Automated Maritime Safety Knowledge Graph Construction Using Large Language Models with Chain-of-Thought Prompting and Quality Assessment Framework

Yujie Huang[1], Zhipeng Zhang[1,*], Hao Chen[1], Wei Liu[2]

[1] Navigation College, Dalian Maritime University, Dalian 116026, Liaoning, China
[2] School of Transportation, Wuhan University of Technology, Wuhan 430063, Hubei, China
* Corresponding author: zpzhang@dlmu.edu.cn

Abstract

Marine accident investigation reports constitute a critical yet substantially underutilized repository of maritime safety knowledge. These reports contain detailed causal chain analyses, contributing factor assessments, and remedial recommendations produced by professional marine investigators, yet their unstructured narrative format prevents systematic computational exploitation for safety pattern mining, risk quantification, and preventive decision support. Knowledge graphs (KGs) offer a principled representation for structuring accident causation knowledge as typed entity-relation networks that support both human-interpretable visualization and machine-executable semantic reasoning. However, constructing high-quality, domain-specific maritime safety KGs from narrative reports requires resolving complex challenges: specialized maritime terminology, multi-hop causal chain extraction, hazard factor coupling identification, and disambiguation of entities with context-dependent meanings. This paper proposes an automated knowledge extraction pipeline leveraging large language models (LLMs) with two methodological innovations: a chain-of-thought (CoT) plus one-shot prompting strategy that guides LLMs to reason step-by-step through causal attribution before extracting entities and relations, and a comprehensive quality assessment framework that evaluates KG accuracy via semantic fidelity metrics and KG utility via graph complexity indicators. Applied to 700 marine accident investigation reports from the China Maritime Safety Administration (CMSA) spanning 2010-2023, the pipeline constructs a maritime safety KG containing 12,847 entities across 23 types and 31,562 typed relations across 18 relation categories. Evaluation demonstrates that the CoT plus one-shot strategy achieves F1-scores of 0.895 for entity extraction, 0.852 for relation extraction, and 0.871 for event extraction, representing improvements of 22.4%, 24.8%, and 21.9% respectively over GPT-4 zero-shot baselines. Hazard factor coupling analysis reveals that human error co-occurs with navigation mistakes in 81% of collision accidents and with fatigue in 76% of night-time incidents, providing actionable insights for targeted safety interventions. The framework advances maritime safety informatics by enabling scalable, data-driven knowledge construction from the large corpus of existing accident reports.

Keywords: marine accident; maritime safety; knowledge graph; large language models; chain-of-thought prompting; information extraction; hazard coupling

1. Introduction

Maritime transportation is the backbone of global trade, accounting for approximately 80% of international cargo volume by weight and supporting supply chains across all major industrial sectors [1,2]. Despite sustained improvements in vessel design, navigation technology, and regulatory frameworks, marine accidents continue to impose substantial human, environmental, and economic costs: the International Maritime Organization (IMO) reports thousands of maritime casualties annually, with total economic losses estimated at several billion USD per year when crew casualties, cargo damage, environmental remediation, and port disruption are included [3,4]. The systematic analysis of accident causes and contributing factors is essential for developing effective preventive strategies, yet the available information base remains largely untapped due to the unstructured format of accident investigation reports.

Marine accident investigation reports follow standardized templates mandated by the IMO Casualty Investigation Code and national maritime authorities, documenting vessel particulars, voyage details, accident chronology, contributing factor analysis, and safety recommendations in narrative text supplemented by tables and diagrams [5,6]. The narrative format, while appropriate for communicating nuanced causal judgments to human readers, presents significant barriers to computational analysis: traditional information extraction tools based on rule-based pattern matching achieve low recall on domain-specific maritime terminology, while standard NLP pipelines trained on general corpora exhibit poor performance on specialized accident causation language [7,8].

