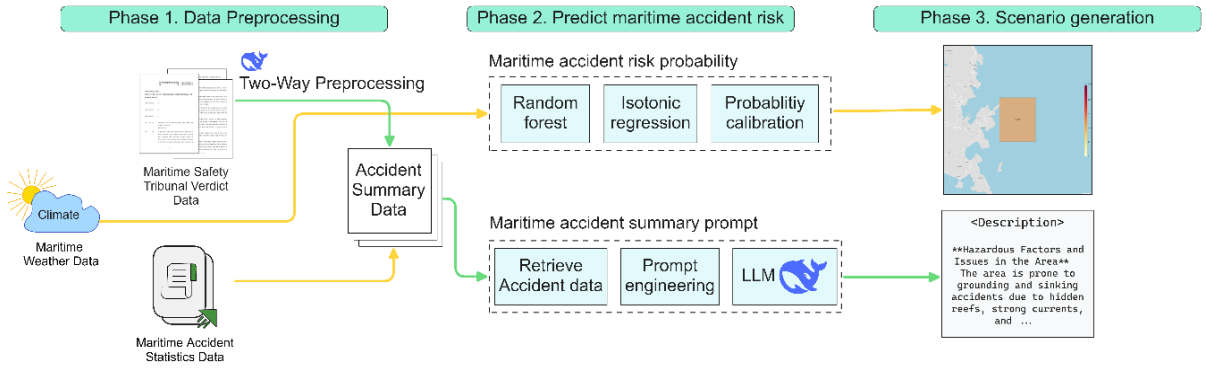

FIGURE 1. Overall framework of the proposed method

Our methodology is structured as a three-step pipeline. First, we devised prompt-engineering templates to instruct a Large Language Model (LLM) to extract only explicit, verifiable facts – such as accident location, vessel specifications, accident type, and damage severity – from unstructured maritime-safety tribunal verdicts, thereby minimizing hallucination. Second, we trained tree-based classifiers to predict the binary occurrence of maritime accidents and applied isotonic regression to their raw probability outputs for bias correction; the resulting calibrated probabilities served as our quantitative risk indicator. Finally, we merged each event's calibrated risk score and structured accident summary with real-time environmental variables (seasonality and time-of-day, wind speed, precipitation, visibility, wave height, and swell parameters) into a single JSON-formatted prompt. Under Retrieval-Augmented-Generation (RAG), Chain-of-Thought (CoT) prompting paradigm, the LLM then unfolds stepwise, interpretable reasoning to generate coherent, context-rich navigational risk scenarios that integrate both quantitative metrics and narrative insights.

Overall, the proposed pipeline adopts a knowledge-driven design by grounding scenario generation in retrieved and structured accident summaries and related evidence through an RAG framework. It is simultaneously scenario-driven in that it explicitly targets the generation of context-aware navigational risk scenarios integrating calibrated risk scores with real-time environmental variables. Furthermore, a reasoning-driven formulation, implemented via CoT prompting, facilitates the stepwise integration of quantitative risk indicators and environmental conditions, resulting in interpretable and actionable maritime risk assessments.

2.1. **Prompt engineering.** In this study, we leverage an LLM to systematically summarize maritime accident reports issued by the Korea Maritime Safety Tribunal, extracting critical information and converting it into a structured, JSON-formatted representation. LLMs offer the flexibility to perform a wide range of tasks via prompt engineering, enabling a single model to execute specialized instructions across diverse domains [8]. We begin by defining the essential elements required for an accident summary – such as accident location, vessel details, accident type, and severity level – directly within the prompt. Based on these definitions, the LLM is guided to parse the free-text tribunal verdicts and output only the specified fields in JSON format. By explicitly coupling each desired element with examples in the prompt, we ensure that the model's output conforms to our target schema. However, tribunal reports are written in natural language, often describing facts implicitly (e.g., "5 km east of Ulleung Island") rather than as precise numeric coordinates. Such implicit expressions require background knowledge or geospatial reasoning that the LLM may not reliably perform, leading to missing or erroneous extractions. Conversely,

encouraging the model to infer unstated details can induce hallucinations – fabricated assertions that cannot be verified against the source text [9].

