CyPortQA: Benchmarking Multimodal Large Language Models for Cyclone Preparedness in Port Operation

Chenchen Kuai[1], Chenhao Wu[2], Yang Zhou[1], Bruce Wang[1], Tianbao Yang[1], Zhengzhong Tu[1], Zihao Li[1*], Yunlong Zhang[1*]

[1]Texas A&M University, College Station, USA
[2]University of California, Los Angeles, USA
{mobility, yangzhou295, tianbao-yang, tzz, scottlzh}@tamu.edu,
{chenhaowu}@ucla.edu, {bwang, yzhang}@civil.tamu.edu

Abstract

As tropical cyclones intensify and track forecasts become increasingly uncertain, U.S. ports face heightened supply-chain risk under extreme weather conditions. Port operators need to rapidly synthesize diverse multimodal forecast products, such as probabilistic wind maps, track cones, and official advisories, into clear, actionable guidance as cyclones approach. Multimodal large language models (MLLMs) offer a powerful means to integrate these heterogeneous data sources alongside broader contextual knowledge, yet their accuracy and reliability in the specific context of port cyclone preparedness have not been rigorously evaluated. To fill this gap, we introduce CyPortQA, the first multimodal benchmark tailored to port operations under cyclone threat. CyPortQA assembles 2,917 real-world disruption scenarios from 2015 through 2023, spanning 145 U.S. principal ports and 90 named storms. Each scenario fuses multi-source data (i.e., tropical cyclone products, port operational impact records, and port condition bulletins) and is expanded through an automated pipeline into 117,178 structured question–answer pairs. Using this benchmark, we conduct extensive experiments on diverse MLLMs, including both open-source and proprietary model. MLLMs demonstrate great potential in situation understanding but still face considerable challenges in reasoning tasks, including potential impact estimation and decision reasoning.

Code&Dataset —
https://github.com/ChenchenMobility/MLLM-Bench-CyPortQA

Introduction

Tropical cyclones[1] (TCs) are becoming more intense and increasingly threaten U.S. ports, causing closures, infrastructure damage, and inland transport disruptions that lead to major economic losses (Li, Zhang, and Wang 2023; Verschuur et al. 2023). Over the past few decades, the number of cyclones rated Category 4 and Category 5 [2] has increased, and climate model forecast even stronger storms by the end of the century (Lipari et al. 2024). Meanwhile, increasing uncertainty on cyclone paths makes it more difficult for port operators to adjust operational plans dynamically and implement timely preparedness to reduce potential infrastructure damage and economic loss (Feehan et al. 2024).

Existing port-preparedness tools rely on diverse meteorological data sources to estimate storm impacts on operations and determine port condition bulletins [3], which in turn regulate vessel movements and facility preparations (USCG 2025). Once a tropical cyclone reaches wind-force thresholds at sea, the National Hurricane Center (NHC) begins issuing forecast products at regular intervals, including probabilistic wind-speed maps, twelve-hour wind-forecast tables, narrative advisories on expected impacts, and graphical cones of uncertainty depicting possible storm tracks (NOAA NHC 2025a,b). Based on these products, NOAA (National Oceanic and Atmospheric Administration) and the USCG (U.S. Coast Guard) evaluate potential impacts on port operations. The USCG then applies a rule-based framework in which alerts are triggered when forecast wind speeds exceed predefined thresholds to determine preliminary operational decisions. Final port-condition bulletins are issued in consultation with port authorities. While this protocol provides a structured and actionable approach, it heavily depends on expert judgment. Additionally, other stakeholders, including vessel owners, terminal operators, and cargo interests, typically remain reactive and await official USCG notices before adjusting their operations (Li, Zhang, and Wang 2023; Balakrishnan, Lim, and Zhang 2022).

Several challenges remain in model-driven real-time decision support for port preparedness, including improving interpretation of multifaceted forecasts, enhancing estimation of potential impacts, and ensuring reliability in decision-making (USCG 2025). Addressing these challenges calls for methods that can understand and reason over diverse and evolving data. Addressing these challenges requires models with robust understanding and reasoning abilities to process diverse, evolving data. Recent advances in MLLMs have demonstrated unprecedented capabilities in integrating het-

*Corresponding authors
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]To be consistent, we use the term *tropical cyclones* as the official designation encompassing hurricanes, typhoons, and cyclones.

[2]According to the Saffir–Simpson Hurricane Wind Scale, which classifies cyclones based on sustained wind speeds.

[3]The USCG issues port condition bulletins such as Whiskey, X-Ray, Yankee, and Zulu under 33 CFR § 165.781 (U.S. Government Publishing Office 2025).