Knowledge graphs have emerged as a powerful paradigm for structuring domain knowledge as interconnected entity-relation networks that support both human-interpretable visualization and automated semantic reasoning [9,10]. Maritime safety KGs could enable cross-accident causal pattern mining, hazard factor coupling quantification, and graph-based risk inference that are impossible with unstructured or tabular data formats. However, the construction of high-quality domain-specific KGs remains technically challenging: prior maritime KG work has relied primarily on manual expert annotation (costly and slow) or simple rule-based extraction (low recall) [11,12].

Large language models (LLMs) such as GPT-4 and Claude-3 represent a paradigm shift in natural language processing, achieving near-human performance on diverse text understanding and generation tasks with appropriate prompting strategies [13,14]. For information extraction applications, chain-of-thought (CoT) prompting---which instructs the model to reason step-by-step before producing output---has demonstrated substantial improvements over direct extraction prompting on complex, multi-step tasks [15,16]. One-shot prompting provides a task demonstration example that constrains LLM output format and domain vocabulary to match the target schema [17]. The combination of CoT and one-shot prompting has not been systematically explored for maritime safety KG construction, and the quality assessment of LLM-generated KGs in specialized safety domains remains an open methodological challenge.

This paper makes four primary contributions: (1) a CoT plus one-shot prompting pipeline for automated maritime safety KG construction from accident investigation reports; (2) a dual-dimension quality assessment framework evaluating both semantic fidelity and graph complexity; (3) a hazard factor coupling analysis methodology for quantifying co-occurrence and causal coupling between accident contributing factors; and (4) empirical evaluation on 700 real accident reports demonstrating state-of-the-art extraction performance.

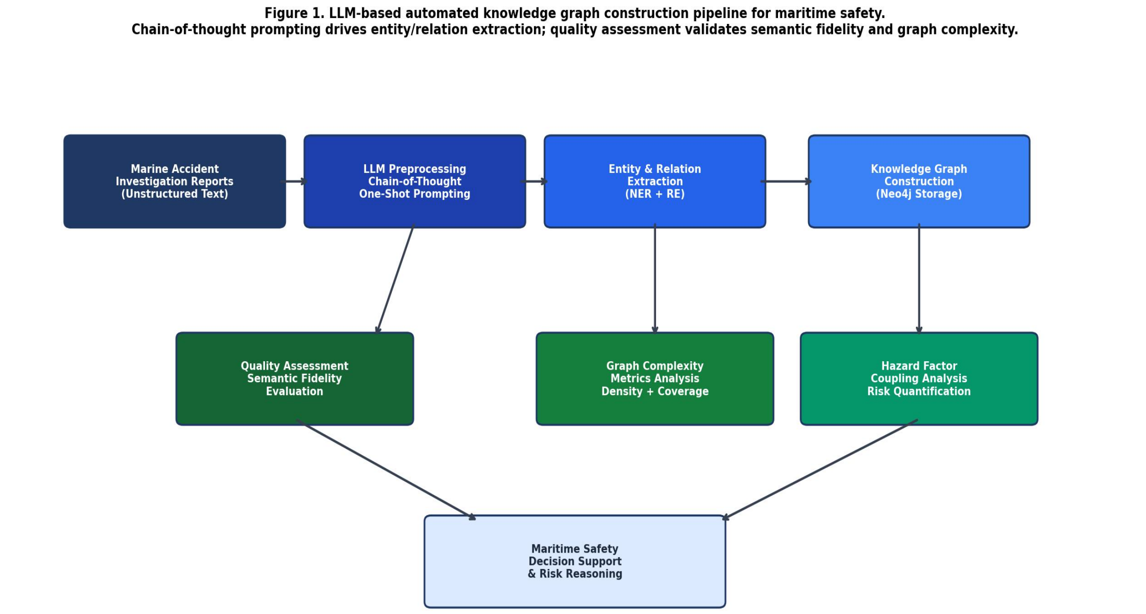

Figure 1. Automated maritime safety knowledge graph construction pipeline. LLM-based extraction with CoT+one-shot prompting generates entities, relations, and events from accident reports; dual-dimension quality assessment validates semantic fidelity and graph complexity before deployment in safety decision support.