To mitigate these issues, we propose a two-way preprocessing strategy. In Step 1, prompts are crafted to elicit only explicit, verifiable facts; any data requiring inference or external knowledge is intentionally omitted to minimize hallucination. In Step 2, any missing implicit data points may be supplemented via a secondary retrieval process, in which the system re-queries the tribunal text or invokes external knowledge sources (e.g., web search, domain-specific QA modules) to obtain or validate the required information. In practice, this secondary retrieval process integrates multiple external data sources, including maritime accident statistics databases, historical and real-time maritime weather data, and targeted web search results, which are selectively queried to supplement missing contextual attributes such as precise location descriptors, environmental conditions, or regional characteristics.

This combined preprocessing and prompting pipeline is illustrated in Figure 1, which depicts the end-to-end flow from raw verdict ingestion through preprocessing, information extraction, supplementary retrieval, and final JSON output. Figure 2 provides a detailed view of the two-step preprocessing logic, highlighting how external data modules – such as maritime accident statistics repositories, meteorological data services, and web-based information retrieval – are orchestrated alongside the LLM to iteratively fill information gaps and validate extracted facts. By adopting this architecture, our system produces structured, high-fidelity accident summaries that maintain both consistency and trustworthiness, thereby laying a robust foundation for downstream maritime risk analysis.

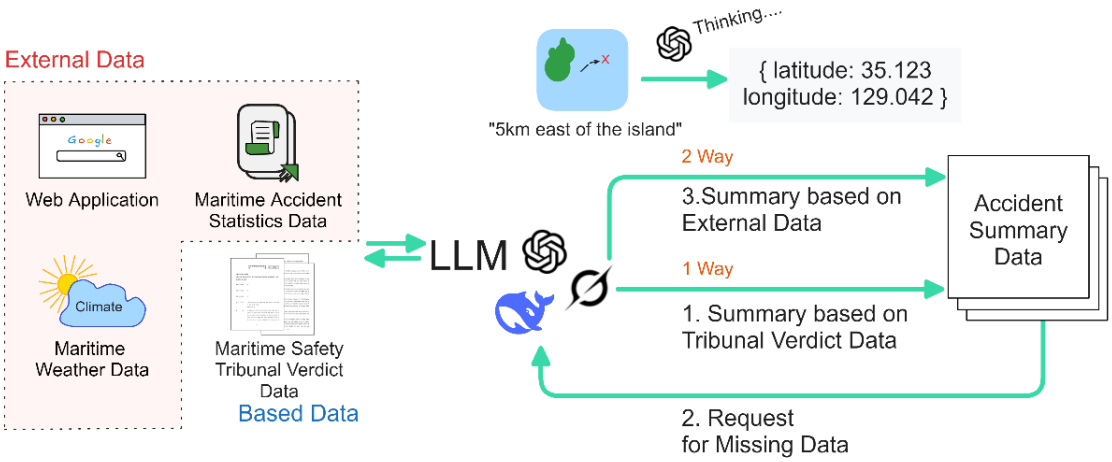

FIGURE 2. Procedure of two-way preprocessing

2.2. **Probability calibration.** Accident risk probabilities are obtained via probability calibration, a requisite procedure that adjusts a model's predicted probabilities, so they align exactly with observed event frequencies, thereby ensuring predictive reliability. In classification settings, calibration is applied whenever the raw probability outputs deviate from true occurrence rates. Unlike conventional approaches that rely solely on binary accident classification, probability calibration enables the model to produce continuous and interpretable risk scores, which are more suitable for risk assessment and decision-making.