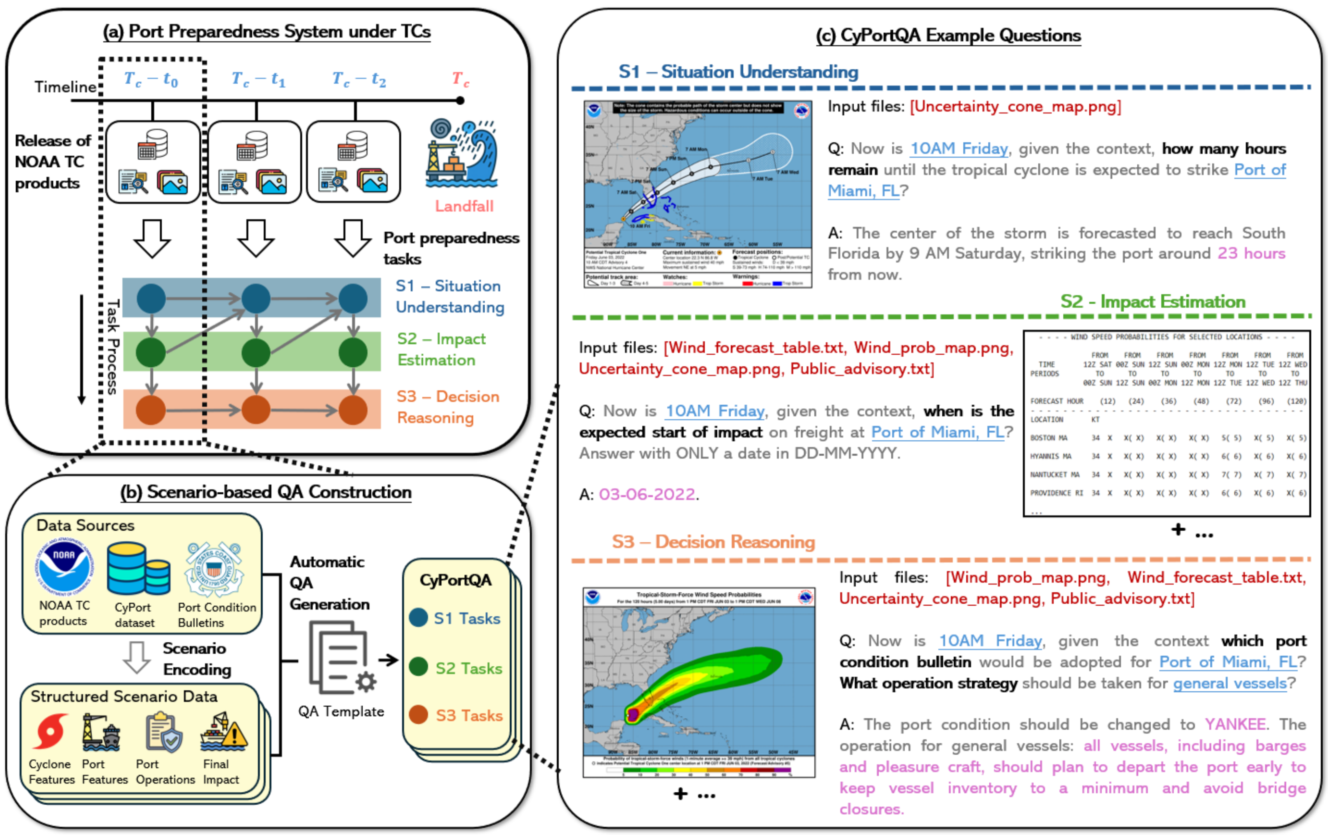

Figure 1: Port Preparedness Framework under Tropical Cyclones and the CyPortQA Benchmark. (a) The time-evolving port preparedness in response to TC, highlighting dynamic decision and key preparedness tasks. (b) Scenario-based QA construction pipeline in CyPortQA, sourcing NOAA TC products, operational performance data and port condition bulletins. (c) Representative CyPortQA examples across three preparedness tasks: S1 – Situation Understanding, S2 – Impact Estimation, and S3 – Decision Reasoning, each with corresponding multimodal inputs and question formats.

erogeneous inputs, such as images (Jiang et al. 2025), structured tables (Sui et al. 2024), and world knowledge (Yu et al. 2023), into coherent contextual outputs. By reasoning jointly across multiple modalities, MLLMs can capture complex relationships, adapt to incomplete or noisy information, and generate actionable insights, making them particularly promising for high-stakes applications in disaster management. To date, no study has systematically integrated dynamic tropical cyclone forecasts with ground-truth port operational impacts and official USCG port-condition bulletins. Without this integration, it is impossible to validate MLLM outputs against real-world outcomes, preventing in-depth performance assessment and slowing the development of automated decision-support tools. Hence, this study aims to address these gaps through the following contributions:

- Developing structured representations of real-world decision scenarios for port preparedness by integrating multisource data (e.g., meteorological forecasts, port operational impact, and USCG operational bulletins).
- Building on these scenarios, we introduce *CyPortQA*, the first multimodal benchmark dataset for port preparedness under cyclones, offering tasks in situational understanding, impact estimation, and decision reasoning.
- Using *CyPortQA*, we benchmark diverse MLLMs including both open-source and proprietary, assessing their performance and uncovering understanding and reasoning gaps in realistic cyclone-related port preparedness.

Related Work

Models for Cyclone Preparedness

Models used to estimate physical damage and disruption to ports caused by tropical cyclones are critical for both pre-event planning and post-event response within the disaster cycle (Cai et al. 2025). The NOAA facilitates cyclone preparedness by publishing a suite of forecast models that provide timely and scientifically grounded predictions of storm tracks and intensities (NOAA 2025a,c; DeMaria et al. 2013; NOAA 2025b; Boussioux et al. 2022). It is equally critical to leverage these public forecast models to inform reliable cyclone impact estimation and support rapid decision making. However, most existing studies focus on long-term planning rather than real-time operational support. For example, Dhanak et al. (2021) developed a simulation-based resilience assessment tool for strategic disaster planning.

Wang et al. (2025b) integrated causal inferences into autoregressive predictions for cyclone intensity forecasting. While these approaches offer important insights, they are not designed for dynamic decision support during active cyclone events. In contrast, Li, Zhang, and Wang (2023) introduced a recommendation algorithm that models cyclones as users and ports as items, using historical interactions to predict the most likely impacted ports. However, this method relies on static hazard indicators such as maximum wind speed and minimum distance to the port, rather than real-time meteorological data, limiting its usefulness for dynamic operational decision making. In many cases, port authorities and terminal operators have to rely on manual interpretation of multimodal weather forecast products rather than integrating them into automated decision support systems (Verschuur, Koks, and Hall 2020).

MLLM in Natural Hazards

Recent advances in MLLMs have demonstrated significant potential for natural hazard assessment by integrating heterogeneous data sources, such as textual reports and imagery, into a unified reasoning framework (Agarwal et al. 2020). Existing work has been increasingly applied MLLMs across three interrelated pillars in natural hazard response: situation understanding (Hughes and Clark 2025; Sun, Wang, and Peng 2023), impact assessment (Sun, Wang, and Peng 2023; Li et al. 2025; Zhou et al. 2021), and decision making (Chen et al. 2024; Yin et al. 2024). Previous studies have focused on leveraging the multimodal reasoning and situational awareness capabilities of MLLMs to address the immediate needs of hazard detection and emergency response under the impact of earthquakes (Ma et al. 2025), hurricanes (Li et al. 2024a; Zhu et al. 2024), and wildfires (Ramesh et al. 2025; Chen et al. 2025). In addition, recent efforts have expanded the scope from a single hazard to multi-hazard environments (Zhou and Liu 2024; Wang et al. 2025a) that support a comprehensive understanding of cascading effects. Despite these capabilities, the use of MLLMs in supporting port preparedness under the threat of tropical cyclones remains largely unexplored.