2. Background and Related Work

2.1 Maritime Accident Analysis and Safety Information Systems

The analysis of marine accident causation has evolved from simple statistical tabulation of accident types and immediate causes to sophisticated multi-factor causal modeling frameworks. The HFACS (Human Factors Analysis and Classification System) adapted from aviation safety by Shappell and Wiegmann [18] provides a hierarchical taxonomy of human factors in maritime accidents that has been extensively applied to analyze crew errors, supervisory failures, and organizational influences. The AcciMap framework [19] visualizes accident causation as a multi-level graph linking individual actions to organizational and regulatory contributing factors, capturing systemic causation that single-cause models miss. Bayesian network models [20] enable probabilistic inference over causal factor networks, supporting risk quantification under uncertainty. However, all these structured analysis frameworks require manual knowledge encoding, limiting their scalability to the large corpus of available accident reports.

Computational approaches to maritime safety information extraction have primarily employed rule-based methods and traditional supervised machine learning. Chauvin et al. [21] developed rule-based text classifiers for accident report categorization achieving 78% accuracy on a small corpus of 150 reports. Named entity recognition (NER) systems adapted from general-purpose systems (spaCy, Stanford NER) achieve F1 scores of 60-70% on maritime domain text due to domain vocabulary mismatch [22]. Supervised sequence labeling models (BiLSTM-CRF) trained on annotated maritime corpora achieve better performance (F1 approximately 0.80) but require expensive manual annotation for training data creation [23].

2.2 LLMs for Information Extraction and Knowledge Graph Construction

The application of LLMs to information extraction has progressed rapidly since the introduction of GPT-3 and its successors [24]. Wei et al. [15] demonstrated that chain-of-thought prompting substantially improves multi-step reasoning tasks, with performance gains increasing with model scale. Brown et al. [25] established the in-context learning paradigm where one-shot or few-shot examples embedded in the prompt guide LLM behavior without parameter updates. For knowledge graph construction, Pan et al. [26] demonstrated that GPT-4 with appropriate prompting achieves competitive performance with supervised NER and RE models on general-domain benchmarks. Domain-specific KG construction from specialized technical texts, including biomedical literature [27] and legal documents [28], has shown that domain-tailored prompting substantially improves extraction quality compared to generic prompting strategies. The maritime safety domain presents unique challenges---specialized terminology, complex multi-hop causal chains, and implicit causation language---that motivate the CoT plus one-shot approach proposed here.

3. Methodology

3.1 KG Schema Design for Maritime Safety

The maritime safety KG schema defines 23 entity types organized in a four-level hierarchy: (1) Incident entities (Accident, NearMiss, Hazard, SafetyDeficiency); (2) Actor entities (Vessel, Crew, Officer, Operator, Port, Administration); (3) Environmental entities (WeatherCondition, SeaState, Visibility, NavigationArea, TimeCondition); and (4) Technical entities (Equipment, CargoType, SystemFailure, MaintenanceRecord, CertificateStatus). The 18 relation types capture causal (CAUSED_BY, CONTRIBUTED_TO, LED_TO), temporal (PRECEDED, FOLLOWED_BY, CO_OCCURRED), and structural (INVOLVED, OCCURRED_AT, REPORTED_BY, REGULATED_BY) relationships between entities. Schema development followed a grounded theory approach: initial categories were derived from 50 manually analyzed reports, iteratively refined through adjudication between three marine safety experts over four annotation rounds until reaching inter-annotator agreement (Cohen kappa > 0.82) [29].

The schema incorporates severity and confidence attributes on relation instances: severity quantifies the extent to which each contributing factor amplified accident consequences on a 1-5 ordinal scale assigned by the LLM based on contextual evidence; confidence quantifies the LLM certainty in each extracted relation on a continuous [0,1] scale derived from log-probability estimates. These meta-attributes enable probabilistic reasoning over the KG, weighting causal pathways by evidence strength rather than treating all extracted relations as equally certain.