In this study, we employ isotonic regression for calibration. Isotonic regression is a non-parametric method that minimizes the squared error between predicted and observed values under a monotonicity constraint. The raw model output probabilities $\tilde{p}$ are adjusted to calibrated probabilities $\tilde{p}_i$ that best approximate the observed outcomes y, as formalized in Equation (1). Here, y_i denotes the observed binary outcome, and $\tilde{p}_i$ represents the

calibrated probability. This procedure removes bias in the probability estimates, ensuring that the predictions accurately reflect the true event frequency. As a result, the calibrated probabilities can be directly interpreted as quantitative accident risk indicators, enabling downstream integration with scenario generation and explainable decision-support modules rather than remaining at a simple accident/non-accident classification level.

$$\min_{\tilde{p}} \sum_{i=1}^{n} (y_i - \tilde{p}_i)^2 \quad \text{subject to} \quad \tilde{p}_1 \leq \tilde{p}_2 \leq \cdots \leq \tilde{p}_n \tag{1}$$

2.3. **Scenario generation.** In the scenario-generation stage, we integrate RAG with CoT prompting to derive immediate navigational risk narratives from calibrated accident-risk probabilities and structured domain knowledge. RAG enables context-aware model adaptation by dynamically retrieving task-relevant information from external knowledge sources at inference time, while CoT prompting elicits stepwise, explicit reasoning that enhances both transparency and prediction accuracy [8].

We construct each LLM input as a JSON object combining real-time dynamic features – seasonality and time-of-day, weather variables (wind speed, precipitation, visibility), and oceanographic conditions (wave height, swell) – with a structured accident summary (tribunal verdict metadata, calibrated probability, and scenario outline). The prompt is augmented with task-relevant cases retrieved from external knowledge bases, grounding the model in the desired input-output structure and depth of inference. Through this retrieval process, historical accident patterns are explicitly incorporated into the reasoning context.

Upon receiving this prompt, the LLM aligns the calibrated risk score with dynamic environmental variables based on the retrieved contextual evidence. It then proceeds through a CoT-style reasoning chain: initially assessing the present risk level based on environmental inputs and subsequently synthesizing this forecast with historical precedent and underlying factors to articulate a comprehensive risk evaluation. This approach transcends mere numeric risk scores by fusing empirical case data with live sensor readings for holistic situational awareness. The final output is structured into three descriptive sections: "Hazard Factors and Issues in the Area", "Predicted Future Accident Type and Likelihood", and "Navigation/Work Precautions". RAG ensures that the generated outputs are consistently grounded in retrieved, verifiable evidence, thereby constraining the LLM's responses to factual accident records and domain knowledge while mitigating hallucinated or unsupported reasoning. In parallel, CoT prompting makes explicit the LLM's multi-step reasoning under complex maritime conditions, thereby improving the practical utility of generated scenarios.

By combining highly calibrated probability estimates from the risk-estimation module with structured accident data from prompt engineering, our prompt design (illustrated in Figure 3) yields advanced navigational risk scenarios that balance interpretability with consistency, delivering actionable insights for maritime decision support.

3. Experiments.

In this section, we present our experimental setup and data descriptions. Additionally, we present the results of generating route-based maritime accident risk scenarios using the proposed framework.

3.1. **Experimental setup.** In this study, we conducted experiments using four distinct datasets encompassing different feature sets and maritime information; the properties of each dataset are summarized in Table 1. The study area was defined as the Exclusive Economic Zone (EEZ) of the Republic of Korea, and all observations were mapped to a uniform level 3 grid (approximately 4.4×5.5 km) – one of the four grid levels specified by the Ministry of Oceans and Fisheries of Korea – based on latitude and longitude. This grid resolution was selected to ensure consistency with officially defined maritime spatial