Methodology

This section presents the construction process of *CyPortQA*. As shown in Figure 1(a), the port-preparedness workflow reacts to real-time multimodal meteorological releases to assess the evolving situation and reason about potential impacts and regulation strategies. Based on this, We sample scenarios at key decision time points, pairing each with its corresponding weather data and operational observations. All inputs are encoded into a structured JSON formate, after which QA templates automatically generate question–answer pairs. This pipeline produces a consistent, scalable benchmark for evaluating multimodal understanding and reasoning in time-critical, high-stake port preparedness contexts.

Data Sources

NOAA Tropical Cyclone Products NOAA tropical cyclone products provide updates every three hours as tropical cyclones move, with time-evolving cyclone information as well as meteorological conditions. As demonstrated in Figure 2, the data is organized in a time sequence following USCG evaluation time and is updated every 12 hours. At each data update interval, we extract four types of forecast products: 1) *5-day Track-Forecast Cone with watch/warning region*, a map showing the projected storm path and coastal alert zones; 2) *Tropical-Storm-Force Wind-Speed Probabilities Map*, showing the spatial likelihood of damaging winds; 3) Plain-text *Tropical Cyclone Public Advisory*, summarizing position, intensity, motion, cyclone watches and warnings; and 4) *Wind Speed Forecast Probabilities Table*, a fixed-width table listing the forecast wind-speed probabilities at selected locations and time bins.

Together, these maps, tables, and advisories represent the core information sources that port operators rely on for real-time situational awareness and decision-making.

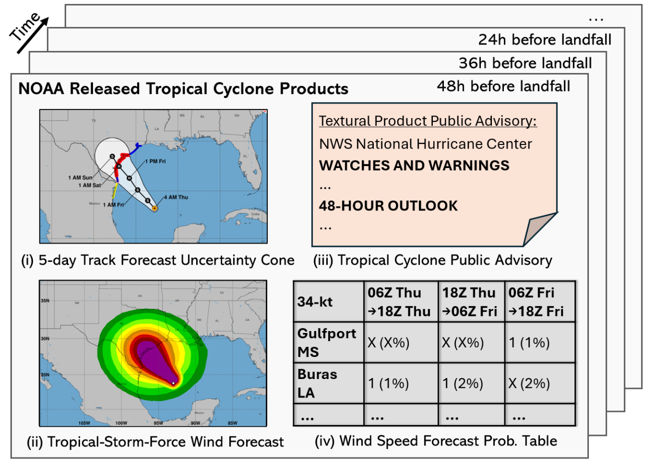

Figure 2: Demonstration of NOAA released tropical cyclone weather products, example data from 2017 Harvey. The data is organized every 12 hours for port operation analysis.

Port Operational Performance CyPort collects 2,197 tropical-cyclone exposures affecting U.S. principal ports from 2015 to 2023 (Kuai et al. 2025). The port operational performance is represented by the commercial-vessel AIS (Automatic Identification System) data captured by onboard navigation-safety devices. For each port's exposure to a tropical cyclone, CyPort documents commercial-vessel activity and impact metrics, including the total reduction in vessel throughput, the onset date of disruption, and the recovery date when normal operations resume. This record offers a fine-grained view of how ports react to tropical cyclones, providing ground-truth data for potential impact estimation tasks.

Port Condition Bulletins As tropical cyclone approaches, U.S. Coast Guard (USCG) districts issue different levels of *Port-Condition* bulletins, *WHISKEY*, *X-RAY*, *YANKEE*, or *ZULU* that prescribe vessel and facility operations. We crawled these bulletins from official USCG X (formerly Twitter) channels (e.g., `@USCG`, `@USCGNortheast`,

`@USCGSoutheast`), parsing for each alert the affected port(s), assigned port condition, release time, and the attached files detailing operational regulations. This information captures the regulatory actions faced by ports during tropical cyclones, serving as a reference for decision reasoning tasks. These data record the regulatory actions taken at time as tropical cyclone approaches.

Together, these three data sources describe real-time port operation scenarios as tropical cyclone approaching by combining real-time forecasts, observed port performance, and regulatory actions. They supply real-world test cases, complete with input meteorological data and verified outputs from ports to rigorously evaluate MLLMs.

Data curation and Scenario Encoding

Tropical cyclone information and port operational performances are sampled every twelve hours, starting 108 hours before the NHC's declared landfall time and ending at landfall (T_c). Each sampling step collects NOAA tropical cyclone products and port condition bulletins, as well as port operational performance (i.e., CyPort dataset). CyPort dataset collects 1,927 cyclone–port exposures from 2015 to 2023. The port impact data is highly imbalanced: only about 21% of these exposures lead to port disruption where freight throughput falls below the historical tenth percentile, and 6% are accompanied by USCG Condition bulletins. To balance the classes, we retain all 410 exposures with impact on ports and select an equal number of exposures from ports within the defined impact buffer that exhibit no significant disruption to balance the classes (Kuai et al. 2025).

To assess ports' real-time preparedness responses, we expand each of the 820 exposures across every 12-hour sampling point with available multimodal weather data, yielding 2,917 scenarios. Each scenario captures a specific cyclone and port at a certain time before landfall, along with weather forecasts, port condition bulletins, and port operational impact.