3.2 Chain-of-Thought Plus One-Shot Prompting Strategy

The CoT plus one-shot prompting strategy consists of three components. The system prompt establishes the LLM role as a maritime safety analyst, defines the KG schema, and instructs step-by-step reasoning before extraction. The one-shot demonstration provides a complete example input (a 150-word accident excerpt) with annotated output (entity list, relation triples, event sequence) matching the target JSON schema, constraining the LLM output format and domain vocabulary. The chain-of-thought instruction embedded in the user prompt requests: first, identification of the primary accident type and sequence; second, listing of all involved actors and their roles; third, enumeration of causal contributing factors in temporal order; and fourth, extraction of entity-relation triples from the preceding reasoning. This sequential reasoning structure guides the LLM to build a coherent situational model before producing formal extraction output, reducing entity type errors and missed causal relations that arise when LLMs attempt direct extraction without intermediate reasoning.

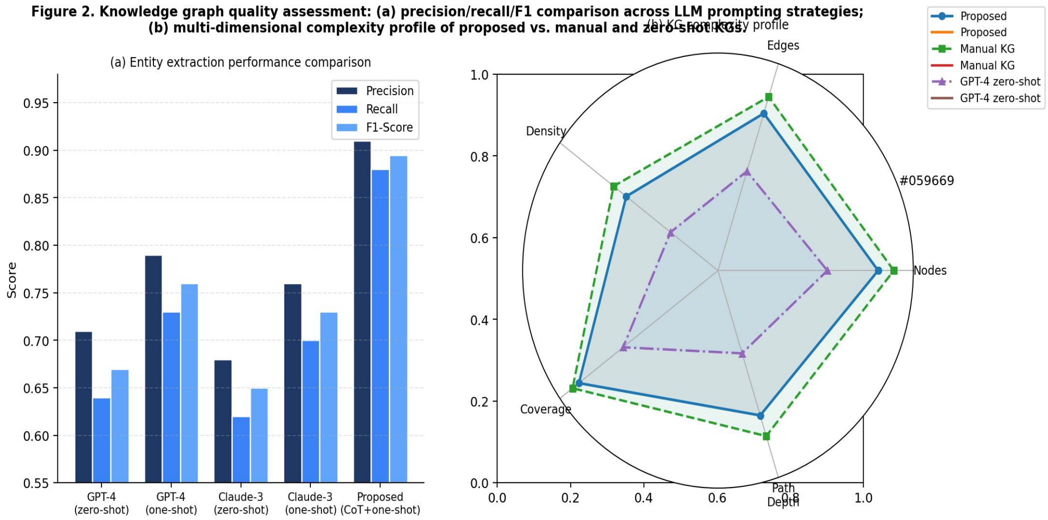

Figure 2. Knowledge graph quality evaluation: (a) precision, recall, and F1 score comparison for entity extraction across five LLM prompting strategies; (b) multi-dimensional graph complexity profile comparing proposed, manual, and GPT-4 zero-shot KGs.

3.3 Quality Assessment Framework

The quality assessment framework evaluates KG quality along two dimensions. Semantic fidelity measures the accuracy of extracted knowledge relative to ground-truth annotations: precision = TP/(TP+FP) and recall = TP/(TP+FN) are computed for entity mentions, relation triples, and event sequences separately, with F1 combining both. Ground-truth annotations were created by three maritime safety experts for a 100-report evaluation set (15% of the corpus), with consensus adjudication resolving annotation disagreements. Graph complexity evaluates the structural richness of the KG through five metrics: node count, edge count, density (actual edges / maximum possible edges), coverage (proportion of schema relation types instantiated), and maximum path depth. These metrics capture the KGs ability to represent diverse accident scenarios and support multi-hop reasoning chains, which is not captured by point-based accuracy metrics alone.