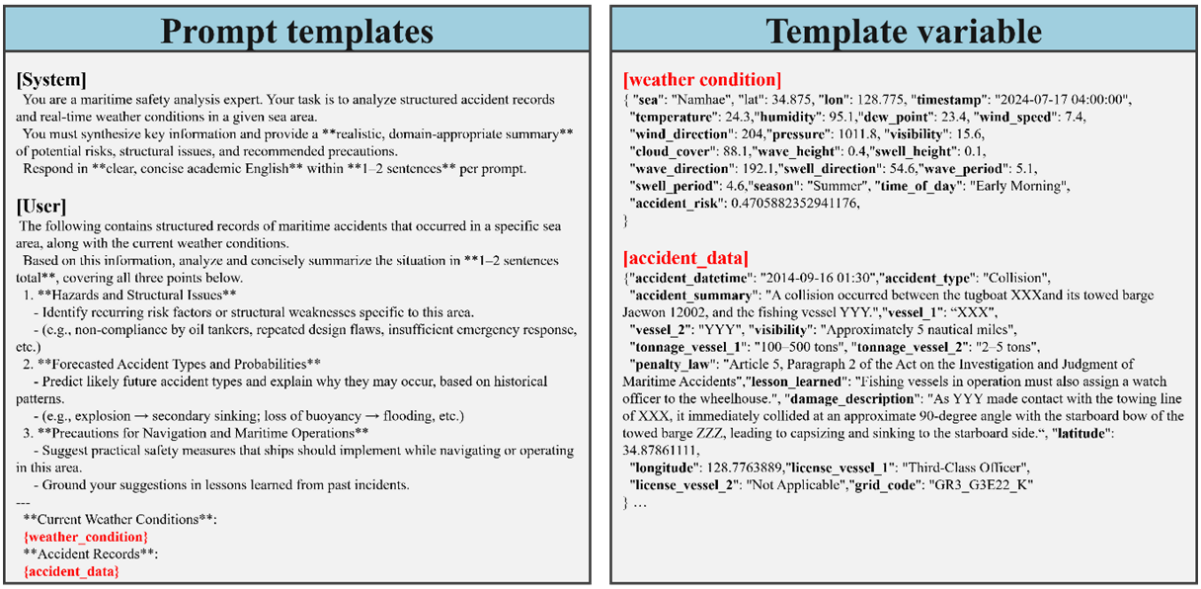

FIGURE 3. Example prompt for scenario generation

TABLE 1. Description of experiment datasets

Dataset	Collected duration	Data format	Frequency	Data source
Maritime accident statistics data	2017.01.01 ∼ 2024.12.31	Tabular	Minutely	Korea Maritime Transportation Safety Authority
Maritime accident judgement reports	1971.01.01 ∼ 2024.12.31	Text	Minutely	Ministry of Oceans and Fisheries
Weather & marine data	2017.01.01 ∼ 2024.12.31	Tabular (Sensor)	Hourly	Visual Crossing
AIS data	2017.01.01 ∼ 2024.12.31	Tabular (Sensor)	Minutely	Global Fishing Watch

standards and to enable reliable matching with traffic density and vessel traffic volume datasets.

The LLM components for both data preprocessing and scenario generation employed the DeepSeek V3 and DeepSeek-R1 models, which were accessed via API calls. All maritime accident risk modeling was implemented using scikit-learn in a Python 3.10 runtime environment.

For accident risk prediction, we employed a stacking ensemble learning framework composed of multiple tree-based classifiers. Specifically, Random Forest, LightGBM, and XGBoost were used as base learners to capture heterogeneous and non-linear relationships among maritime traffic, environmental, and accident-related features. The outputs of these base models were subsequently combined using a logistic regression meta-classifier, which learns an optimal aggregation of base predictions while preserving probabilistic interpretability. The hyperparameters of each base learner were optimized through cross-validation-based tuning on the training set. For Random Forest, the number of trees was increased to ensure ensemble diversity while maintaining stable generalization performance. LightGBM and XGBoost hyperparameters were tuned with respect to the number of estimators, tree depth, and learning rate to balance model capacity and overfitting control.

3.2. **Experiments results.** In this section, we present a probability calibration study to assess the reliability of predicted maritime accident risk probabilities within the proposed framework. Calibration is essential for transforming raw classification scores into meaningful risk probabilities that can support operational decision-making. We evaluate calibration performance using the Brier score, which jointly measures probabilistic accuracy and calibration quality, along with reliability curves that visualize deviations from ideal probability alignment.