We convert each scenario's data, including the meteorological data, API-crawled port bulletins, and CyPort data records, into a unified JSON format using ChatGPT o3. Domain experts then independently cross-checked a random 10% sample of scenarios and found strong alignment with the automated outputs, verifying the accuracy and reliability of the structured representations. Port performance impacts are taken directly from the structured CyPort dataset. Each JSON record specifies a cyclone identifier, a port identifier, and the number of hours before landfall, forming the metadata for our QA pipeline, Appendix 1 lists all scenario encoding prompts and JOSN template.

Automatic QA Generation

The automatic QA generation process is grounded in a generalized template framework. For each scenario, we employ a curated set of 48 QA templates, into which scenario metadata can be systematically inserted to generate precise QA pairs. Each template is designed to ensure that the answer can be directly validated from the scenario encoding, thereby preserving label accuracy while evaluating the MLLM's ability to understand and reason over one or more input modalities. The resulting QA pairs collectively span a wide range of tasks aligned with key stages of preparedness: *situation understanding (S1)*, *impact estimation (S2)*, and *decision reasoning (S3)*.

QA generation runs only when a template's required fields are present in a scenario's encoded JSON. Whenever the criteria are met, the system instantiates the QA pair and adds it to CyPortQA. Because port-condition bulletins are not issued in most scenario, S3-type decision reasoning questions receive sparser coverage. In Table 1, CyPortQA ultimately contains 117,178 QA pairs spanning all stages of the cyclone-preparedness workflow. To mitigate potential template bias, each question was paraphrased into several semantically equivalent variants for consistency testing. QA templates and paraphrasing details are provided in Appendix 2, and the design criteria for each stage are presented below.

Situation Understanding (S1) Understanding-oriented QA templates are designed around four key dimensions essential to accurate disaster resilience insights: *spatial awareness*, *temporal understanding*, *exposure interpretation*, and *uncertainty quantification*. These dimensions align with core components of situational awareness during disaster preparedness and response (Lei et al. 2025).

Spatial awareness items test whether an MLLM can reason about geographic relationships from meteorological products, such as whether a port lies within a cyclone's forecast cone or inside a specified wind-probability zone. Because preparedness is time-critical, the QA templates also probe temporal understanding (e.g., hours remaining until landfall). Exposure interpretation templates assess how well the model gauges a port's level and seriousness of exposure. Finally, uncertainty quantification QAs evaluate the model's ability to work with probabilistic information, such as estimating the likelihood that a port will be affected.

Impact Estimation (S2) Impact estimation QA templates are categorized into three distinct subtasks, *time of impact*, *recovery duration*, and *impact severity*, that examine an MLLM to fuse evolving weather data with port attributes and forecast when freight operations will be disrupted, how long recovery will take, and how intense the disruption will be. These ground-truth labels are derived from vessel AIS records within the CyPort dataset, which capture disruptions in freight activity during tropical cyclones. Potential impact estimation in the benchmark verifies whether MLLMs can translate evolving weather into freight impact predictions, which helps commercial vessels reroute proactively and minimize economic loss (Li, Zhang, and Wang 2023).

Decision Reasoning (S3) Decision reasoning is the final task of the benchmark, encompassing two critical tasks: determining the appropriate port condition bulletin to issue, and generating port-level operational strategies. These strategies include restrictions on general vessel traffic, ocean-going commercial vessels, and port facility operations as the cyclone approaches. Since ports serve as critical links in maritime supply chain, post-cyclone decision making requires precise calibration, as both over-reaction and under-reaction can lead to cascading disasters. Therefore, two ad-

Task	Ability	# of QA	Input Modality	Question Type
S1.1	Spatial Awareness	46507	I / T / X	TF / MC / SA
S1.2	Temporal Understanding	40707	I / T / X	TF / MC / SA
S1.3	Exposure Interpretation	31885	I / T / X	TF / MC / SA
S1.4	Uncertainty Quantification	29066	I / T / X	TF / MC / SA
S2.1	Time of Impact	5820	I + T + X	SA
S2.2	Recovery Duration	2910	I + T + X	SA
S2.3	Impact Severity	8730	I + T + X	MC / SA
S3.1	Condition Alert	610	I + T + X	MC
S3.2	Operation Planning	822	I + T + X	SA

Table 1: Task breakdown in the CyPortQA benchmark. Each task targets a specific understanding or reasoning ability and is associated with a set of QA pairs grounded in multimodal inputs. Input modalities include I (Image), T (Table), and X (Text). Question types include TF (True/False), MC (Multiple Choice), and SA (Short Answer). The symbol '/' denotes alternatives (or), while '+' indicates combination (and).

	S1.1				S1.2			S1.3			S1.4			
Model	TF	MC	SA-Num	SA-Desc	TF	MC	SA-Num	MC	SA-Num	SA-Desc	TF	MC	SA-Num	SA-Desc
LLaVA1.6	0.52	0.27	0.11	0.24	0.51	0.27	0.26	0.31	0.30	0.26	0.49	0.26	0.30	0.01
LlaMA3.2	0.60	0.30	0.20	0.30	0.65	0.30	0.22	0.37	0.48	0.30	0.48	0.21	0.48	0.05
Gemma3	0.51	0.29	0.33	0.30	0.49	0.28	0.20	0.32	0.31	0.22	0.47	0.21	0.31	0.05
Qwen2.5-VL	0.48	0.54	0.37	0.29	0.56	0.42	0.24	0.60	0.56	0.36	0.66	0.40	0.38	0.06
Mistral Small	0.72	0.49	0.42	0.44	0.74	0.42	0.27	0.52	0.59	0.37	0.75	0.37	0.59	0.06
ChatGPT4o	**0.81**	0.54	**0.53**	**0.54**	**0.83**	**0.54**	**0.40**	**0.61**	**0.63**	**0.45**	**0.83**	**0.55**	**0.79**	**0.16**
Gemini2.5	0.74	**0.57**	0.40	0.47	0.78	0.51	0.29	0.60	0.59	0.38	0.82	0.41	0.59	0.15

Table 2: Performance comparison of MLLMs on situation understanding tasks across question-type. True/false and multiple-choice items are scored by exact accuracy. Short-answer numeric (SA-Num) items use tolerance-based accuracy. Text-description items (SA-Desc) are graded by an LLM judge on a 0–1 scale (↑).

ditional metrics, over-reaction and under-reaction rates, are incorporated in this stage of evaluation.

