Safe
Maritime
Transportation
System

SafeMTS Report: Applying Large Language Models to Maritime Near-Miss Safety Data Analysis

U.S. Department of Transportation
Office of the Secretary of Transportation

Bureau of Transportation Statistics

About the Bureau of Transportation Statistics

Leadership

Rolf R. Schmitt, *Deputy Director*, *Office of Director*

Publication Management

Allison Fischman, *Director*, *Office of Safety Data and Analysis*

About This Report

Project Manager

Amanda Lemons

Contributors

Michael Bohlman, Honglei Dai, Curtis Doucette, Allison Fischman, Bahadir Inozu, Benjamin Irwin, Amanda Lemons, Will Nabach, Peter Schaedel, Brian Sumner

Acknowledgements

Special thanks to SafeMTS member companies within the maritime industry for sharing safety data used in this report and providing periodic feedback on results. BTS would also like to thank Todd Ripley, Kevin Kohlmann, and Thane Gilman for sharing their valuable input and perspective.

Report DOI

10.21949/rdm1-cf20

Publication Date

December 2025

Title

SafeMTS Report: Applying Large Language Models to Maritime Near-Miss Safety Data Analysis

Key Words

Maritime safety; near-miss; incident prevention; predictive analytics; large language models

Abstract

This report describes the methods, models and results of the application of large language models to maritime near-miss safety data to increase analytical efficiency and accuracy among diverse industry data within the SafeMTS (Safe Maritime Transportation System) program.

Recommended Citation

United States Department of Transportation, Bureau of Transportation Statistics. *SafeMTS Report: Applying Large Language Models to Maritime Near-Miss Safety Data Analysis*. Washington, DC: 2025. https://doi.org/10.21949/rdm1-cf20.

BTS information service contact information:

Ask-A-Librarian https://transportation.libanswers.com/

Phone 202-366-DATA (3282)

Quality Assurance Statement
The Bureau of Transportation Statistics (BTS) provides high quality information to serve government, industry, and the public in a manner that promotes public understanding. Standards and policies are used to ensure and maximize the quality, objectivity, utility, and integrity of its information. BTS reviews quality issues on a regular basis and adjusts its programs and processes to ensure continuous quality improvement.

Notice
This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The U.S. Government assumes no liability for its contents or use thereof.

Table of Contents

List of Figures

List of Tables

Executive Summary

The SafeMTS (Safe Maritime Transportation System) program, sponsored by the U.S. Department of Transportation's Maritime Administration (MARAD) and operated in collaboration with the Bureau of Transportation Statistics (BTS) and the maritime industry, enables the commercial maritime industry to voluntarily and confidentially share near-miss safety information to identify early warnings, develop preventive measures, and reduce the risk of serious incidents. The confidentiality of data collected by SafeMTS is protected by BTS under the Confidential Information Protection and Statistical Efficiency Act (CIPSEA, 44 U.S.C. 3561−3583). BTS collects and analyzes SafeMTS maritime safety data and shares results in published reports, like this one.

The 2023 program pilot validated the promise of SafeMTS, demonstrating the program's capacity to collect valuable data while maintaining the confidentiality protections promised to industry participants. The pilot report also identified areas for further research and development, including exploration of ways to improve the efficiency of analysis, both to enable companies to operationalize safety learnings faster, and to facilitate growth and scalability of the program.

In 2024, SafeMTS began testing advanced data-science methods, including secure, CIPSEA-protected applications of large language models (LLMs), to increase analysis speed and accuracy without adding reporting burden and to help scale the program. These technologies are seen as tools to amplify human expertise, shifting safety practice from reactive to more predictive, adaptive, and continuously learning systems. This report documents the methods, models, results, accomplishments, and remaining gaps to guide future improvements to SafeMTS.

The primary objective of the project was to determine whether AI-enabled analysis would be effective in combining and analyzing raw data on near-miss events submitted by companies participating in SafeMTS. To accomplish this, the project team trained and applied LLMs with the expertise of maritime subject matter experts to extract SafeMTS field values with high potential learning value, such as causal factors. Overall, the model significantly reduced the time spent reviewing and extracting information from the maritime near-miss reports, with high accuracy. Applying the model to extract discrete causal factors, for example, resulted in at least one causal factor identified for nearly all of the more than 19,000 events in the study dataset, a substantial increase from about 2% of events with discrete causal factors identified from SME review alone. The model was also applied to further evaluate key risk areas such as high potential consequence events.

The report includes potential next steps informed by this effort for improving data quality, accelerating timely insights, and fostering ongoing industry engagement to support SafeMTS' growth and effectiveness. These include continued development of advanced analytical capabilities and enhanced collaboration with maritime industry stakeholders to shape further research.

1. Introduction

The SafeMTS (Safe Maritime Transportation System) program enables the commercial maritime industry to voluntarily and confidentially share near-miss safety information to identify early warnings, develop preventive measures, and reduce the risk of serious incidents. The program aims to fill an industry-identified gap in sharing of information on maritime precursor safety events by securely collecting, analyzing, and sharing learnings from near-miss events across the industry.

SafeMTS is a collaboration between the U.S. Department of Transportation's (DOT) Maritime Administration (MARAD) and Bureau of Transportation Statistics (BTS), in partnership with the maritime industry. MARAD is the sponsoring agency, and BTS collects and protects submitted data under the legal authority of the Confidential Information Protection and Statistical Efficiency Act (CIPSEA). BTS also analyzes the data and shares results in published reports, including ad hoc reports like this one.

In 2023, SafeMTS conducted a pilot phase, which demonstrated the program's capacity to collect valuable data while maintaining the confidentiality protections promised to industry participants. This effort helped define core program elements and effectively implemented core functions. The published pilot report included a data key, initial findings, and the process for adding new company participants. As detailed in the pilot report, the pilot phase also identified challenges to full implementation of the program, including the resource-intensive nature of analyzing this type of information to develop safety insights, identify risks, and allow companies to benchmark their safety performance in a timely way.

In 2024, SafeMTS began an effort to test the feasibility of using more advanced data science techniques—emerging artificial intelligence (AI) methods and large language models (LLMs)—within secure, CIPSEA-protected environments, to address these challenges and improve the efficiency and accuracy of SafeMTS data collection and analysis. This effort is an important step toward increasing value to stakeholders and data providers while not increasing the burden on them to collect and share this data, and toward scaling the program to more participants within the maritime industry.

The rapid evolution of AI technologies—particularly the rise of large language models, autonomous agents, and agentic workflows—is transforming how organizations ensure safety and reliability. These advanced systems can continuously analyze data, detect emerging risks, and suggest or possibly initiate corrective or preventive actions, creating a self-reinforcing cycle of learning and improvement. Organizations can move from reactive to predictive and adaptive safety management. Rather than replacing human expertise, these technologies amplify human awareness, decision-making, and foresight, setting the foundation for a new era of intelligent, resilient, and continuously learning safety systems. This report describes this work, including the methodology and AI models applied, the results and key achievements, and gaps and improvements for future work and full program implementation.

2. Analysis Method, Evaluation, and Results

The SafeMTS pilot highlighted the need for methods that allow for faster, more efficient processing and analysis of disparate maritime near-miss data. Examples of challenges identified in the pilot include inconsistent data definitions and formats across companies, which made it difficult to aggregate and compare information. Many records lacked completeness, with missing discrete data on causal factors, preventive actions, and root causes—key elements needed for meaningful safety learning.

As a result, the SafeMTS team explored the application of currently available AI/LLM tools. Detailed below is the methodology used in applying these tools to the current SafeMTS data submissions, as well as technical details, sample data analysis, and a discussion of significant results and limitations. This AI/LLM approach supports more efficient, valuable safety results that can reach the maritime industry faster.

2.1. METHODOLOGY

The primary objective of the project was to determine whether LLMs would be effective in combining and analyzing raw text data submitted by companies participating in SafeMTS. To accomplish this, the project team defined the following major steps:

- preparing a clean, useable data set, which included integrating data from multiple companies and identifying and removing duplicative data;
- extracting data from text fields to populate missing information; and
- training an AI/LLM using the expertise of maritime subject matter experts (SMEs), by testing various prompt engineering techniques on target fields

The most resource-intensive step in this process was determining if the LLMs could accurately predict structured SafeMTS field values based on free-text narratives and descriptions. For example, this task involved extracting fields such as “Causal/Contributing Factors” and “System/Equipment Involved” from narrative incident reports. The team treated this element as a multi-class classification problem, where the goal was to assign the correct category labels to predefined fields based on the unstructured input.

Traditional classification models typically require large, labeled training datasets to achieve high accuracy. These labels are often generated through manual annotation processes, which can be time-consuming and costly. In contrast, recent advancements in natural language processing (NLP), particularly with the emergence of LLMs, offer promising alternatives. LLMs have demonstrated strong performance in zero-shot learning, where the model performs classification without task-specific training, as well as in few-shot learning scenarios, where limited examples are provided within the prompt. Details of these various methods used on SafeMTS data are covered further in section 2.1.4, Extracting SafeMTS Field Values Using AI Models.

2.1.1. Data Integration and Standardization

The foundational work toward developing an AI-enabled model to analyze SafeMTS data began with consolidating the structurally diverse datasets submitted from participants into a unified

analytical framework. The team developed a centralized data lake[1] to integrate these disparate sources, each of which presented considerable structural variability, with individual files containing anywhere from 16 to 55 fields. While some sources included fewer fields, offering more streamlined data, others presented significantly higher field counts, many of which contained redundant elements, though often with richer data.

To address inconsistencies in field naming conventions across the contributing datasets, a semantic translation framework—internally referred to as the “Rosetta Stone” framework—was developed. This framework applied content-based alignment techniques, as exact field name matches were generally unavailable across sources. For example, the team standardized critical descriptive fields across multiple naming variations such as “description,” “incident description,” “event description,” “summary description,” “near-miss description,” and “details.” This content-driven mapping approach allowed for effective unification of key data elements regardless of the originating structural and terminology differences.

During the integration process, the team assembled datasets with both simple and more complex structures. The translation framework proved highly adaptable at both ends of this spectrum, facilitating scalable and repeatable integration processes for future data contributions. Throughout this process, data integrity was strictly maintained through robust validation procedures, ensuring the fidelity of the consolidated information.

Despite the high level of variation across the submissions, the team successfully created a final integrated data set comprising 103 distinct fields, and including 19,161 individual records, covering 8 companies and a date range of January 1, 2020, through December 31, 2024. This unified data lake was a foundational platform for advanced AI analytics, enabling standardized performance tracking, cross-functional insights, and the development of scalable reporting mechanisms. Overall, the initiative established a robust technical and methodological foundation for enterprise-wide data standardization, significantly improving the consistency, reliability, and usability of organizational data assets.

2.1.2. Data Quality Assurance and Deduplication

As part of the broader data integration effort, the team implemented a dedicated data quality assurance and deduplication process to ensure the accuracy and reliability of the consolidated dataset. Due to the wide variety of systems and processes that companies use to collect data, duplicate near-miss reports could result from such situations as capturing the same incident report across several involved vessels, a lack of unique identifier on individual records, or documenting corrective and other follow-up actions separately from original reports. A structured methodology was established to identify and appropriately handle duplicate records without compromising legitimate data.