To justify the choice of isotonic regression, we conduct a comparative analysis against alternative calibration strategies, including no calibration and Platt scaling. Table 2 reports the Brier scores obtained under different calibration methods for individual tree-based classifiers and ensemble models. Without calibration, all models exhibit relatively higher Brier scores, indicating over- or under-confident probability estimates. Platt scaling improves calibration performance across models; however, isotonic regression consistently achieves lower Brier scores, particularly for ensemble-based predictors.

Table 2. Comparison of probability calibration methods using Brier score

Model	Brier score (No calibration)	Brier score (Platt scaling)	Brier score (Isotonic regression)
Random Forest	0.1523	0.1412	0.1378
LightGBM	0.1551	0.1436	0.1399
XGBoost	0.1544	0.1436	0.1394
Weighted Voting	0.1496	0.1428	0.1369
Stacking	0.1427	0.1318	**0.1256**

Among single models, Random Forest, LightGBM, and XGBoost show comparable calibration performance under isotonic regression, with Brier scores ranging from 0.137 to 0.140. Ensemble-based approaches achieve consistently lower Brier scores, indicating improved probability reliability. In particular, the weighted voting ensemble reduces the Brier score to 0.1369, while the proposed stacking ensemble achieves the lowest Brier score of 0.1256.

These results demonstrate that isotonic regression is well suited for calibrating ensemble-based risk predictors, as it flexibly captures non-linear distortions in predicted probabilities without assuming a parametric form, unlike Platt scaling. When combined with the stacking ensemble, isotonic calibration substantially improves the reliability and trustworthiness of maritime accident risk probabilities, supporting its selection in the proposed framework.

3.3. **Case study: Based on the exclusive economic zone of the Republic of Korea.** In this section, we present a case study demonstrating the application of LURKER in the Korean sea region. We consider hypothetical vessel routes crossing multiple grid cells in the Namhae Sea and East Sea, and analyze the associated navigational risks along each trajectory. Figures 4 and 5 illustrate representative results, including the generated natural-language risk interpretation for a selected grid cell, the vessel route, and the corresponding risk heatmap.

Figure 4 illustrates a case study in the Namhae Sea, where the highlighted grid is located near Jangseungpo, an area with a high incidence of collisions and onboard fires. By integrating current meteorological conditions (e.g., low visibility) with historical accident patterns, LURKER identifies key risk factors such as limited visibility and traffic congestion, and generates actionable recommendations including reduced speed and enhanced radar monitoring.

Figure 5 shows a case study in the East Sea near Gurayongpo Port, a coastal area historically prone to grounding and sinking accidents due to submerged reefs and strong