Experiments

We conduct experiments on the proposed CyPortQA benchmark to evaluate MLLM performance. Diverse baselines (five open-source and two proprietary models) are tested to ensure comprehensive model coverage.

- `LLaVA-1.6-7B` (Liu et al. 2023)
- `Llama-3.2-Vision-11B` (Dubey et al. 2024)
- `Gemma-3-12B` (Team et al. 2025)
- `Qwen-2.5-VL-7B` (Bai et al. 2025)
- `Mistral-Small-3.1-24B` (Mistral AI Team 2025)
- `ChatGPT-4o` (OpenAI 2025)
- `Gemini-2.5-Flash-Lite` (Comanici et al. 2025)

Questions and Evaluation Metrics

Close-ended questions This question type includes True-or-False and multiple-choice formats, each with a single correct answer. Accuracy is used as the evaluation metric, calculated as the proportion of MLLM responses that exactly match the ground-truth answer.

Open-ended Questions Open-ended questions require short answers, which are either numeric values (i.e., specific numbers, dates, or durations) or descriptive responses (e.g., operational suggestions). To evaluate the numeric outputs, we apply tolerance-based accuracy (Tian et al. 2025), where a prediction is considered correct if it falls within a predefined acceptable range based on domain understanding (e.g., Recovery duration is considered correct when it falls within ± 1 day of the ground truth). For descriptive answers, we adopt Multi-LLMs-as-Judge approach, where four top-ranked LLM judges (Li et al. 2024b) were used: GPT-o3, Gemini-2.5-Pro, Claude-Sonnet-4, and GPT-o4-mini. The final evaluation score of the Multi-LLM judge is computed as the average across all four models. For S3.2 task, the judges are also required to classify responses to one of the following three categories: under-reaction, proportionate reaction, or over-reaction. The final classification is determined by majority vote. Implementation details for LLM judges are provided in Appendix 3.

Evaluation of situation understanding (S1)

The results across all situation understanding dimensions are presented in Table 2. Top-performing MLLMs demonstrate high accuracy, with spatial awareness, temporal understanding, and uncertainty quantification all exceeding 80% accuracy for True/False questions. This highlights the poten-

Model	S2				S3	
	S2.1	S2.2	S2.3 (MC)	S2.3 (SA-Num)	S3.1	S3.2
LLaVA1.6	0.11	0.09	0.26	0.16	0.22 (0.43/0.27)	0.36 (0.63/0.11)
LlaMA3.2	0.30	0.27	0.26	0.24	0.29 (0.61/0.05)	0.41 (0.43/0.17)
Gemma3	0.41	0.28	0.29	0.31	0.10 (0.37/0.10)	0.27 (0.68/0.12)
Qwen2.5-VL	0.17	0.43	0.46	0.44	0.21 (0.62/0.17)	0.40 (0.65/0.11)
Mistral Small	0.12	0.12	0.24	0.28	0.12 (0.09/0.07)	0.05 (0.70/0.10)
ChatGPT4o	0.55	0.70	0.44	0.34	**0.35** (0.37/0.27)	**0.53** (0.62/0.02)
Gemini2.5	**0.69**	**0.72**	**0.48**	**0.44**	0.22 (0.58/0.22)	0.41 (0.67/0.08)

Table 3: Performance comparison of MLLMs on reasoning tasks across question-type. Columns S3.1–S3.2 as decision reasoning questions results are demonstrated as *Acc* (*UR*/*OR*), where *Acc* is mean accuracy or score, *UR* is the under-reaction rate (MLLM recommends a less stringent regulation than required), and *OR* is the over-reaction rate (MLLM recommends a more stringent regulation than required).

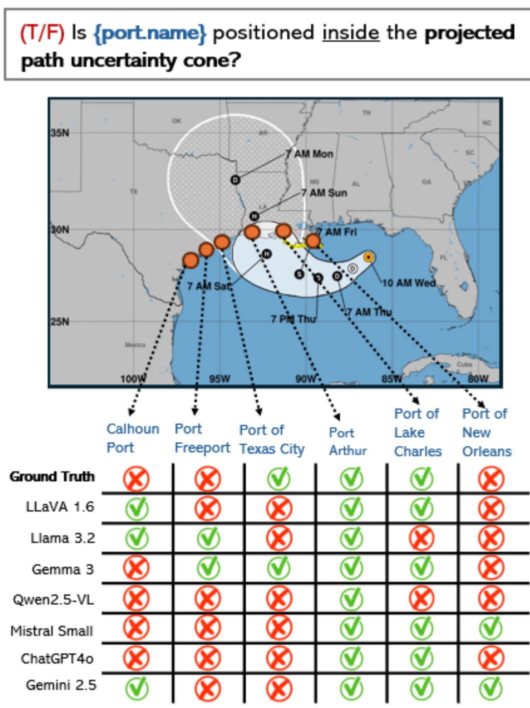

Figure 3: Spatial Awareness & Exposure Interpretation Gaps. A tick indicates a 'yes' response (the port lies within the cyclone's uncertainty cone), while a cross indicates a 'no' response (the port lies outside).

tial ability of MLLMs in understanding the Tropical Cyclone scenarios and extracting key port preparedness information. Descriptive uncertainty questions score poorly, revealing that the models struggle to convert multimodal inputs into precise probabilistic risk statements.

Overall, proprietary models outperform open-source counterparts. ChatGPT-4o achieves the highest accuracy across tasks involving all abilities. Gemini 2.5 Flash-Lite shows competitive performance on spatial understanding multiple-choice questions and performs comparably well. Among open-source MLLMs, Mistral-Small and Qwen-2.5-VL demonstrate stronger performance compared to LLaVA-1.6, LLaMA-3.2, and Gemma-3, particularly in exposure interpretation and uncertainty quantification tasks.