The initial phase of duplicate detection employed Excel-based exact text matching focused on the incident description fields. This allowed for a preliminary identification of potential duplicates across sources. However, due to common documentation practices such as copy-paste entry—particularly in standardized reporting formats—the team used a manual evaluation process to

[1] A **data lake** is a repository (often centralized) that stores large volumes of data in their native, raw, or minimally processed form, allowing for multiple data types (structured, semi-structured, unstructured) and deferred transformation (i.e., “schema-on-read” rather than “schema-on-write”).

supplement automated detection. This step was crucial in distinguishing between true duplicates and distinct incidents that happened to share similar phrasing or language.

To confirm a record as a duplicate, three key fields were required to be identical: vessel identifier, date/time, and incident description. Only records meeting all three criteria were separated from the main dataset. Records that contained matching descriptions but differed in vessel or time of occurrence were retained, recognizing them as separate, valid events. This conservative approach ensured that valuable incident data was not mistakenly discarded, thereby preserving analytical completeness. By maintaining both rigor and nuance in record evaluation, the initiative ensured that the final dataset remained clean, contextually rich, and fit for reliable analysis and future decision-making.

In addition to removing duplicate records, three core data issues were identified which complicated the mapping process. Clarifying their nature and remediation tactics was necessary to improve overall data utility:

- **Correct values placed in incorrect fields** – Some records contained valid values but were placed in the wrong fields. These could be corrected or converted relatively easily.
- **Invalid values in incorrect fields** – Certain records contained both invalid values and incorrect field placement, requiring more complex reassignment and mapping.
- **Incorrect or irrelevant entries** – Some data points were meaningless (e.g., blank spaces or noise) and needed to be removed.

This algorithmic triage successfully addressed the complexity of heterogeneous real-world data by automating cleaning, reassignment, and mapping, thereby ensuring a reliable foundation for analysis.

2.1.3. Creation of an Annotated Testing Dataset

To evaluate the performance of the models used in extracting structured information from unstructured incident reports, a manually annotated testing dataset was created. A total of 200 records were randomly selected from the aggregated SafeMTS data lake. Each record was then randomly assigned to SMEs for independent labeling. The SMEs were provided with the unstructured near-miss event descriptions and were tasked with selecting appropriate values for a list of prioritized SafeMTS fields. These selections came from standardized dropdown menus, ensuring consistency across annotations. Importantly, while the SMEs did have access to additional structured metadata, they were asked to focus on the interpretation of the narrative descriptions. The narrative descriptions constitute the richest and most consistent field of data across stakeholders within the SafeMTS program. The resulting SME-labeled dataset served as a “ground truth” reference set for evaluating the accuracy and reliability of the AI-based field extraction methods.

2.1.4. Extracting SafeMTS Field Values Using AI Models

As noted toward the start of the Methodology section above, LLMs offer promising alternatives to using labeled training datasets and show a strong performance in *zero-shot*, *meta prompted,* and *few-shot* learning, achieving accurate classification without extensive task-specific training. These prompt engineering techniques were tested across various leading LLMs, to evaluate their performance on extracting structured SafeMTS field values from unstructured near-miss reports. These models were tested without prior fine-tuning, relying instead on the prompt

engineering techniques to guide their responses. Table 1 provides a summary of the comparative results of different models across target fields. Models were run exclusively in BTS' protected CIPSEA environment.

Table 1. Tracking Performance

LLM	Input Data	Context	Task	Error Rate	At Least One Match
Llama 3.2	Narrative	None given	Causal Factors	10%-25%	46%-53%
Llama 3.2	Narrative	Descriptions	Causal Factors	10%-20%	50%-58%
Llama 3.2	Narrative	Descriptions	Grouped Causal Factors	10%-25%	50%-58%
Llama 3.2	Narrative	Descriptions+ Few shot examples	Grouped Causal Factors	10%-25%	56%-63%
Maverick	Narrative	Task Description	Grouped Causal Factors	0%	69%-79%
Maverick	Narrative	Few shot examples	Grouped Causal Factors, Selected High Potential Near-Misses	0%	78%-92%
Maverick	Narrative + Short Descriptions + Titles	Few shot examples	Grouped Causal Factors, Selected High Potential Near-Misses	0%	83%-94%

SOURCE: U.S. Department of Transportation, Bureau of Transportation Statistics, SafeMTS.

2.1.4.1. Prompt Design

The performance of LLMs in classification tasks is highly dependent on the structure and content of the input prompt. The initial approach employed *zero-shot classification*, in which the LLM was prompted to extract the target SafeMTS field without being provided with any labeled examples. In this setting, the model was given only the event description and the name of the field to be extracted, relying solely on its pre-trained knowledge to infer the appropriate label.

To enhance performance, the team introduced a *meta-prompting* strategy, which involved augmenting the prompt with task descriptions and synthetically generated examples for each valid value in the target SafeMTS fields. These descriptions and examples were constructed with the assistance of a more advanced LLM, enabling the creation of structured guidance to improve the model's interpretability of the classification task. By clearly defining the semantic boundaries of each label, the meta-prompt helped the model produce more accurate and consistent outputs.

Subsequently, *few-shot classification* was tested by incorporating labeled examples from the SME-annotated dataset directly into the prompt. These examples consisted of actual near-miss descriptions paired with their corresponding human-extracted SafeMTS field values. This approach provided the LLM with contextual references, allowing it to generalize better across similar inputs and improving field extraction accuracy in complex or ambiguous cases.

Together, these prompt engineering strategies enabled a systematic evaluation of the effectiveness of the different learning settings (zero-shot, meta-prompted, and few-shots) and in guiding LLMs to extract structured data from the narratives of near-miss reports.

Table 2 provides more details on how the tested prompts evolved when the task given was to extract causal factors from narrative fields, also discussed further in section 2.1.4.2.

Table 2. Breakdown of Various Prompts Tested

Prompt Content	Prompt Description and Content	Result
Acceptable Values	The initial prompt contained a list of acceptable values the model could pull from the event description for a target field (such as causal factors) without further input. An example prompt for extracting the values for the Causal/Contributing Factors field: "Extract the Causal/Contributing Factors from the near-miss description. Acceptable values for causal/contributing factors are: 1) Act of Violence, 2) Carelessness, 3) Complacency/Laziness Near-miss event description: On ATB, the activation/reset switch for the watertight door system on the bridge was accidentally bumped into the closed position while a chair was being moved, causing the three watertight sliding doors in the engineering spaces to auto-close. ..."	When this prompt structure is used with Llama 3.2 to predict the 132 SME labeled records in the initial phase of the project, in 46-53% of the records the model had at least one matching prediction.
Acceptable Value Descriptions	The prompt was next revised to include brief descriptions of the SafeMTS acceptable values with made up examples: Equipment/Material Failure Description: Malfunction or failure of any equipment, machinery, tools, or vessel components. Example: The bilge alarm failed to activate during flooding, despite being tested days prior.	Depending which analysis is performed, there was a 0-10% boost in performance. The most significant impact was on causal factors.
Grouped Acceptable Values	Next, grouping acceptable values was tested to limit the model's confusion. For example, the set of four acceptable values "Carelessness", "Complacency/Laziness", "Distraction", and "Fatigue" were grouped as single value: "Mental Lapse". This reduced the number of acceptable values for causal/contributing factors from 23 to 11.	Depending which analysis is performed, there was a 0-15% boost in performance.
Few-Shot Examples	Next, few-shot learning was tested. Selected records from the dataset were provided as input to the model with their corresponding SME selected values. Example prompt: "Below provided are near-miss descriptions and the corresponding causal/contributing factors. Use these when forming the answer. Near-miss Event Description: "Aboard the MV during general vessel operations, the auxiliary system was believed to have a battery issue so the Second Mate secured the power at the breaker panel to work on the issue. ... Corresponding causal/contributing factors: Equipment/Material Failure"	There were significant performance improvements for all models. When Llama 3.2 was provided with 1, 2, and 8 randomly selected examples for extracting grouped causal/contributing factors, the model obtained at least one matching prediction for 48%, 58% and 69% of records, respectively. When larger models were tested, including Maverick (which has more than x100 the parameters of Llama 3.2), there were significant performance gains. For example, the Maverick model obtained at least one matching causal/contributing factor prediction for 88% of records.
Chain-of-Thought Prompt	The model was directed to provide step-by-step reasoning explaining its choices.	Improved performance on a case-by-case basis. Functional prompt structure was not tested independently.

Prompt Content	Prompt Description and Content	Result
Additional Unstructured Text	Often, companies provide additional unstructured input that provides useful information such as the fields “Short Description” and “Title”. We adjusted the prompt and provided these two fields in addition to the near-miss description.	Improved performance on a case-by-case basis. It was not tracked independently.

SOURCE: U.S. Department of Transportation, Bureau of Transportation Statistics, SafeMTS.

2.1.4.2. Causal Factor Extraction Using Prompts

For initial testing, it was necessary to prioritize the data to be extracted, and it was determined that “causal factors” was a critical field that would provide the most value back to stakeholders. As described above, the team developed and tested several prompts to assess their ability to extract causal factors and initial results achieved only modest accuracy (approximately 20–30%). After review with SMEs, the prompts were refined to include descriptive details such as personnel roles, definitions, and examples, which improved performance to approximately 60%. The next phase introduced curated and random examples to further guide the model. Through iterative testing with different numbers and types of examples, performance advanced significantly, ultimately reaching accuracy levels in the high 80% range for causal factor extraction. One of the sample near-miss event descriptions from the SafeMTS guidance was used as input for testing the extraction of causal factors, as shown in Table 3:

Table 3. Extraction of Causal Factors

SafeMTS Sample Near-Miss Event Description	Extracted Causal Factors
On ATB, the activation/reset switch for **the watertight door system on the bridge was accidentally bumped** into the closed position while a chair was being moved, causing the three watertight sliding doors in the engineering spaces to auto-close. The local alarm and 20 second delay notified personnel in that space of a closing. ***Two electrical extension cords being run through two of the watertight doors were severed.***	The watertight door system switch was accidentally bumped → **Carelessness**
Persons on the bridge were unaware until notified by personnel in affected spaces due to there being no alarm on the bridge panel. The only indication was the small light at each door symbol on the panel. Upon realization of the closure, personnel were notified of the unintended closing of the watertight doors. The activation switch was reset on the bridge and the watertight doors were electrically opened at the local operation switch.	Bridge personnel were unaware because no alarm existed on the bridge panel → **Poor/Insufficient Communications and Poor/Bad Design**
The previous day welding leads had been run through the doors and could have caused a worse incident. A temporary cover was immediately placed over the spring-loaded activation switch as an added security measure to prevent accidental operation by being bumped or brushed up against. A more permanent cover will be fabricated. I recommend we add a new procedure to lock-out the watertight doors from being operated remotely whenever ***cables are routed temporarily through these doors.*** All cables should always be removed whenever work is not being carried out."	Temporary routing of cables through watertight doors without safeguards → **Poor Planning and Failure to Follow Procedure**

SOURCE: U.S. Department of Transportation, Bureau of Transportation Statistics, SafeMTS.

2.2. TECHNICAL PROCESSES AND MODEL EVALUATION

Integrating LLMs into the data ingestion and analysis workflow required key modifications, including the development of a centralized data lake and the reassessment of data quality procedures. Further technical details of the extract-transform-load (ETL) process, and how outputs of the LLM were evaluated for accuracy are discussed in this section.

2.2.1. ETL/Data Lake

The ETL process was constructed to ensure that incoming datasets from multiple companies could be standardized, validated, and integrated into the data lake for analysis in the most efficient, automated manner possible, while still maintaining strict data integrity and protections. The steps below detail further how the data was stored and integrated.