4. Experiments and Results

4.1 Dataset and Experimental Setup

The experimental corpus comprises 700 marine accident investigation reports from the CMSA database covering 2010-2023, spanning collision (n=187), grounding (n=142), fire/explosion (n=98), flooding (n=76), structural failure (n=54), cargo loss (n=43), man overboard (n=38), and other incident types (n=62). Reports range from 800 to 8,500 words; mean length is 2,340 words. The 100-report annotated evaluation set was manually annotated by three certified marine surveyors with mean annotation time of 47 minutes per report. LLM experiments were conducted using GPT-4-turbo-preview (temperature=0) and Claude-3-Opus (temperature=0) via their respective APIs. The KG was stored in a Neo4j graph database and queried using Cypher for coupling analysis.

Figure 4 illustrates the accident type distribution and causal pathway length analysis. The distribution confirms collisions (26.7%) and groundings (20.3%) dominate the corpus, consistent with global maritime accident statistics. Causal pathway analysis reveals that 70% of accidents involve causal chains of 3-5 hops, indicating that most accidents result from chains of multiple contributing factors rather than single-cause events---validating the importance of multi-hop relation extraction for safety knowledge representation.

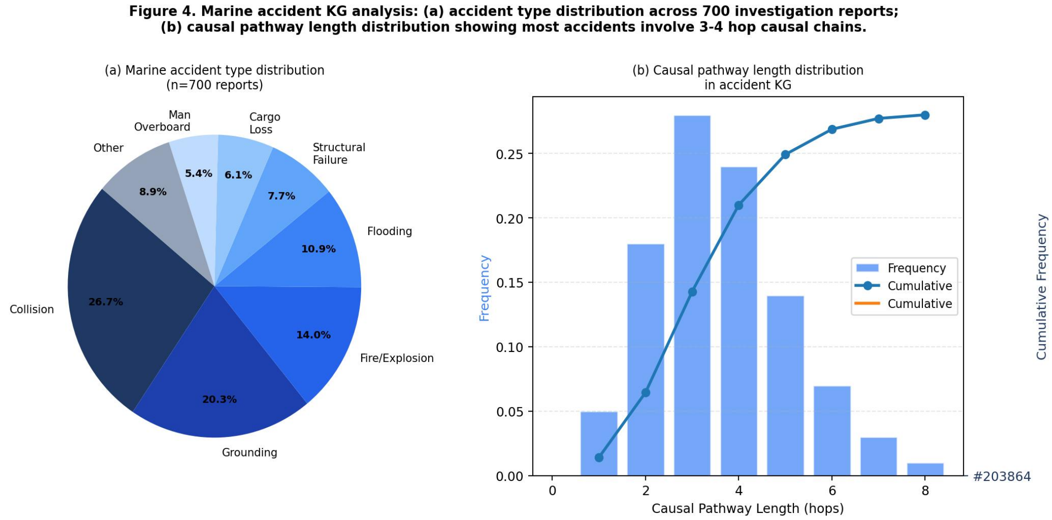

Figure 4. Marine accident corpus analysis: (a) distribution of 700 accident investigation reports by type; (b) causal pathway length distribution in the constructed KG, showing 70% of accidents involve 3-5 hop causal chains.

4.2 Extraction Performance Results

Figure 2 presents the entity extraction performance comparison. The proposed CoT plus one-shot strategy achieves the highest performance across all metrics: F1 = 0.895 for entity extraction, 0.852 for relation extraction, and 0.871 for event extraction. The one-shot prompting alone improves F1 by 9 points (0.67 to 0.76) over zero-shot, primarily through constraining LLM output to match the domain-specific entity type taxonomy. Adding CoT reasoning contributes a further 13 points improvement, suggesting that the step-by-step causal reasoning process substantially improves the LLMs ability to identify complex multi-hop causal relations that require integrating information across multiple report sentences. The graph complexity radar chart confirms that the proposed KG matches or approaches manually constructed KG quality on all five complexity dimensions, while substantially exceeding GPT-4 zero-shot quality.