Error Analysis for situation understanding Figure 3 shows that the MLLMs correctly identify Port Author, which is located well within the uncertainty cone, demonstrating their ability to perform coarse-grained spatial localization. However, performance degrades when ports are located on or near the edge of the uncertainty cone. For instance, only Gemma-3 correctly classifies Port of Texas City, underscoring how boundary ports confound other models. Although geographically distant from the cone, Calhoun Port are incorrectly labeled 'within' by three models. Smaller ports may have less geographical knowledge in MLLMs, and spatial-awareness questions involving them are more often misinterpreted. Taken together, these errors indicate that MLLMs still lack the fine-grained spatial understanding in port operations.

Temporal understanding (Figure 4) reveals another gap. Across the selected lead times (12h, 24h, 48h, and 72h to landfall), prediction accuracy tends to improve as the time approaches landfall due to the reduced forecast uncertainty in later stages. However, the smallest deviation from ground truth is reached at 24 h rather than 12h for most models (e.g., ChatGPT-4o, Gemini-2.5, and LLaVA-1.6). These MLLMs tend to predict a time close to a full day (i.e. 24h). This is revealed by earlier findings that MLLMs exhibit a prior distribution or bias which favors certain outputs or decisions (McCoy et al. 2024). This imprecision or "preference" can lead to inaccurate judgments in port preparedness decisions.

Evaluation of reasoning tasks (S2&S3)

Results for reasoning tasks are shown in Table 3. The evaluated MLLMs show clear differences in impact estimation tasks (S2), with Gemini outperforming all other models in multiple-choice accuracy. Furthermore, Qwen2.5-VL demonstrates better performance than ChatGPT-4o and all other open-source models. For S3 tasks, ChatGPT-4o shows a much higher likelihood of issuing the correct bulletin at the appropriate time, whereas its correctly issued bulletin still remains insufficient for real-world operational deployment.

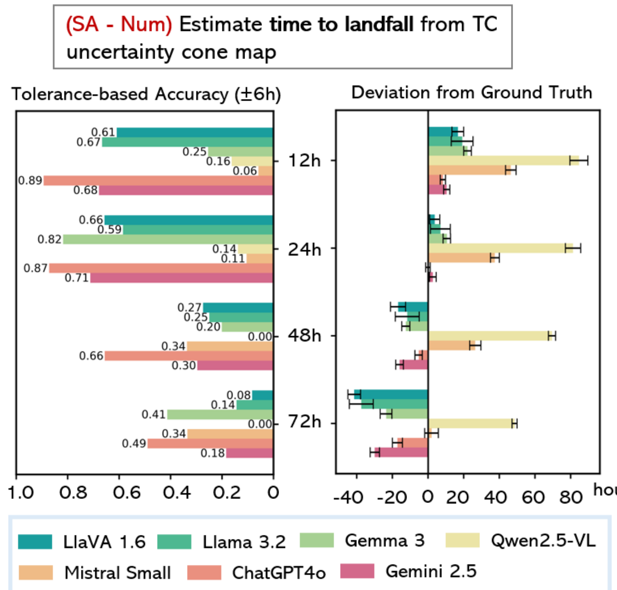

Figure 4: Temporal-understanding gaps. Performance aggregated at 72, 48, 24, and 12 h before landfall. Left panel: tolerance-based accuracy; right panel: mean deviation from ground truth with 95 % confidence intervals.

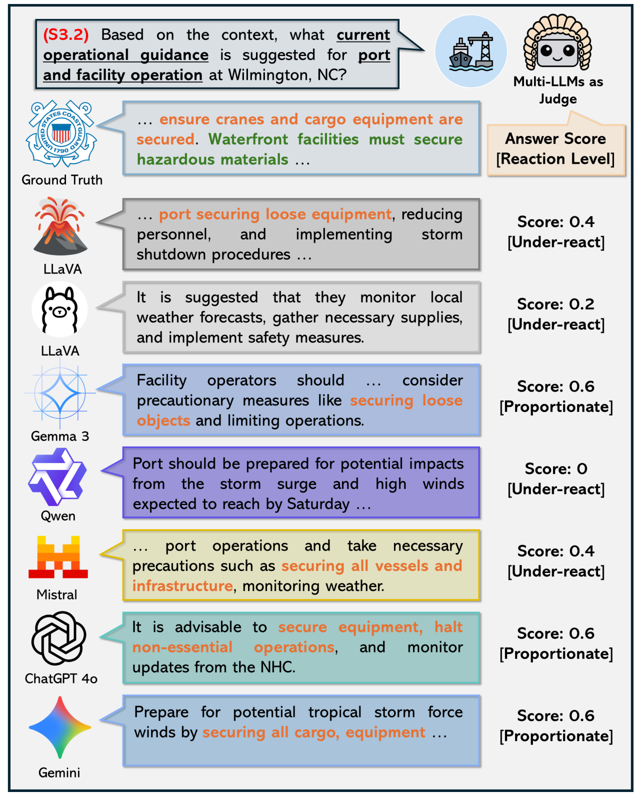

Figure 5: Responses from MLLMs for decision reasoning tasks (port and facility operation instructions) under a single scenario. Evaluation results from the LLM-as-a-judge include a numerical score and classify each response as either an under-reaction, normal reaction, or over-reaction.

In addition, most models tend to underestimate the severity of situations, frequently assigning lower level bulletins when confronted with potential threats. This pattern is particularly evident in Gemini-2.5, LLaMA-3.2, and Qwen-2.5-VL. In practical applications, such under-reactions can lead to delayed or insufficient preparation, potentially exposing vessels, ports, and key infrastructure to unexpected damage.

Error Analysis for decision reasoning Figure 5 presents a scenario that reveals reasoning gaps in MLLM-generated operational responses. In this case, none of the baseline models correctly address the regulatory requirements needed for coordinated port and facility preparedness. In particular, all models fail to mention "securing waterfront facility" that form a key part of port readiness protocols.