- **Extract:** New data files are received and stored in their original format. Each file is checked against a “Rosetta Stone” mapping to verify consistency of structure and column names. If the structure matches existing records, the file is moved to the next step of processing. If the structure has any variation and does not match previously submitted files, a new mapping entry is created in the Rosetta Stone mapping before processing.
- **Transform:** Using the Rosetta Stone mapping, the dataset columns are aligned and standardized. The process ensures that all fields follow the agreed naming conventions and that no additional or missing columns cause discrepancies. Supporting files are updated to capture metadata and maintain traceability.
- **Load:** Once validated and standardized, the data is appended to the data lake. Duplicates are also identified and reviewed during this step, and version control is maintained through incremented file versions. The data lake is reviewed to confirm that record counts align with expectations and that no duplicates or structural inconsistencies remain.

2.2.2. Accuracy and Evaluation Metrics (LLM Metrics)

Several measures are used to test how often the AI models identify the correct categories from near-miss events and how reliable those predictions are. The metrics below are used to evaluate the performance. These are provided for the general setting where fields can have multiple correct values (e.g., causal factors).

- **Accuracy:** Calculates the percentage of model-produced values that match the values provided by the SMEs. For example, out of all the available answers the AI could have given for causal factors, this is the percent that the model determined correctly when compared to SME determined values. For some fields, the number of correct values is only one (e.g., actual severity level), and for other fields, multiple values could be correct (e.g., causal factors). To be 100% accurate, the model must match the exact number of values as well as the specific values themselves. For fields that only permit one value and have no missing data, this will be the same as Precision and At Least One Match (discussed below).
- **Precision:** Precision tallies the total number of values that matched with SME-selected values. Precision is not affected by how many values the SME provided, only whether a value produced by the model is a match or not.

- **At Least One Match:** Calculates the number of records where the model produced at least one value matching a SME-determined value for that record. If the model predicted one correct value for every record, this measure would have a value of 100%. This is a more forgiving measure used when a record might have several correct answers by simply checking if the model obtained at least one correct value for that record.

In summary, **accuracy** gives the overall success rate, **precision** shows how trustworthy the AI's answers are, and **at least one match** shows whether the AI was at least partly on the right track for each case.

2.3. DATA ANALYSIS AND RESULTS

2.3.1. Value Extraction and Population

For this analysis, eight companies' data were analyzed, which included 19,161 near-miss and hazard recognition events that occurred between January 1, 2020, and December 31, 2024. Table 4 and Table 5 show an example of the progression from each company's raw data submissions and manual SME processing (which occurred during the SafeMTS pilot phase), to advanced AI-driven standardization, for a select set of fields. The evolution demonstrates how SafeMTS moved from fragmented original data toward a more unified dataset capable of supporting broader analysis and correlation studies. It is important to keep in mind that the resulting dataset is still incomplete, but the process undertaken, particularly in regards time-savings, represents the capability of using LLMs to significantly move disparate data towards timely, useful safety learnings.

Table 4 reflects the initial SafeMTS pilot raw data submission completeness, including values extracted by rudimentary scripts and SME manual reviews. Cells in green represent instances where SMEs populated data through manual review during the SafeMTS pilot, up to the number shown in the cell.[2] Non-green cells show the number of values present in original data files. While many fields shown can have multiple values assigned per record/event, only the total number of records that were able to be populated with at least one value in the applicable field is shown in the "Total Values Submitted or Extracted" column.

Importantly, the number in each cell does not reflect the accuracy of the values; to realize maximum learning value, many values would require further standardization (corrected, moved to another field, etc.) beyond what was originally provided or populated by a SME. This does not necessarily mean that original company values or SME inputs were incorrect, but rather reflects the potential for improvement to categorization, preciseness, and consistency during LLM processing. For example, in Table 4, "Task Performed" is an original company field that contained critical data, though the field name itself did not align with the SafeMTS established data dictionary fields. Values from Task Performed were extracted by the LLM to populate other fields, such as "Operations/Activity Ongoing," "Near-Miss Classification," and "Causal Factors," that more accurately reflected the original values. However, not all data contained in Task Performed was harnessed to potentially fill gaps in other data fields. Using a field such as Task Performed again to attempt to populate other fields, such as "System/Equipment Involved" is planned for the next phase of the program.

[2] The number does not represent how many values were originally present or populated specifically by a SME, but rather the total.

Table 4. Dataset Completeness by Company – Pilot Phase

Data Field	C1	C2	C3	C4	C5	C6	C7	C8	Total Values Submitted or Extracted	Percent of 19,161 Events
Vessel Type	11254		5	189	316	576	10	10	12360	64.5%
Near Miss Classification	10729	4491	40	455	316	51	10	10	16102	84.0%
Location on Vessel (High Level)	9106	1859	87	452	316	29	9	10	11868	61.9%
Location on Vessel (Detailed Lvl)		1412		0		53			1465	7.6%
Operations|Activity Ongoing		1		189	316	54	10	10	580	3.0%
Task Performed	8958	1494	39	448					10939	57.1%
System|Equipment Involved	6989		37	187	316	53	2	9	7593	39.6%
Observing Personnel Type			38	188	314	52	10	10	612	3.2%
Factor Preventing Further/Worse incident	244		38	189	316	52	10	10	859	4.5%
Potential Consequence			39	189	316	52	10	10	616	3.2%
Actual Consequence				2	34	12	3	1	52	0.3%
Potential Severity	2073			450					2523	13.2%
Actual Severity				452	316				768	4.0%
Causal | Contributing Factors		10		4	313	44		10	381	2.0%
Root Cause				0		28		8	36	0.2%

SOURCE: U.S. Department of Transportation, Bureau of Transportation Statistics, SafeMTS.

While the first structured view of the dataset was created during the pilot phase, it was a resource-intensive process with potential for efficiency and accuracy gains through the application of LLMs. SMEs have the expertise to deeply understand event scenarios and the factors surrounding them, but an LLM has the additional benefit of speed. Applying LLMs in a targeted way, within a structured process that integrates SME guidance and verification, represents potential for meaningful improvements to learning value and efficiency.

Table 5 below reflects the introduction of LLMs to process and analyze the dataset. Cells in blue represent values that were either newly retrieved from narrative fields or normalized. This step transformed the dataset into a more consistent, comprehensive, and analysis-ready form. Importantly, there were some changes made to the pilot data key to better sort and categorize data. The complete list of changes is contained in Appendix E.

As shown in Table 5, through LLM processing, causal factor values were extracted for nearly every record, resulting in a 98 percentage point increase in completeness. The table also shows that the completeness of certain fields decreased. This is largely due to changes made to the data key during the LLM work to improve consistency and model performance, which resulted in standardizing previously inconsistent values to better reflect the information captured by reporters. Some values previously contained in “Location on Vessel (High Level)” were moved to “Location on Vessel (Detailed Level)”, for example.

As shown in the table, values for certain fields were more readily extracted for some companies than others. For example, no information on the vessel type or system/equipment involved could be extracted for one company, whereas this information was extracted for the other companies. This shows that while an LLM can be successful in extracting certain values, there are still areas where its capabilities may be limited, likely due to the content and richness of narratives provided, on which the LLM is heavily dependent. This is an area where investments in data quality improvements may significantly enhance the dataset and analytical outputs.

Table 5. Dataset Completeness by Company – Post LLM Processing

Data Field	C1	C2	C3	C4	C5	C6	C7	C8	Total Values Standardized	Percent Change
Vessel Type	12358		5	185	316	197	10	9	13080	3.8%
Near Miss Classification	8754	175	21	311	292	43	10	8	9614	-33.9%
Location on Vessel (High Level)	1102	703	11	139	36	37	4	1	2033	-51.3%
Location on Vessel (Detailed Lvl)	8089	1752	81	338	280	46	6	10	10602	47.7%
Operations|Activity Ongoing	7847	720	39	454	310	54	10	10	9444	46.3%
System|Equipment Involved	6305		37	187	284	50	10	10	6883	-3.7%
Observing Personnel Type			38	155	314	52	7	7	573	-0.2%
Factor Preventing Further/Worse incident	244		38	186	316	52	10	10	856	0.0%
Potential Consequence	2673		39	436	316	52	10	10	3536	15.2%
Actual Consequence				4	33	11	3	1	52	0.0%
Potential Severity				450					450	-10.8%
Actual Severity				452	316				768	0.0%
Causal | Contributing Factors	11331	4491	2026	455	315	522	9	10	19159	98.0%
Root Cause				0		3	1	7	11	-0.1%

NOTE: “Percent Change” reflects the percentage point change from the percent of total values submitted or extracted as shown in Table 4.
SOURCE: U.S. Department of Transportation, Bureau of Transportation Statistics, SafeMTS.

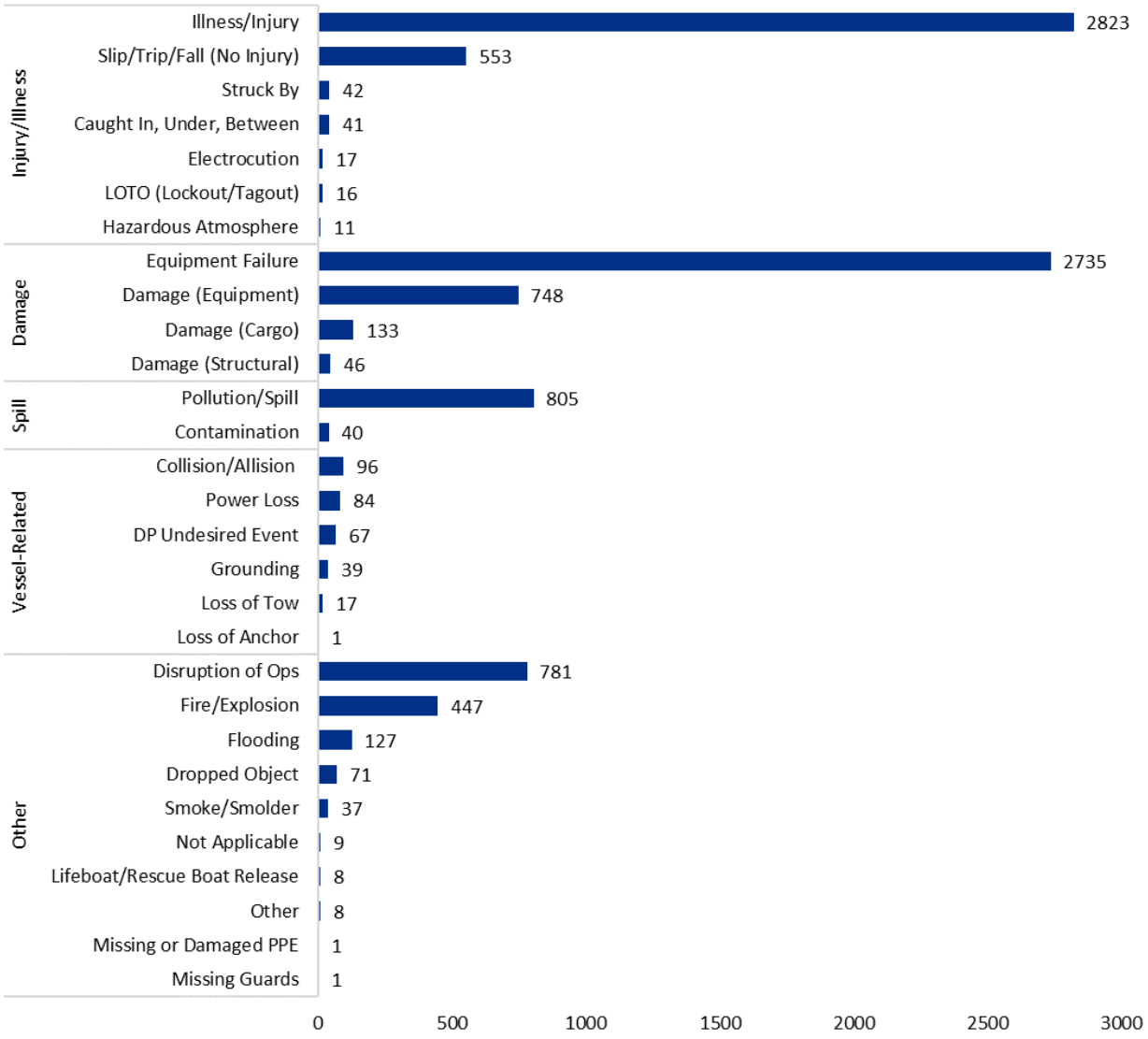
2.3.2. Analysis of LLM Populated Data

The analysis presented in this section reflects the LLM-processed dataset, including the values imputed by the AI model. Importantly, the results, trends, and observations shown are based on only this limited dataset and should not be interpreted as representative of the entire maritime industry. Rather, this section is an illustration of analytics that can be developed when a sufficient sample of data providers fully participate in SafeMTS. These results are exploratory and intended to guide further validation and analysis rather than support firm conclusions.