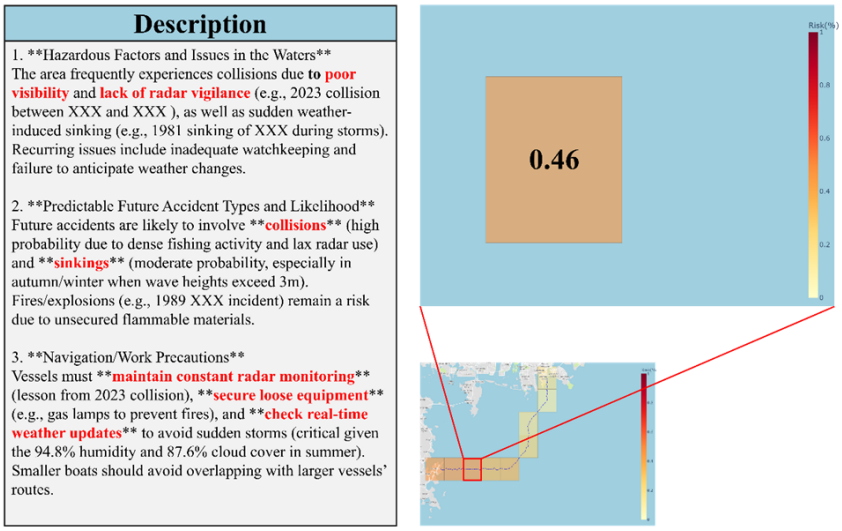

FIGURE 4. Case study results of the LURKER in the Namhae Sea

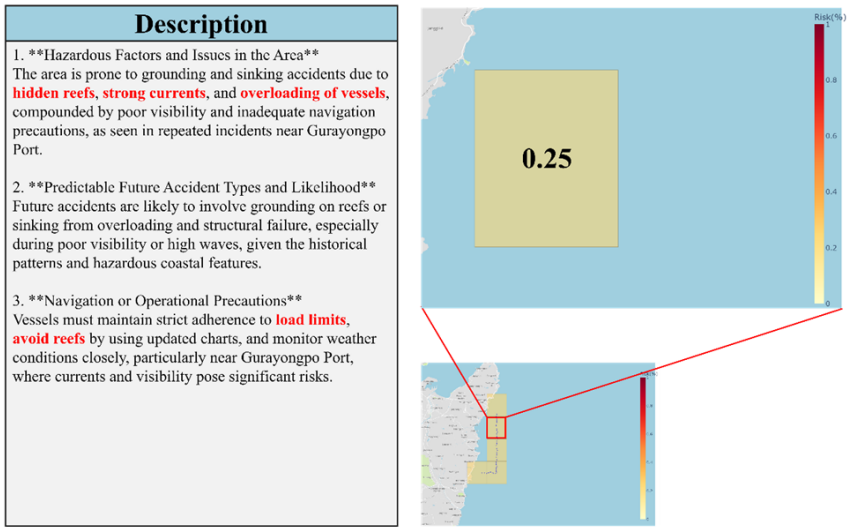

FIGURE 5. Case study results of the LURKER in the East Sea

coastal currents. Based on environmental conditions and spatial accident records, LURKER classifies the grid as a high-risk zone and issues warnings emphasizing updated nautical charts, real-time weather monitoring, and strict adherence to load regulations.

Unlike conventional risk predictors, LURKER goes beyond probability estimation by jointly providing quantitative risk scores, causal explanations, and actionable natural-language guidance within a unified framework, supporting proactive and context-aware navigational decision-making.

3.4. **LLM-as-Judges: Based on the exclusive economic zone of the Republic of Korea.** In this section, we quantitatively evaluate the quality of maritime risk scenarios generated by LURKER using multiple LLM-as-Judge [10] models. Table 3 summarizes the quantitative LLM-as-Judge evaluation results across three judge models (GPT-4.1, Claude-sonnet-4.5, and Gemini-2.5-pro) under different ablation settings. Across all judges, the Full (ours) configuration consistently achieves the highest scores on most evaluation metrics, indicating that integrating environmental conditions, and historical accident records leads to more coherent and operationally realistic maritime risk scenarios.

In contrast, ablated variants exhibit clear performance degradation. The w/o condition setting shows reduced Relevance and Factuality, highlighting the importance of current

Table 3. Quantitative LLM-as-Judge evaluation results

	GPT-4.1				Claude-sonnet-4.5				Gemini-2.5-pro			
	Full (ours)	w/o acc	w/o cond	Base	Full (ours)	w/o acc	w/o cond	Base	Full (ours)	w/o acc	w/o cond	Base
Comprehensiveness	**8.8**	7.9	7.7	7.9	**8.6**	7.5	7	6.8	**8.6**	7.2	6.7	6.8
Knowledge ability	**8.8**	7.8	7.9	7.8	**8.1**	7	6.8	6.3	**8.8**	7.4	7.1	6.8
Correctness	**9.1**	8.2	8	8.1	**8.6**	7.8	7.4	7.1	**9.1**	7.4	8.7	6.4
Relevance	**8.9**	8	7.7	7.8	**9**	8.1	6.4	7.1	**9.7**	9.2	6.5	6.8
Diversity	**8.8**	7.9	7.7	8	**8.1**	7.1	7.4	6.7	**9**	8.1	7.3	7.4
Logical Coherence	**9.2**	8.3	8.3	8.3	**9**	8.1	8	7.7	**9.6**	8.8	9.2	7.9
Factuality	**8.9**	7.9	7.9	7.8	**8.1**	7.2	6.8	6.6	**8.6**	6.7	7.3	6.1
Overall	**8.9**	8	7.9	7.9	**8.5**	7.6	7.1	6.9	**9**	7.9	7.6	6.9