The missing portions of the responses reveal insufficient knowledge of waterfront facility regulations, core guidance for port operations. Domain knowledge thus becomes a key bottleneck preventing MLLMs from effectively identifying the specific procedures and interdependent actions. This deficiency further results in vague, high-level responses that generalize or oversimplify operational requirements, making the operations underestimating the urgency of real-world decisions. These results highlight the importance of grounding model outputs in accurate situational understanding and domain-specific knowledge. Effective decision support requires MLLMs that integrate domain expertise to translate weather observations into precise, actionable guidance.

Conclusion

We present CyPortQA, a multimodal benchmark for cyclone preparedness in port operations. Constructed from nine years of real-world disruption scenarios across 145 major U.S. ports and 90 named cyclones, it integrates verified data from NOAA tropical cyclone products, port operational impact records, and USCG port-condition bulletins. Using CyPortQA, we evaluate seven MLLMs (both open-source and proprietary) on three escalating tasks: situation understanding, impact estimation, and decision reasoning. Our experiments indicate that proprietary MLLMs outperform open-source models on both understanding and reasoning tasks. While these models demonstrate strong potential in situation awareness to support port preparedness, they reveal notable limitations in advanced reasoning tasks such as precise impact estimation and actionable decision support. We release CyPortQA to inspire further research on reliable, LLM-assisted emergency decision-support tools that enhance critical infrastructure resilience and operational effectiveness, especially under natural disasters.

Limitations and Future Works: Although CyPortQA provides a comprehensive benchmark, it is constructed from U.S.-based tropical cyclone cases. As a result, its applicability to regions with different cyclone patterns and port operations has yet to be tested. Future work should focus on expanding the dataset to include global ports and diverse hazard events, enabling cross-regional evaluation.

Acknowledgments

We thank Mr. Blakemore of the USCG for his invaluable insights on port disaster preparedness. We also thank undergraduate Brady Foster for data collection and curation. This research is funded by the Center for Freight Transportation for Efficient & Resilient Supply Chain (FERSC) through the U.S. Department of Transportation (69A3552348338). All datasets used in this study (i.e., AIS, NOAA, CyPort, and USCG Port Condition bulletins) are publicly available and redistributable, with no privacy concerns.

References

Agarwal, M.; Leekha, M.; Sawhney, R.; and Shah, R. R. 2020. Crisis-dias: Towards multimodal damage analysis-deployment, challenges and assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 346–353.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Balakrishnan, S.; Lim, T.; and Zhang, Z. 2022. A Methodology for Evaluating the Economic Risks of Hurricane-Related Disruptions to Port Operations. *Transportation Research Part A: Policy and Practice*, 162(4): 58–79.

Boussioux, L.; Zeng, C.; Guénais, T.; and Bertsimas, D. 2022. Hurricane forecasting: A novel multimodal machine learning framework. *Weather and forecasting*, 37(6): 817–831.

Cai, E.; Chen, X.; Keeney, R. G.; Zuckerman, E.; O'Connor, B.; and Grabowicz, P. A. 2025. Identifying and Investigating Global News Coverage of Critical Events Such as Disasters and Terrorist Attacks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 307–323.

Chen, R.; Wang, C.; Sun, Y.; Zhao, X.; and Xu, S. 2025. From perceptions to decisions: Wildfire evacuation decision prediction with behavioral theory-informed llms. *arXiv preprint arXiv:2502.17701*.

Chen, Z.; Shamsabadi, E. A.; Jiang, S.; Shen, L.; and Dias-da Costa, D. 2024. Integration of large vision language models for efficient post-disaster damage assessment and reporting. *arXiv preprint arXiv:2411.01511*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

DeMaria, M.; Knaff, J. A.; Brennan, M. J.; Brown, D.; Knabb, R. D.; DeMaria, R. T.; Schumacher, A.; Lauer, C. A.; Roberts, D. P.; Sampson, C. R.; et al. 2013. Improvements to the operational tropical cyclone wind speed probability model. *Weather and forecasting*, 28(3): 586–602.

Dhanak, M.; Parr, S.; Kaisar, E. I.; Goulianou, P.; Russell, H.; and Kristiansson, F. 2021. Resilience assessment tool for port planning. *Environment and Planning B: Urban Analytics and City Science*, 48(5): 1126–1143.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Feehan, C. J.; Filbee-Dexter, K.; Thomsen, M. S.; Wernberg, T.; and Miles, T. 2024. Ecosystem damage by increasing tropical cyclones. *Communications Earth & Environment*, 5(1): 674.

Hughes, A. L.; and Clark, H. 2025. Seeing the Storm: Leveraging Multimodal LLMs for Disaster Social Media Video Filtering. In *Proceedings of the International ISCRAM Conference*.

Jiang, J.; Ma, C.; Song, X.; Zhang, H.; and Luo, J. 2025. Corvid: Improving Multimodal Large Language Models Towards Chain-of-Thought Reasoning. *arXiv preprint arXiv:2507.07424*.

Kuai, C.; Li, Z.; Zhang, Y.; Wang, X. B.; Lord, D.; and Zhou, Y. 2025. U.S. Port Disruptions under Tropical Cyclones: Resilience Analysis by Harnessing Multiple-Source Dataset. *arXiv preprint*.

Lei, Z.; Dong, Y.; Li, W.; Ding, R.; Wang, Q.; and Li, J. 2025. Harnessing large language models for disaster management: A survey. *arXiv preprint arXiv:2501.06932*.

Li, H.; Yang, Z.; Ma, Y.; Bin, Y.; Yang, Y.; and Chua, T.-S. 2024a. Mm-forecast: A multimodal approach to temporal event forecasting with large language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2776–2785.

Li, L.; Li, D.; Ou, Z.; Xu, X.; Liu, J.; Ma, Z.; Yu, R.; and Deng, M. 2025. LLMs as World Models: Data-Driven and Human-Centered Pre-Event Simulation for Disaster Impact Assessment. *arXiv preprint arXiv:2506.06355*.