Just 1.4% of the originally submitted data contained standardized values in the near-miss classification field. While up to 84% of records were able to be additionally populated through SME review during the pilot phase, many of those values needed to be further standardized to realize maximum learning value. Figure 1 below shows that 9,804 near-miss classification values were able to be standardized across approximately half the data set, or 9,614 cases (50.2%), using the LLM processes. For this analysis, an event was able to have more than one near-miss classification. In future phase, further refinements will be developed to select the best value.

Of the 9,804 standardized values, “illness/injury” was the largest category, applied on 29.4% of events that had a value for near-miss classification, and equipment failure the second most frequent, at 28.4% of the same subset of events. These results are similar to the pilot phase data, but as shown in the chart, the categories have been modified slightly based on the newly analyzed cases. For near-miss classification, there were also nine cases where a value of “not applicable” was applied by the model. This means that the original value supplied was not appropriate for near-miss classification but might be able to be evaluated further for populating other, more applicable data fields.

Figure 1. Near-Miss Classification Values Extracted/Standardized by LLMs

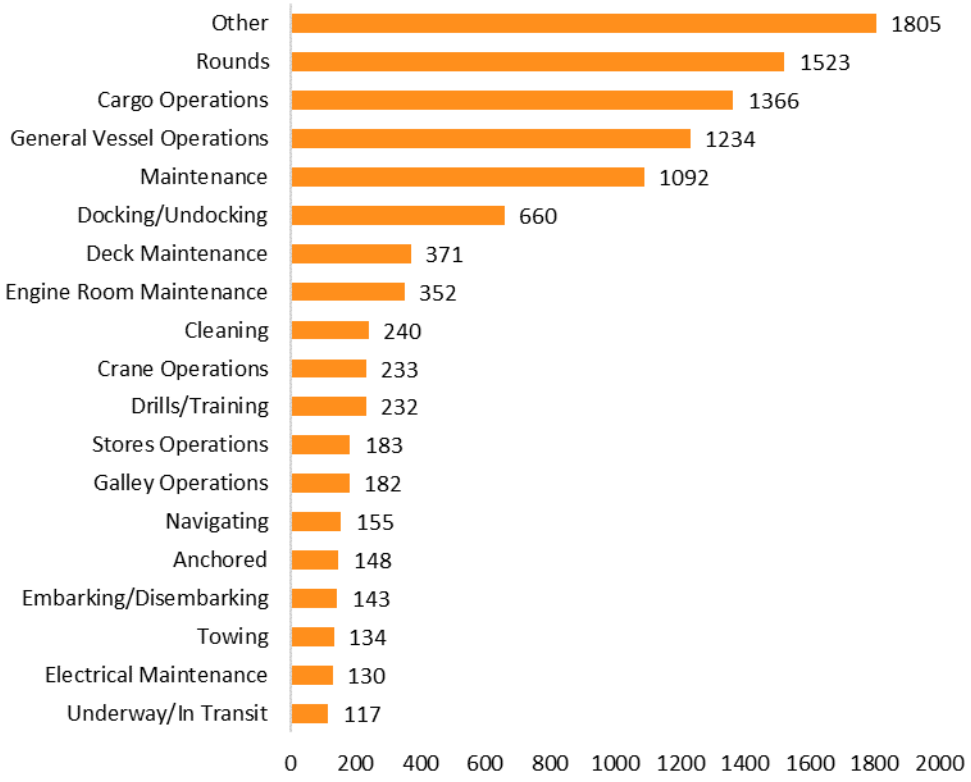


SOURCE: U.S. Department of Transportation, Bureau of Transportation Statistics, SafeMTS.

Just 0.01% of the originally submitted data contained standardized values in the operation/activity ongoing field, with up to 580 values able to be populated through SME review during the pilot phase. Figure 2 below shows that 11,460 operation/activity ongoing values were standardized across 9,444 events, using the LLM processes. For this analysis, an event can have more than one operation/activity ongoing. In total, there were 80 unique operation/activities defined by the model (excluding “not applicable”). Figure 2 includes only those operations/activities values that were applied to at least 1.0% of the 9,444 records. The categories most frequently reported are similar to the pilot data set, with the exception of “Other”, which represents the largest group (19.1%) in this data set, reflecting that many values were moved into the Other category when changes to the data dictionary were made to further group some of the operations/activities. As well, this chart shows that “Rounds” makes up a much larger group than in the pilot data set, and this reflects further refinement of the data key, as many of these cases were previously categorized as “Inspection.” There were six “not

applicable" values assigned to records, though not shown on the chart (0.1% of the 9,444 records).

Figure 2. Operation/Activity Ongoing Extracted/Standardized by LLMs

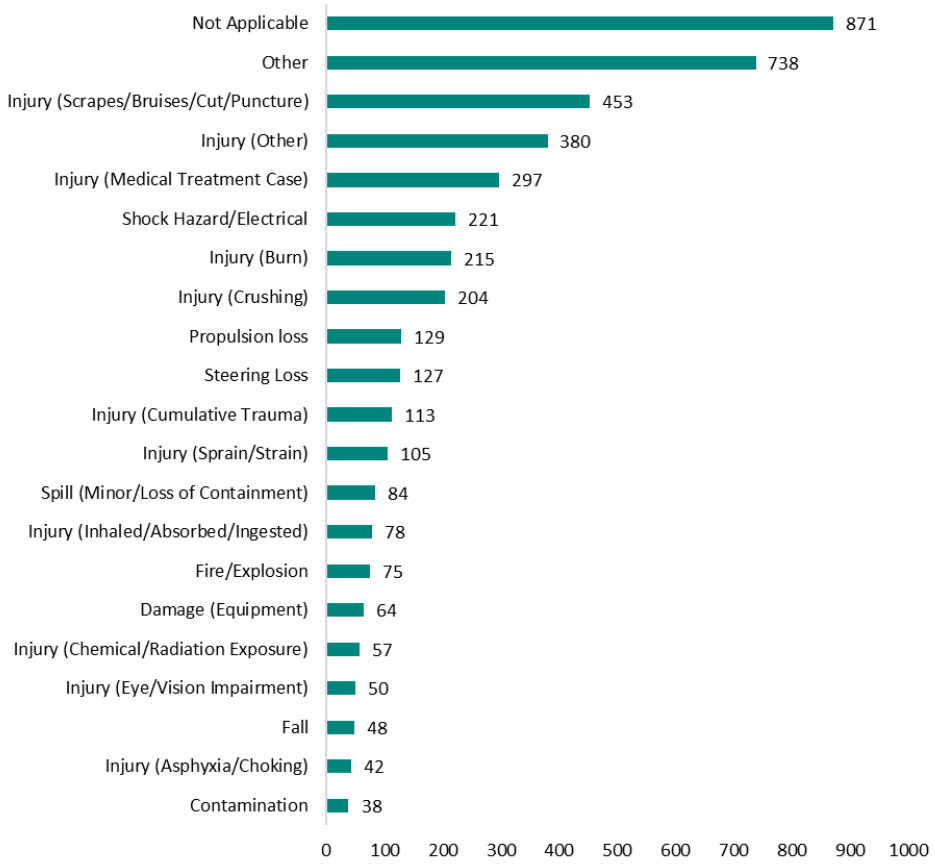


SOURCE: U.S. Department of Transportation, Bureau of Transportation Statistics, SafeMTS.

Of originally submitted data, 0.6% contained standardized values in the potential consequence field, with up to 3.2% (616 values) able to be populated through SME review during the pilot phase. Figure 3 below shows that 4,582 potential consequence values were able to be standardized across 3,536 records (15.2%), using the LLM processes. For this analysis, an event can have more than one potential consequence. In total, there were 30 unique potential consequence values that were defined by the model (excluding "not applicable"). Figure 3 includes only the potential consequence values that were applied to at least 1.0% of the 3,536 records. Though not all types of injuries are shown on the chart, similar to the pilot data set, the combined total of various types of injuries made up the majority of potential consequences identified, applied to 57.2% of records with a potential consequence value.

Similar to the operation/activity ongoing field, "Other" was labeled as a potential consequence in many cases (20.9%), reflecting that many values were moved into the "Other" category when changes to the data dictionary were made to further group potential consequences. "Not applicable" was applied to 24.6% of records that received a value for potential consequence, meaning that the many original values submitted in this field were not appropriate as a potential consequence, and may be used in further research to populate other fields.

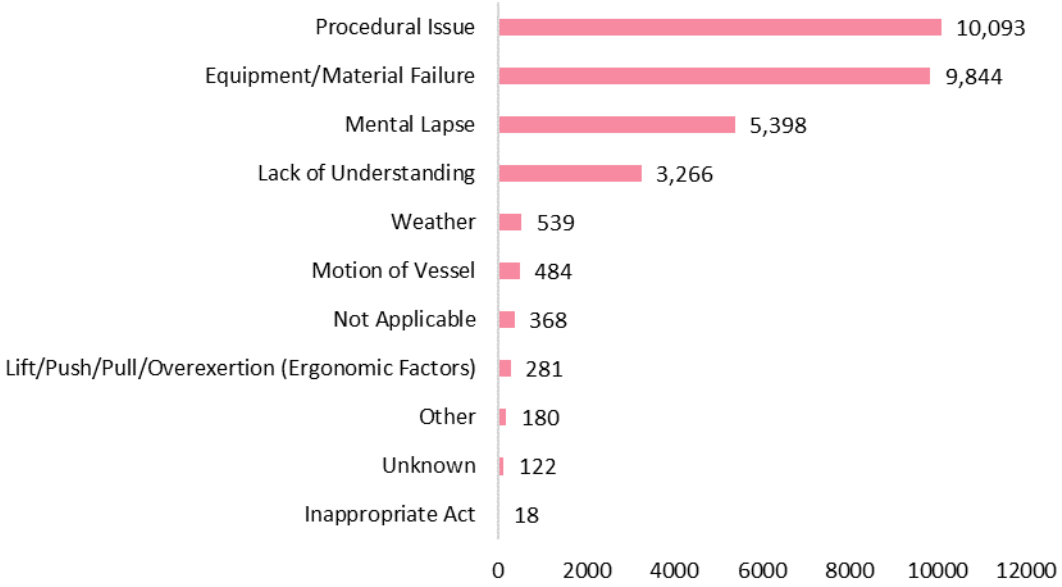
Figure 3. Potential Consequence Extracted/Standardized by LLMs



SOURCE: U.S. Department of Transportation, Bureau of Transportation Statistics, SafeMTS.