*Full uses all contextual inputs; w/o accident removes historical accident records; w/o condition removes environmental conditions; Base denotes a context-free baseline.

situational information, while the w/o accident setting yields lower Knowledgeability and Logical Coherence due to the absence of empirical accident evidence. The Base configuration records the lowest overall scores, reflecting the limited effectiveness of generic, minimally grounded scenario generation. Despite differences in absolute scoring tendencies among judge models, the relative ranking of all variants remains consistent, demonstrating the robustness of the evaluation results.

4. **Conclusions.** We propose an LLM-based framework that integrates maritime AIS data, meteorological measurements, historical accident statistics, and tribunal verdict summaries to unify accident risk probability estimation with dynamic scenario generation. A probability calibration module – combining tree-based classifiers with isotonic regression – yields well-calibrated risk scores, while prompt designs leveraging RAG and chain-of-thought techniques ensure systematic, stepwise inference during scenario synthesis. By mapping results onto a spatial grid, our approach delivers intuitive, visual representations of area-specific risk levels and narrative scenarios. Crucially, the framework not only highlights key hazard factors and predicts accident likelihood and type, but also generates actionable navigational and operational advisories – thereby functioning as a comprehensive decision support tool for maritime safety management. In our future research, we will augment our framework by integrating maritime traffic and density data, as well as grid-level historical accident-damage metrics, to further enhance the reliability of our risk models. We will also explore the feasibility of a real-time, dynamically updated safe-route recommendation system based on these refined risk assessments.

Acknowledgment. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00218913, 50) (RS-2023-00242528, 50).

REFERENCES

[1] G.-H. Shin and H. Yang, Maritime accident prediction in Busan port using machine learning: An integrated approach with maritime accident reports and VTS data, *Ocean Engineering*, vol.316, 119968, 2025.

[2] A. Shintani et al., Development of marine accident probability prediction model for pleasure boats using ship accident database in central part of Seto Inland Sea, *Ocean Engineering*, vol.322, 120460, 2025.

[3] C.-W. Choe et al., Development of spatial clustering method and probabilistic prediction model for maritime accidents, *Applied Ocean Research*, vol.154, 104317, 2025.

[4] Z. H. Munim et al., Predicting maritime accident risk using automated machine learning, *Reliability Engineering & System Safety*, vol.248, 110148, 2024.

[5] P. Brandt et al., Maritime accident risk prediction integrating weather data using machine learning, *Transportation Research Part D: Transport and Environment*, vol.136, 104388, 2024.
[6] Y. Yang et al., Geographical spatial analysis and risk prediction based on machine learning for maritime traffic accidents: A case study of Fujian sea area, *Ocean Engineering*, vol.266, 113106, 2022.
[7] D. Kim et al., A study on the derivation of sea grid and severity indicators based on maritime accident data, *Journal of Society for e-Business Studies*, vol.29, no.1, pp.57-72, 2024.
[8] P. Lewis et al., Retrieval-augmented generation for knowledge-intensive NLP tasks, *Advances in Neural Information Processing Systems*, vol.33, pp.9459-9474, 2020.
[9] L. Huang et al., A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, *ACM Transactions on Information Systems*, vol.43, no.2, pp.1-55, 2025.
[10] J. Gu et al., A survey on LLM-as-a-Judge, *arXiv Preprint*, arXiv: 2411.15594, 2024.