Li, T.; Chiang, W.-L.; Frick, E.; Dunlap, L.; Wu, T.; Zhu, B.; González, J. E.; and Stoica, I. 2024b. From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and Bench-Builder Pipeline. *arXiv preprint arXiv:2406.11939*. Submitted 17 Jun 2024; Revised 14 Oct 2024.

Li, Z.; Zhang, Y.; and Wang, B. 2023. Prediction of Port Recovery Time after a Severe Storm. Technical Report Final Report 69A3551747130, U.S. Department of Transportation, Maritime Transportation Research and Education Center, College Station, TX.

Lipari, S.; Balaguru, K.; Rice, J.; Feng, S.; Xu, W.; K. Berg, L.; and Judi, D. 2024. Amplified threat of tropical cyclones to US offshore wind energy in a changing climate. *Communications Earth & Environment*, 5(1): 755.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023. Improved Baselines with Visual Instruction Tuning. *arXiv preprint*.

Ma, Z.; Li, L.; Li, J.; Hua, W.; Liu, J.; Feng, Q.; and Miura, Y. 2025. A Multimodal, Multilingual, and Multidimensional Pipeline for Fine-grained Crowdsourcing Earthquake Damage Evaluation. *arXiv preprint arXiv:2506.03360*.

McCoy, R. T.; Yao, S.; Friedman, D.; and Griffiths, T. L. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41): e2322420121.

Mistral AI Team. 2025. Mistral AI Documentation. https://docs.mistral.ai/. Product documentation; accessed 29 Jul. 2025.

NOAA. 2025a. National Hurricane Center Experimental Tropical Cyclone Forecast Cone Graphic. https://www.nhc.noaa.gov/experimental/cone/. Accessed: June 2025.

NOAA. 2025b. NOAA TC Wind Speed Probability Loops . https://www.weather.gov/mfl/tcwsploop. Accessed: June 2025.

NOAA. 2025c. NOAA text products. https://www.weather.gov/mlb/text. Accessed: June 2025.

NOAA NHC. 2025a. Tropical Cyclone Advisory and Graphics Archive. https://www.nhc.noaa.gov/archive/. Dataset, accessed 29 Jul. 2025.

NOAA NHC. 2025b. Tropical Cyclone Surface Wind-Speed Probability Graphics. https://www.nhc.noaa.gov/aboutnhcgraphics.shtml. Product documentation; accessed 29 Jul. 2025.

OpenAI. 2025. ChatGPT (GPT-4o model). https://chat.openai.com/. [Online; accessed July 2025].

Ramesh, M.; Sun, Z.; Li, Y.; Zhang, L.; Annam, S. K.; Fang, H.; and Tong, D. 2025. Assessing WildfireGPT: a comparative analysis of AI models for quantitative wildfire spread prediction. *Natural Hazards*, 1–14.

Sui, Y.; Zhou, M.; Zhou, M.; Han, S.; and Zhang, D. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 645–654.

Sun, Y.; Wang, C.; and Peng, Y. 2023. Unleashing the potential of large language model: Zero-shot vqa for flood disaster scenario. In *Proceedings of the 4th International Conference on Artificial Intelligence and Computer Engineering*, 368–373.

Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Tian, K.; Mao, J.; Zhang, Y.; Jiang, J.; Zhou, Y.; and Tu, Z. 2025. NuScenes-SpatialQA: A Spatial Understanding and Reasoning Benchmark for Vision-Language Models in Autonomous Driving. *arXiv preprint arXiv:2504.03164*.

U.S. Government Publishing Office. 2025. 33 CFR § 165.781 – Regulated Navigation Area; Coast Guard Sector San Juan Captain of the Port Zone. https://www.ecfr.gov/current/title-33/chapter-I/subchapter-P/part-165/section-165.781. Code of Federal Regulations; accessed 2 Aug. 2025.

USCG. 2025. Severe Weather Contingency Plan for Marine Terminals. Technical report, Maryland Port Administration. Revised May2025.

Verschuur, J.; Koks, E.; and Hall, J. 2020. Port disruptions due to natural disasters: Insights into port and logistics resilience. *Transportation research part D: transport and environment*, 85: 102393.

Verschuur, J.; Koks, E. E.; Li, S.; and Hall, J. W. 2023. Multi-hazard risk to global port infrastructure and resulting trade and logistics losses. *Communications Earth & Environment*, 4(1): 5.

Wang, J.; Xuan, W.; Qi, H.; Liu, Z.; Liu, K.; Wu, Y.; Chen, H.; Song, J.; Xia, J.; Zheng, Z.; et al. 2025a. DisasterM3: A Remote Sensing Vision-Language Dataset for Disaster Damage Assessment and Response. *arXiv preprint arXiv:2505.21089*.

Wang, X.; Chen, K.; Liu, L.; Han, T.; Li, B.; and Bai, L. 2025b. Global tropical cyclone intensity forecasting with multi-modal multi-scale causal autoregressive model. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Yin, K.; Liu, C.; Mostafavi, A.; and Hu, X. 2024. Crisissense-llm: Instruction fine-tuned large language model for multi-label social media text classification in disaster informatics. *arXiv preprint arXiv:2406.15477*.

Yu, J.; Wang, X.; Tu, S.; Cao, S.; Zhang-Li, D.; Lv, X.; Peng, H.; Yao, Z.; Zhang, X.; Li, H.; et al. 2023. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*.

Zhou, Y.; Li, Z.; Meng, Y.; Li, Z.; and Zhong, M. 2021. Analyzing spatio-temporal impacts of extreme rainfall events on metro ridership characteristics. *Physica A: Statistical Mechanics and its Applications*, 577: 126053.

Zhou, Y.; and Liu, P. 2024. Assessing multi-hazards related to tropical cyclones through large language models and geospatial approaches. *Environmental Research Letters*, 19(12): 124069.

Zhu, J.; Dang, P.; Cao, Y.; Lai, J.; Guo, Y.; Wang, P.; and Li, W. 2024. A flood knowledge-constrained large language model interactable with GIS: enhancing public risk perception of floods. *International journal of geographical information science*, 38(4): 603–625.