Just 0.05% of the originally submitted data contained standardized values for causal factors, with up to 2.0% that were extracted by SMEs during the pilot phase. Figure 4 below shows that 30,593 high level causal factor values were able to be extracted or standardized across almost the entire data set (19,159 events), using the LLM processes. An event can have more than one causal factor; there was an average of 1.2 causal factors identified per event. The figure shows that Procedural Issue was extracted as a causal factor for 10,093 reported events, followed closely by Equipment/Material Failure, listed for 9,844 events. The third and fourth most frequently listed causal factors were related specifically to human error: Mental Lapse applied to 28.2% of reports and Lack of Understanding applied to 17.0% of reports.

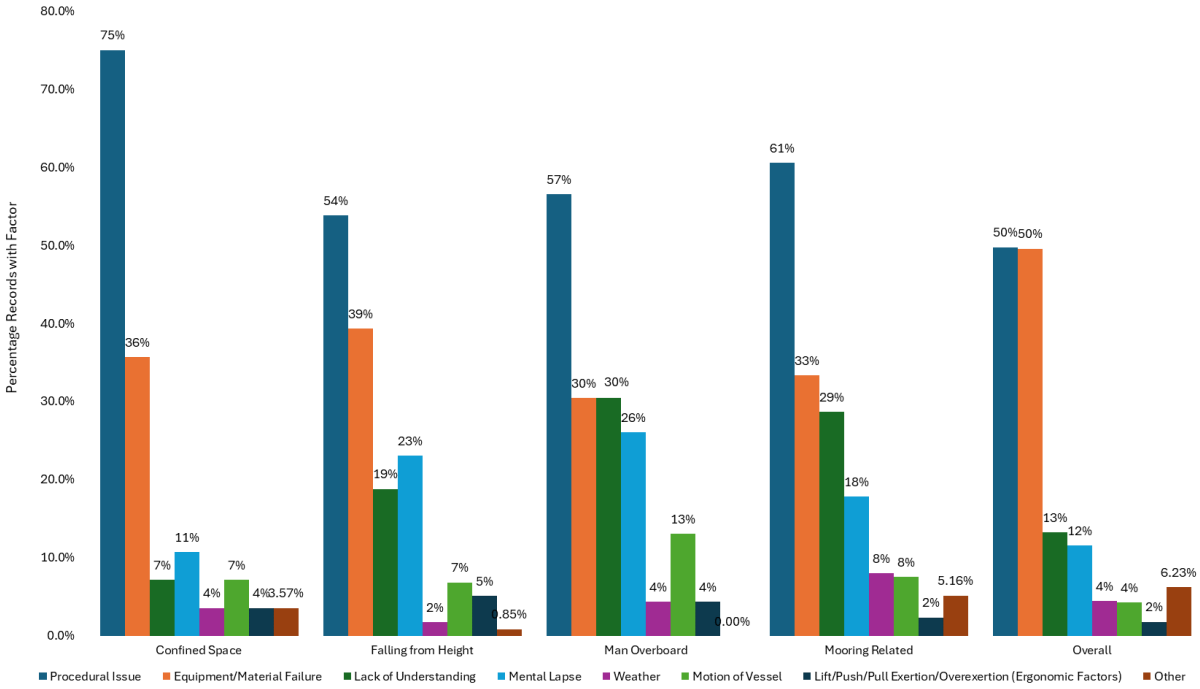
Figure 4. High Level Causal Factors Extracted/Standardized by LLMs



SOURCE: U.S. Department of Transportation, Bureau of Transportation Statistics, SafeMTS.

Four key high-potential (HiPo) event types were initially identified in this analysis as critical to examine: “Confined Space,” “Falling from Height,” “Man Overboard,” and “Mooring Related.” As shown in Figure 5, the causal factors for these types of events follow a similar pattern to causal factors overall, with procedural issue and equipment/material failure being the leading causal factors for each of these event groups. However, compared to the overall dataset, procedural issue makes up a much higher percentage of events than equipment/material failure, and both lack of understanding and mental lapse are associated much more frequently with the three other groups compared to “Confined Space.”

Figure 5. High Potential Events Extracted by LLMs



SOURCE: U.S. Department of Transportation, Bureau of Transportation Statistics, SafeMTS.

2.3.3. Identified Gaps

2.3.3.1. Richness of Narratives

Two critical questions to consider when evaluating the sufficiency of data for producing meaningful outputs include: 1) Is there *enough* data to learn from it?, and 2) Is the data rich enough to address the stated business problem? While AI-enabled approaches show strong potential, their effectiveness remains constrained by data sufficiency and diversity, as current SafeMTS datasets are sector-heavy and unevenly distributed, underscoring the need for broader vessel types, operating contexts, and failure modes to improve model robustness, transferability, and trust. .

While LLMs are extremely useful for extracting values from event descriptions, it is heavily dependent on the richness of the narratives. To improve data outputs, it may be pertinent to improve the depth and breadth of event descriptions and other narrative fields if the SafeMTS program is going to continue to rely almost exclusively on those fields. To this end, outputs from the pilot phase of the program included draft guidance for near-miss narratives. It stated that a complete narrative should include what happened, when and where it occurred, who was involved, and the immediate corrective action taken. Stronger narratives also explain causal factors, planned long-term corrective actions, potential consequences had the event not been intercepted, and any relevant human factors.

The pilot report also encouraged use of 14 framing questions, provided in Appendix B—several informed by Human and Organizational Performance (HOP) principles developed by Todd

Conklin[3]—to prompt richer context. Adoption of HOP-style prompts among participants was very limited in the initial datasets. One approach could be to use the questions as optional aids, rather than prerequisites. SafeMTS is seeking enhancements that better trace the sequence of events and illuminate barrier/defense weaknesses—especially human-factor elements that are increasingly important for preventing similar occurrences.

Given that a substantial proportion of maritime accidents involve human factors, there is increasing emphasis within the industry on understanding how human performance, decision-making, communication, workload, fatigue, training, and organizational context influence safety outcomes. Actively encouraging the capture of human-factors–related details within event narratives—and designing data-collection processes and interfaces that make this information easier to provide—would improve data completeness, analytic consistency, and interpretability. Although some human-factors information may be identified during investigations or follow-up activities, capturing it at or near the time of the event reduces hindsight bias, improves accuracy, and enhances the overall reliability and value of the SafeMTS dataset.

One critical question that has not been regularly included in most submitted data (included in 14% of records) is “What might have occurred had the event not been caught or stopped? What is the worst consequence that could have occurred had one or more of the safety barriers preventing the event failed or not been in place?” While framing questions 11, 12, 13, and 14 may be too involved to answer quickly at the event capture stage for a seafarer, they may be appropriate for corporate level follow-up review, that could be submitted to SafeMTS, increasing the data richness of individual events. To be most effective, such follow-up would need to be near real-time.

For comparison, the Aviation Safety Reporting System (ASRS) encourages reporters to address what they believe caused the problem and how to prevent recurrence, with emphasis on (1) the chain of events—how the problem arose, contributing factors, how it was discovered, and corrective actions taken—and (2) human-performance considerations—perceptions, judgments, decisions, factors affecting performance quality, and actions or inactions.

Other approaches that would improve data richness could be considered, such as traditional accident investigation focuses on *what happened, how it happened, and why it happened*—a linear model aimed at accountability and order. By contrast, leading safety scholars emphasize *human-centered learning, storytelling, and social perspectives on risk*, fostering

[3] The SafeMTS pilot report encouraged the use of 14 framing questions to support event analysis. Several of these questions are informed by Human and Organizational Performance (HOP) principles, including: (8) *What would you change? Which new controls, defenses, or capacity should be added to mitigate potential hazards?* (11) *What happened the way you thought it would happen?* (12) *What surprised you?* (13) *Which hazards did you identify, and which hazards did you miss?* and (14) *Where did you have to “make do,” improvise, or adapt?* Other questions reflect more traditional, control-oriented approaches to safety and risk management and are included to support broader analytical and regulatory needs**.** Collectively, the HOP questions are intended to explore three parts of an event: (1) *The context, that is, everything that took place before the event happened.* (2) *The consequence is what happened, nothing more and nothing less.* (3) *The retrospective way the organization views the event, post-consequence, that is, how the organization simplifies and linearizes the event into an understandable and analyzable problem to be fixed.*

resilience and cultural growth by reframing events beyond blame and attribution (Dekker, Conklin, 2022).[4]

2.3.3.2. Data Quality

For data integrity and understanding, it is important that all acronyms be spelled out on first use or supported by an accompanying key. Inconsistent or undefined acronyms can lead to misinterpretation by analysts, reviewers, and automated analytic tools, particularly when similar abbreviations carry different meanings across organizations, vessel types, or operational contexts. Clear definition of terminology improves shared understanding, supports reliable cross-event comparisons, and reduces ambiguity in downstream analysis. This practice is especially important in LLM-supported workflows, where accurate extraction and classification depend on unambiguous language and consistent terminology.

As well, metadata and information about data collection systems is critical for combining data from varied sources. Without definitions, context, or understanding of how reports are collected and stored, data can be misinterpreted, transformed incorrectly, or its use could be delayed or incomplete. This can be mitigated by establishing a standardized framework for data definitions and collection protocols to ensure consistency across all inputs.

While LLM models will continue to improve and be used in data extraction, SME evaluation will remain critical for data quality. However, SME review throughput (raw near-miss reports per hour) depends heavily on report quality and completeness. For average submissions, one SME could review and correct all 24 fields in the SafeMTS data key at an average of three reports per hour. Because many reports lacked key details, SMEs often had to infer context based on their expertise. With an LLM's pre-processing step that structures reports and populates missing context for SME verification, this throughput could reasonably double to about 6 event reports per hour.

[4] This view is consistent with the Safety-II and Learning from Normal Work literature, which argues that safety is created through everyday human adaptation and system interactions, and that learning is enhanced by understanding normal operations, narratives, and social context rather than relying solely on linear, blame-oriented causal models.

3. Potential Next Steps

Industry stakeholders expressed that the greatest value outputs from the program would be near real-time feedback and key risk area analysis, including HiPo events. Using LLMs to aggregate and analyze SafeMTS data represents an opportunity to realize these outcomes at a greater speed, producing more useful safety information quickly.

The pieces critical to the data pipeline in addition to using LLMs in accomplishing these outputs include data quality and timeliness, as well as a scalable process. Refining and expanding analytical capabilities is one part of the role SafeMTS will take to move in this direction, and working with participants to collect more timely and higher quality data is another. The potential next steps that follow are informed by input from SafeMTS participating companies.

3.1. EXPANDING ANALYTICAL CAPABILITIES

1. **Further Extract Missing Data Values:** LLMs were trained to automatically identify and extract critical missing data elements directly from narrative text. This approach enabled recovery of substantial amounts of information that were not available in discrete data fields in original data submissions. The extracted data were mapped to corresponding SafeMTS fields, ensuring compatibility with the established data schema. The LLMs successfully reconstructed missing fields for many incidents, significantly improving the overall completeness, consistency, and analytical value of the dataset.

   Potential Next Step: Expand validation of the accuracy of LLM-extracted fields for the extended set of data fields through automated cross-checks with SMEs. Enhance the model to include low-frequency and lower-criticality fields and to expand extraction across the extended number of categories.

2. **Additional Accuracy Improvement:** Through iterative improvements in data entry clarification and LLM-enabled analysis, the model was able to extract values for fields with an accuracy that has risen from less than 40% to more than 80%. This increase demonstrates the effectiveness of structured feedback loops and advanced analytical methods.

   Potential Next Step: Incorporate data quality scoring as a formalized benchmark, with goals set at or above 90% record completeness.

3. **Expand AI-Enabled Data Reviews:** AI tools were implemented to automate data classification and initial reviews, reducing manual workload while improving accuracy and preserving security. This freed SMEs to focus on higher-value insights rather than repetitive tasks.

   Potential Next Step: Expanding these capabilities in the next phase might include continuous feedback loops with SMEs for progressive accuracy gains.

4. **Expand Analysis of High-Potential (HiPo) Events:** LLMs have demonstrated the ability to detect HiPo events within incident data. This capability adds significant value by surfacing hidden risks that might otherwise go unnoticed.

Potential Next Step: Build upon this capability in the next phase, extending testing across broader datasets, such as NTSB and USCG accident data, and validating against expert judgment. Several HiPo categories have been suggested by industry members, including:

- Loss of Power
- Loss of Main Propulsion
- Steering Failure (in closed waters or open waters during a high storm)
- Grounding
- Collision/Allision
- Man Overboard, including inland waterways
- Fall from Height
- Death or Injury related to:
  - Confined Space
  - Mooring
  - Dropped object
  - Lifeboat launch/recovery
  - Hazardous atmosphere
  - Large chemical/oil over the side

3.2. WORKING WITH STAKEHOLDERS TO COLLECT MORE TIMELY AND HIGHER QUALITY DATA

1. **Further Refine the Data Taxonomy and Data Dictionary:** The data taxonomy and dictionary were reviewed and refined to support greater clarity and alignment during the LLM phase. This effort strengthened consistency, reduced ambiguity, and enhanced compatibility with algorithmic and AI-driven standardization. It is recommended that continued refinements be coupled with collaborative engagement of companies within Working Groups, encouraging voluntary alignment with recognized standards, such as ASTM F3256, while minimizing reporting burdens. SafeMTS can serve as a testing ground for the data dictionary to be a living document that is repeatedly enhanced with direct input from participants and stakeholders, to drive meaningful safety improvement efforts.

2. **Set Up Industry Working Group:** Incorporating participant and stakeholder input is critical to future development of SafeMTS. One or more working groups may be formed to prioritize topics based on stakeholders' needs as well as to inform other areas of SafeMTS program implementation. Future working group topics include areas of research, including key risk area analysis, ways to move towards real-time feedback, and a phased data quality improvement approach. These are detailed further below.

3.3. FUTURE RESEARCH TOPICS

Future research topics, such as the Success Path method, which provides a structured way to identify and protect the critical barriers that must succeed to keep operations safe, should be considered (Chen et al., 2022). Another topic to explore might be Learning From Normal Work (LFNW), which captures how frontline staff actually get the job done, including the adaptations and constraints they face even when nothing goes wrong (Nazaruk, 2025). In addition, incorporating positive learning data fields and potentially integrating the model with dashboards to enable continuous updates and feedback-driven improvement, is an area for exploration.

Positive learning fields represent the areas and conditions where SafeMTS can proactively learn from successful operations—not just from deviations or failures. These fields highlight how people and systems adapt effectively to variability, maintain safety, and ensure consistent performance. Understanding these patterns strengthens resilience and supports continuous improvement across the maritime safety ecosystem (Nazaruk, 2025). The identification of positive learning fields allows SafeMTS to shift from a reactive to a proactive learning posture, emphasizing how normal operations succeed under real-world pressures. Examples are shown in Table 6.

Table 6. Positive Learning Fields

Positive Learning Field	Description
Routine operations under workload pressure	Examining how personnel maintain safety and throughput despite high demands.
Adaptive responses to changing conditions	Understanding safe and effective adjustments when conditions deviate from plan.
Cross-team coordination	Learning from effective collaboration across departments or organizations.
Consistently successful high-variability tasks	Studying operations with natural complexity that still perform reliably.
Error recovery and resilience in action	Analyzing instances where potential issues were identified and resolved before escalation.
Innovation and local improvement	Capturing grassroots improvements and safe workarounds that enhance system performance.

SOURCE: Adapted from “Learning from Normal Work” by Marcin Nazaruk (Nazaruk, 2025)

By systematically identifying and analyzing these positive learning fields, SafeMTS can codify adaptive expertise, reinforce what works well, and promote a culture that values learning from success as much as from failure. This insight supports the broader goal of developing a high reliability data ecosystem that enhances maritime safety and operational performance.

On the other hand, the Success Path Method takes a unique “positive learning” approach by identifying what needs to go right for success, rather than only focusing on what can go wrong. This shifts the perspective from purely reactive risk management to proactive success planning. In the oil and gas industry, the Success Path evaluation has been applied to well control barriers, equipment configurations, and operational procedures (Cunningham et al., 2022). A potential next step includes further research to apply this approach to SafeMTS data, possibly in integration with LFNW as well as advocated by Shell’s “Fail Safe” approach.

3.3.1. Real-time or Near Real-Time Feedback for Rapid Learning

Across the maritime sector, there is a recognition that paperwork and compliance activities, while necessary, do not by themselves keep ships moving safely. What matters most in practice are the day-to-day actions of crews, the quality of leadership, and the decisions made on deck. Studies highlight that safety performance improves when reporting systems are used not just for compliance, but as real-time feedback loops that guide training, coaching, and problem-solving (Kim & Gausdal, 2020). When approached this way, reporting shifts from a routine administrative burden to a tool that directly strengthens operational reliability and crew safety.

Real-time or near real-time feedback to companies would allow for rapid learning, improved data quality, and effective corrective actions. The advantages of near-real time analysis include the ability to request clarifications while events are fresh in memory, a reduction in the “shelf-life

problem” where the value of data diminishes over time, and greater relevance and applicability of insights for policy, training, and operational change. Timely feedback is seen as critical for training and prevention, whereas delays of weeks or months often mean missed opportunities to act. Being able to quickly clarify missing inputs will be a critical piece in moving towards real-time feedback. This will require more timely data from companies to SafeMTS. A procedure in which SafeMTS can communicate back to participants quickly with any questions about the data could also be considered, subject to available resources.

There are various ways to move towards real-time feedback, such as a data collection application which might include instant data clarification requests or condition monitoring, each detailed in this section.

One example of where real-time feedback could be particularly effective in reducing the time from incident to analysis is the “Equipment/Material Failure” category. While true causes usually require forensic analysis, with access to numerous sources of data, this category can be more immediately actionable with two simple follow-up questions, for example, in a shipboard app, whenever an equipment or component issue is reported:

1. Is there an inspection or maintenance plan for this equipment/component?
2. If yes, was the plan followed?

The answers to these questions provide the immediate next steps that could be taken to address the issue in near real-time:

- If a plan exists and was followed, then review whether the plan needs to be modified.
- If a plan exists but was not followed, then communicate with the relevant parties to determine if there was a misunderstanding, communication error, or other human error that can be corrected.
- If no plan exists, then determine whether one should be created.

These types of immediate question follow-ups in a more automated system are an example of what a near real-time learning loop could look like. This is difficult to achieve in batch-mode processing but straightforward in a real-time or near real-time system.

3.3.1.1. Data Collection Mobile Application

To operate a safe shipping company, crews and organizational leaders must manage their time effectively while dynamically addressing risks and maintaining financial viability. Managers need timely visibility into emerging vulnerabilities so they can act before weak links become failures. Achieving this requires a near-miss reporting process that is both efficient for field personnel and supported by more timely processing and analysis.

A current challenge is that collecting and documenting high-quality near-miss incident information onboard vessels may demand more time than crews can realistically provide. During the SafeMTS pilot phase, a data collection mobile application was considered as a potential next step, along with narrative guidance, a data entry form, and other tools, to facilitate near-

miss data reporting. Similarly, a Ship Operations Cooperative Program (SOCP[5]) survey confirmed industry support for such a tool, noting its potential to resolve the time-effort conflict. Such a tool may be more useful to some companies within the maritime industry than others; for example, it may be more useful to passenger vessel operators or smaller fleet operators than larger companies with established data collection systems. A potential next step in this area for SafeMTS is to explore options with participants for facilitating higher quality data collection.

3.3.1.2. Condition Monitoring

Another option to moving closer to real-time feedback is to fuse condition monitoring with narrative data. Through building a scalable AI model, condition-monitoring streams (e.g., temperatures, pressures, vibration, alarms) can be ingested from critical vessel equipment and be aligned with near-miss/incident narratives. This fusion lets the model detect emerging degradation patterns, flag plausible precursors, and generate ranked, equipment-specific recommendations—without prescribing data-collection logistics in this report. The payoff is actionable signal over noise: earlier warnings, fewer unplanned outages, tighter maintenance windows, including shifting from time-based maintenance approaches to ones based on live data feeds.

There have been recent industry moves in this direction of live-capture data and optical/sensor data, and a shift from paperwork to real work on the vessel. A recent review examines current predictive maintenance (PdM) practices across maritime systems, emphasizing that no single approach fits all applications. While deep learning models outperform traditional methods, their success depends on large, high-quality datasets—currently scarce in the maritime sector. The study highlights challenges with centralized machine learning, including privacy and scalability risks, and identifies federated and transfer learning as promising decentralized alternatives. It also stresses the need for explainable models to meet regulatory and trust requirements. Overall, PdM offers strong potential to extend vessel lifespans and enhance safety (Kalafatelis et al., 2025).

Another example is a case study which used data from the U.S. Coast Guard's Marine Information for Safety and Law Enforcement (MISLE) system, artificial intelligence, and machine learning to predict maritime incidents. It demonstrated how near-miss counts and vessel/waterway type data were used with LLMs to predict serious casualties, achieving high accuracy (92-99%) (Madsen, 2024). This is a topic that a SafeMTS working group could discuss, particularly as it applies to participants' capabilities.

Some example outputs that LLMs could produce as it applies to condition monitoring:

- Degradation scores and anomaly alerts by system/asset (propulsion, pumps, generators).
- "Precursors watchlist" linking sensor anomalies to similar patterns in past narratives.
- Prioritized, condition-based maintenance actions with confidence levels and expected risk-reduction.
- Key Performance Indicator lift estimates (e.g., reduction in downtime, near-miss recurrence, reduction in unnecessary maintenance activities, fuel/consumables).

[5] The Ship Operators Cooperative Program (SOCP) is a member-driven organization of industry leaders to promote and improve the maritime industry through collaboration, facilitation, recommendation, and innovation.

Access to condition monitoring data of critical vessel equipment is helpful for gaining accurate insights into potential weak links that can lead to near-misses or accidents. Such data provides a real-time and historical view of equipment performance, degradation trends, and anomalies that may not be visible through incident reports alone. By integrating condition monitoring information with other data sources, organizations can identify emerging risks earlier, prioritize maintenance efforts more effectively, and develop proactive interventions that enhance safety and reliability across operations.

3.3.1.3. Data Clarification Requests

As mentioned above, a system or procedure that could quickly respond to incomplete reports—prompting the user to provide missing fields—would move SafeMTS closer to near real-time feedback and learnings. If that feedback loop happened within the same day or shortly thereafter, data completeness and usefulness would significantly increase.

This might look like a communication procedure established with SafeMTS and participating companies that is enhanced by AI tools. When used, such tools would output information quickly to a SME or analyst, who can then communicate with a SafeMTS participating company. The challenge here lies in the prioritization of communications across parties; near real-time feedback requires near real-time communication between parties.

In the long-term, this process could be superseded by more advanced automation, such as LLM capabilities built into a shipboard data collection app, for example. This might look like entering a near-miss report or narrative in which the tool then immediately prompts the user for missing or additional details. Such practices are being developed to enhance the efficiency of other safety reporting systems such as the Aviation Safety Reporting System (ASRS) (see, e.g., Ray et al., 2023).

Such a combined approach would give SafeMTS both a top-down view of barrier reliability and a bottom-up view of operational realities. Achieving this level of insight requires a shift away from narrative-only data. Research suggests that relying on narrative-only descriptions is insufficient for pinpointing systemic weak links, even when using frictionless technology like voice-activated apps. By adopting an approach of integrating diverse data sources, the necessary full forensic detail can be captured to move beyond the limitations of verbal accounts.

3.3.2. Phased Data Quality Improvement Approach

Another potential topic for discussion of a SafeMTS working group would be approaching data quality improvement through an incremental approach similar to other industries such as aviation and healthcare. For example, the aviation sector uses a systematic, feedback-driven model for improving data quality, emphasizing the integration of human expertise, automation, and learning from operational data. Within this framework, incremental data quality improvement is achieved through iterative validation, structured feedback loops, and continuous refinement of both data and analytical tools (Ray et al., 2023).

In healthcare, a best practice emerged from the Agency for Healthcare Research and Quality (AHRQ) Near-Miss Reporting and Improvement Tracking project by using incremental data-quality improvement through feedback and simplification. Participating primary care practices enhanced reporting accuracy and completeness by standardizing a short electronic form, providing monthly feedback and reminders, promoting anonymous, non-punitive participation, and linking each near-miss report to follow-up quality-improvement actions. This continuous

cycle—simplify, report, review, and act—gradually improved the reliability, usefulness, and learning value of safety data over time (Crane et al, 2017).

In taking advantage of these best practices, the working group may consider a structured roadmap for improving data quality and recognize that it needs to be prioritized and sustainable, addressing foundational issues first before optimizing analytical and operational performance. A sample phased approach including the critical data quality elements is detailed below. The data quality elements are further covered in Appendix C.

Phase 1 – Foundational Integrity (Critical Priorities)
Objective: Build a trustworthy data foundation by prioritizing completeness, followed by accuracy, consistency, and format.

- **Completeness:** Ensure all critical fields (incident type, causal/contributing factors, incident type, date/time, location, personnel, and operational conditions, personnel role, operational conditions) are present. Missing or incomplete data can require extensive SME time and prevents valid analysis. Some key data may exist only in logbooks, PDFs, or other systems, should be integrated incrementally using AI-assisted extraction tools within available resources. Complete data enables meaningful analysis and risk detection.
- **Accuracy**: Remove typographical and entry errors, duplicates, and misclassifications to ensure factual precision.
- **Consistency:** Standardize data structures, taxonomies, and definitions across systems (e.g., vessel IDs, equipment names, time formats). Recognize that complete data in special formats (e.g., unstructured PDFs) adds friction. Converting to usable structured formats should follow once completeness is achieved.

The outcome of these is a reliable baseline dataset that stakeholders can trust for baseline operational insights and decision-making.

Phase 2 – Analytical Maturity (Enhanced Detail and Timeliness)
Objective: Enrich the dataset to allow deeper, more actionable analysis while ensuring data reflects real-time operations.

- **Granularity:** Add more detailed contextual data (e.g., sea state, weather, experience level, fatigue indicators).
- **Reliability:** Remove ambiguity by defining data elements precisely and ensuring they align with operational realities.
- **Timeliness/Freshness:** Shorten the lag between data collection and dashboard availability to enable near real-time anomaly detection and risk response.

The outcome of the additional elements is a dynamic, context-rich dataset supporting advanced analytics and predictive modeling for proactive risk management.

Phase 3 – Operational Excellence (Sustainability and Usability)
Objective**:** Institutionalize data governance and usability practices that sustain long-term improvement and user engagement.

- **Relevancy:** Continuously verify that every data element serves a defined operational or learning purpose.
- **Availability/Accessibility:** Deliver the right data to the right people rapidly through secure dashboards and APIs, shifting from data retrieval to automatic delivery of actionable insights.

- **Usability:** Simplify column names and formats to make datasets intuitive for analysts and end-users.
- **Metadata:** Treat metadata as a continuous-improvement function. As new ship types, unmanned vessels, or operational activities emerge, update definitions, structures, and codes to maintain consistency and usability.

The outcome of these additional data elements is a sustainable data ecosystem where users can easily find, understand, and trust the data, enabling ongoing performance improvement.

The key benefits of this phased approach include:

- Incremental Progress: Focuses on achievable improvements that build upon one another.
- Cross-Functional Engagement: Encourages collaboration between data stewards, analysts, and operational teams.
- High Return on Investment: Early improvements (completeness, accuracy, consistency) yield immediate gains in analytical reliability.
- Long-Term Sustainability: Embeds data governance and literacy as part of the organization's culture.

Bias is an additional quality dimension to be considered. For example, established NMAC (near-midair collision) safety-reporting systems are voluntary and based on subjective reports from pilots, mechanics, controllers etc. Because of that, the data are subject to reporting bias—i.e., not all events are reported, and the subset that is reported may reflect factors like awareness of the system, willingness to report, perceived severity, and context (Dy and Mott, 2024). Addressing bias might look like establishing fairness and inclusivity as governance principles. Auditing datasets for systemic bias or exclusion; ensuring ethical sourcing, anonymization, and transparency in LLM training data.

Appendix A. Current Data Key Update

Detailed testing and review necessitated changes to the SafeMTS data key to increase LLM accuracy for producing meaningful output. For the System/Equipment Involved field, some values were edited, new ones were added, and then all values were further grouped to add an additional level/field. Below is a summary list of the new grouped categories, as well as a more detailed label with all listed changes. Additional modifications are suggested to be added as part of the phased data enhancement that would be a working group topic/work plan.

- Cargo Handling Systems
- Deck Machinery
- Electrical Systems
- Emergency & Safety Equipment
- Fuel Systems
- Hull & Structure
- Living Quarters / Hotel Services
- Lubrication Systems
- Miscellaneous
- Navigation and Communication
- Off the ship
- Propulsion and Maneuvering Systems
- Sea Water Cooling Systems
- Ships Stores
- Tools/Scientific Equipment

Table 7. New System/Equipment Field Values

(NEW) High Level System/Equipment Involved	System/Equipment Involved (Edits shown in red)	Change Detail
Propulsion and Maneuvering Systems	Main Propulsion	
	Main Reduction Gear	
	Propeller, Shafting, & Stern Tube	EDIT
	Steering Systems	
	Helm	
	Bow/Stern Thruster	NEW
	Main Boiler	NEW
	Dynamic Positioning System	NEW
	Thrust Bearing	NEW
	Hydraulic Clutch	NEW
	Transmission	NEW
	Intercon Coupler	
Fuel Systems	Fuel Tank	
	Fuel Purification	
	Fuel Service	

(NEW) High Level System/Equipment Involved	System/Equipment Involved (Edits shown in red)	Change Detail
	Bunkering/Fuel Loading Systems (non-cargo related)	NEW
Lubrication Systems	Lube Oil	
	Lube Oil Purification	
	Thermal Oil	
	Lube Oil Storage	NEW
Sea Water Cooling Systems	Sea Water Cooling Systems	EDIT
	Sea Water Heat Exchangers	NEW
	Main & Aux Condensers	NEW
	Sea Water Treatment/Marine Growth Prevention Systems	NEW
Utility Systems	Auxiliary Engine	
	Aux Steam Systems	EDIT
	Hydraulic System	
	Compressed Air	
	Heating/Ventilation & Air Conditioning Systems (HVAC)	EDIT
	Auxiliary Steam Boilers	NEW
	Potable/Distilled Water Systems	NEW
	Sewage Collection & Treatment Sys	NEW
	Ballast Systems	EDIT
	Oily Water Separator (OWS)	
	Heat Exchangers - Non Seawater	NEW
	Bilge Water Systems	
	Cargo Gear	
	Cargo Pump	
	Cargo Tank/Hold	
	Cargo Piping, Hoses & Valves	
	Stripping and Transfer	

(NEW) High Level System/Equipment Involved	System/Equipment Involved (Edits shown in red)	Change Detail
Cargo Handling Systems	Vapor Recovery	
	Refrigeration Systems (for cargo)	NEW
	Fork Lift	
	Cargo Being Carried	
	Inert Gas Generator Systems	EDIT
	Reefer Container	
	Refrigerated Cargo Holds	
	Hatch (Cargo Access)	
	Cranes/Davits/Lifting apparatus (cargo related)	NEW
Living Quarters / Hotel Services	Galley related equipment	NEW
	Refrigeration boxes & all equipment related to ships stores	NEW
	Clothes Washer/Dryer	
	Comminuter-Food Grinder	
	Freezer/Chill Box/Storeroom	
	Garbage	
	Hospital	
	Incinerator	
Electrical Systems	Electrical Distribution Systems Medium/High Voltage (>= 440 VAC)	
	Electrical Distribution Systems Low Voltage (< 440 VAC)	
	Centralized Control Systems (Low Voltage)	
	Switchboards	
	Electrical/Motor Control Devices	
	Electrical Devices/Appliances	
	Electrical Generators	
	Lighting/Electrical Fixtures	
Navigation and Communication	Communication Equipment	
	Navigation Systems	

(NEW) High Level System/Equipment Involved	System/Equipment Involved (Edits shown in red)	Change Detail
	Navigation Lights	
	GMDSS	
	Documentation	
Tools/Scientific Equipment	Cargo Stowage System/Stability Mgt System (Software)	EDIT
	Test Equipment	
	Research Equipment	
	ROV	
	Diving Equipment	
	Power Tools/Tools	
	Vendor/Customer Equipment	
	Welding/Hot Work Equipment	
Hull & Structure	Gangway/Accommodation Ladder	EDIT
	Ladder/Stairs (everywhere on ship)	
	Landing Chair	
	Door/Hatch (People or Equipment Access - Not associated with cargo spaces)	
	Railing	
	Ramp	
	Stern Ramp	
	Cranes/Davits/Lifting apparatus (non-cargo related)	
	Hold	
	Mast/Rigging	
	Void Space	
	Deck Plating	
Deck Machinery	Stern Ramp	
	Pilot Ladder	
	Anchoring Machinery	
	Mooring Equipment	
	Towing Equipment	
	Gangway Winch	
Emergency & Safety Equipment	Lifeboat/Rescue Boat	
	Other Lifesaving Equipment	
	Fire Fighting Equipment	EDIT
	PPE	

(NEW) High Level System/Equipment Involved	System/Equipment Involved (Edits shown in red)	Change Detail
Ships Stores	Cleaning fluids/materials	
	Chemicals	
	Provisions	
	Spare Parts	
	Ship Paints	
	Specialty Chemicals F/Maintenance	NEW
Off the ship	Aids to Navigation	
	Dockside	
	Dock Equipment	
	Other Vessel	
Miscellaneous	Not reported/Not Identified	
	Other (Not In the list)	

Appendix B. LLM Technical Details

This section defines the metrics used to evaluate the performance of the LLMs. These definitions are provided for the general setting where fields can have multiple correct values.

Consider a target SafeMTS field with *K* acceptable values. Given a dataset with *N* near-miss records, matrix $\boldsymbol{Y} \in \{\boldsymbol{0}, \boldsymbol{1}\}^{N\times K}$ denotes the SME labels where $\boldsymbol{Y}_{i,j} = 1$ if the i^{th} record has the j^{th} acceptable value labeled. LLM predictions are similary denoted as a matrix $\widehat{\boldsymbol{Y}} \in \{\boldsymbol{0}, \boldsymbol{1}\}^{N\times K}$ where $\widehat{\boldsymbol{Y}}_{i,j} = 1$ if the LLM predicted j^{th} acceptable value for the i^{th} record. Notice, in SafeMTS fields which allow multiple values such as "Causal/Contributory Factors", each row of the matrix $\boldsymbol{Y}$ $\boldsymbol{Y}$ (or $\widehat{\boldsymbol{Y}}$) can contain multiple non-zero entries. For the k^{th} acceptable value, true positive is defined as $TP_k = \sum_{i=1}^{N} \boldsymbol{Y}_{i,k}\widehat{\boldsymbol{Y}}_{i,k}$, false negative is defined as $FN_k = \sum_{i=1}^{N} \boldsymbol{Y}_{i,k}\left(1 - \widehat{\boldsymbol{Y}}_{i,k}\right)$, and the false positive is defined as $FP_k = \sum_{i=1}^{N}\left(1 - \boldsymbol{Y}_{i,k}\right)\widehat{\boldsymbol{Y}}_{i,k}$.

Recall of the k^{th} acceptable value is given by:

$$k^{th}\ value's\ recall\ = \frac{TP_k}{TP_k + FN_k}$$

Micro-averaged recall is given by summing the nominator and denominator over all possible values, i.e.,

$$micro\ averaged\ recall\ = \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} TP_k + FN_k}$$

Micro-averaged recall equals is essentially equal to the total number of correctly predicted values divided by the total number of SME labels, hence we refer to it as overall accuracy.

Precision of the k^{th} acceptable value is given by:

$$k^{th}\ value's\ precision\ = \frac{TP_k}{TP_k + FP_k}$$

Micro-averaged precision, analogous to the micro-average recall, can similarly to be computed as follows:

$$micro\ averaged\ precision\ = \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} TP_k + FP_k}$$

Notice that micro-averaged precision is equivalent to the total number of correctly predicted values divided by the total number of predictions. As an alternative metric, we also computed fraction of records with at least one match is computed as $\frac{\sum_{i=1}^{K} \vee_{k=1}^{K} \boldsymbol{Y}_{i,k}\widehat{\boldsymbol{Y}}_{i,k}}{N}$ where $\vee$ denotes the logical OR operation. When using few-shot learning, the evaluation metrics are computed using the subset of the records that are not included as few-shot examples.

We used Ollama to serve the models locally and used ollama-python to prompt the models. To promote more reproducible results, we fixed the random seed to 42 and the temperature to 0.

Appendix C. Quality Dimensions

Elements critical to consider for high data quality are described below:

1. **Accuracy**: Is the data correct, precise, error-free (no typos, no duplicate entries, no human-generated errors)?
2. **Completeness**: Is the data complete (no missing information, no missing or empty fields, no unexplained acronyms). Incomplete data leads to information gaps and slows down the value creation process at best, often terminates it with no useful outcome.
3. **Consistency**: Is the data represented in the same way in/across the datasets? Are there any changes in the format and/or structure of the data, or in the names of the attributes used? Data standardization is a must for creating consistency in data.
4. **Granularity**: Data should have enough granularity and detail for the type of task that needs to be executed.
    A. **Blue-line data (Work as Imagined)** is needed for a proper comparison of the effects of several operational conditions such as sea state, weather, actual work hours and rest time etc. on injuries, injury causes, and other parameters of interest. [Conklin].
    B. Several factors which could be correlated with incidents are not present or they are present but with far too many missing values, such as personnel experience level, location of event, operational condition, and many more.
5. **Relevancy**: Do we know why we are storing this information? What is the purpose? Is it needed and relevant? There is no point in collecting information that is of no use. Irrelevant information creates confusion, and is a waste of time, money, and precious storage space.
    A. The company needs to make sure that every field in each record serves an actual business use case.
    B. This business use case needs to be communicated to the people, who actually touch the data. When people understand a data point's purpose, they are far more likely to enter that data correctly and completely.
    C. Critical Data Fields for specific reports/analyses need to be known, screened for accuracy as the data is entered and their accuracy needs be valued. For NLP analyses, we look at text incident title, business line, vessel name, date of the incident for time dependent analyses for example. Vessel name inconsistencies increased the level of effort.
6. **Reliability**: Data should not be ambiguous, vague, and should not contain contradictory information. Trusting data is the only way of building confidence in the information extracted from the data.
7. **Timeliness/Freshness**: How up to date is the data? What is the time lag between data creation and data publishing? In high-risk environments, the data analysis should be carried out immediately after the data has been collected. Anomaly detection before it becomes a threat is of paramount importance in preventing costly outcomes.
8. What is the lag between when data is collected and when it is presented on dashboards or available for analyses?
9. **Availability/Accessibility**: The right data should be available to the right people in the organization in a fast and easy way. This could be measured by whether the data can be accessed from the database(s) via an API.
10. **Usability**: How easy is it to work with the data? Data files should have meaningful, understandable, and relevant column names.

11. **Metadata**: Is the data well described? Do you have data describing the data stored? Metadata is a system catalog – a repository that lists all data files, their relationships, the attributes used in data files along with their intended use, format, structure, and the data type adopted. Well-planned and executed work on preparing the metadata resolves many data problems and saves time spent on searching for relevant data.

Appendix D. References

Black, H., and L. De Wolf. 2025. “Learning from Ferry Near-Misses: Human Factors-Driven Topic Modelling for Accident Prevention.” Conference presentation at IMHFS 2025, University of Strathclyde, Glasgow, United Kingdom.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. “Language Models Are Few-Shot Learners.” arXiv preprint arXiv:2005.14165. https://doi.org/10.48550/arXiv.2005.14165.

Chen, Ben, Mark Cunningham, David Grabaskas, Bruce Hamilton, and Sinem Perk. 2022. Success Path Method: Implementation Guidance. Lemont, IL: Argonne National Laboratory.

Crane, Steven, Phillip D. Sloane, Nancy C. Elder, Lauren W. Cohen, Natascha Laughtenschlager, and Sheryl Zimmerman. 2017. Implementing Near-Miss Reporting and Improvement Tracking in Primary Care Practices: Lessons Learned. Rockville, MD: Agency for Healthcare Research and Quality. https://www.ahrq.gov/patient-safety/reports/liability/crane.html.

Cunningham, Mark, Ben Chen, David Grabaskas, Bruce Hamilton, and Sinem Perk. 2022. Success Path Method: Definitions and Technical Requirements. Lemont, IL: Argonne National Laboratory.

Dekker, Sidney, and Todd E. Conklin. 2022. Do Safety Differently. ISBN 9798413008652.

Dy, C. T., and C. H. Mott. 2024. “Evaluating Near Midair Collision Reporting Systems Using Aircraft Surveillance Data: A Case Study at a University Airport.” Journal of Air Transport Management 124: 102088. https://doi.org/10.1016/j.jsr.2024.09.004.

Hamilton, B., et al. 2018. Risk-Based Evaluation of Offshore Oil and Gas Operations Using a Success Path Approach. Lemont, IL: Argonne National Laboratory. https://www.bsee.gov/sites/bsee.gov/files/research-reports/5009ac.pdf.

Inozu, Bahadir, and Curtis Doucette. 2021. “Using AI Technologies for Dynamic Risk Management.” Proceedings of the Marine Safety & Security Council: The Coast Guard Journal of Safety & Security at Sea 78 (2): 50–54. Washington, DC: U.S. Coast Guard.

Inozu, Bahadir, and Peter Schaedel. 2022. “Near Miss App Development – Part I: Survey.” Distribution Near Miss Report, prepared for Ship Operations Cooperative Program (SOCP), January 20, 2022. Washington, DC: Maritime Administration (MARAD) / SOCP.

Kalafatelis, A. S. 2025. “Towards Predictive Maintenance in the Maritime Industry.” Journal of Marine Science and Engineering 13 (3): 425. https://doi.org/10.3390/jmse13030425.

Kim, Tae-eun, and Anne Haugen Gausdal. 2020. “Leaders’ Influence Tactics for Safety: An Exploratory Study in the Maritime Context.” Safety 6 (1): 8. Basel: MDPI. https://doi.org/10.3390/safety6010008.

Krstinić, Damir, Marko Braović, Luka Šerić, and Dubravka Božić-Štulić. 2020. “Multi-Label Classifier Performance Evaluation with Confusion Matrix.” Computer Science & Information Technology (CSIT): 1–14. Chennai: AIRCC Publishing Corporation. https://doi.org/10.5121/csit.2020.100801.

Madjarov, Gjorgji, Dragi Kocev, Dejan Gjorgjevikj, and Sašo Džeroski. 2012. “An Extensive Experimental Comparison of Methods for Multi-Label Learning.” Pattern Recognition 45 (9): 3084–3104. Amsterdam: Elsevier. https://doi.org/10.1016/j.patcog.2012.03.004.

Madsen, Peter M., Robin L. Dillon, and Evan T. Morris. 2024. “Using Near Misses, Artificial Intelligence, and Machine Learning to Predict Maritime Incidents: A U.S. Coast Guard Case Study.” Risk Analysis. https://doi.org/10.1111/risa.15075.

Nazaruk, Marcin. 2025. Learning from Normal Work: How to Reduce Risk When Nothing Goes Wrong. London: Psychology Applied. ISBN 978-1068196416.

Ray, Archana Tikayat, Anirudh Prabhakara Bhat, Ryan T. White, Van Minh Nguyen, Olivia J. Pinon Fischer, and Dimitri N. Mavris. 2023. “Examining the Potential of Generative Language Models for Aviation Safety Analysis: Case Study and Insights Using the Aviation Safety Reporting System (ASRS).” Aerospace 10 (9): 770. https://doi.org/10.3390/aerospace10090770.

United States. Department of Transportation. Bureau of Transportation Statistics. 2020. “SafeMTS: Report on the Pilot.” Washington, DC: U.S. Department of Transportation. https://doi.org/10.21949/1529811.

Wang, Lian, Weijia Xu, Jinghan Jia, Weisheng Li, Ziqi Chen, Yifan Yang, et al. 2024. “Prompt Engineering in Consistency and Reliability with the Evidence-Based Guideline for LLMs.” npj Digital Medicine 7 (1): 41. London: Nature Publishing Group. https://doi.org/10.1038/s41746-024-01029-4.

Woltman, A. 2011. “Shell Process Safety Initiatives Target Sweet Spot for Reducing Risk.” Presented at the IADC Advanced Rig Technology Conference & Exhibition, Houston. Drilling Contractor, September 20, 2011. https://drillingcontractor.org/shell-process-safety-initiatives-target-sweet-spot-for-reducing-risk-11223.