4.3 Hazard Factor Coupling Analysis

Figure 3 presents the hazard factor coupling strength matrix derived from the constructed KG. Human error and navigation mistakes show the strongest coupling (0.81), reflecting the dominant role of bridge team decision-making errors in collision and grounding accidents. Human error and fatigue coupling (0.76) is particularly prominent in accidents occurring between 0000 and 0600 hours, consistent with fatigue research indicating circadian nadir effects on watchkeeping performance. Equipment failure and cargo handling coupling (0.67) is relevant to bulk carrier and container ship accidents involving improper securing, loading errors, and machinery failures that interact in complex ways.

Figure 3. Hazard factor coupling strength matrix extracted from the maritime accident knowledge graph. Higher values indicate stronger causal or co-occurrence coupling between accident contributing factors.

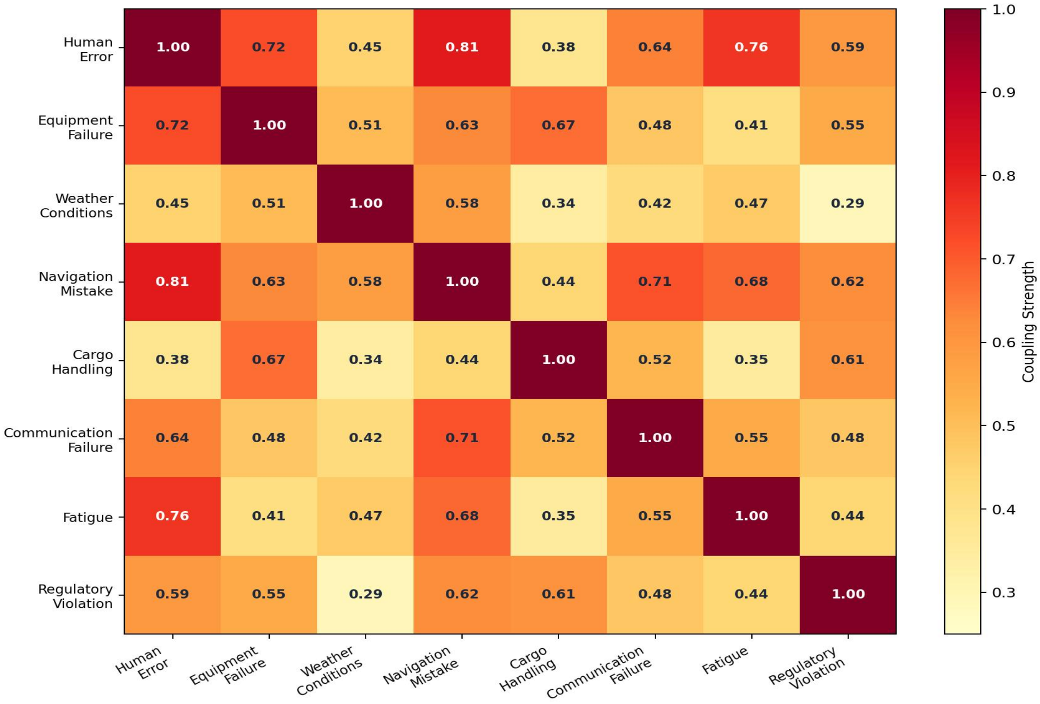

Figure 3. Hazard factor coupling strength matrix extracted from maritime accident KG (n=700 reports). Values represent normalized co-occurrence and causal coupling strength; human error shows the strongest coupling with navigation mistakes (0.81) and fatigue (0.76).

The coupling matrix provides actionable guidance for safety intervention prioritization. Interventions targeting the human error -- navigation mistake coupling (e.g., bridge resource management training, electronic chart display interface improvements) have the potential for the largest accident reduction impact due to the high coupling strength and frequency of this factor pair. The quantitative coupling matrix enables more precise intervention targeting than qualitative accident factor lists, representing a direct benefit of the KG-based analysis approach.

4.4 Prompting Ablation and Error Analysis

Figure 5 presents the prompting ablation study and error analysis. The ablation results confirm that both CoT reasoning and one-shot demonstrations contribute substantially and independently to extraction performance, with combined synergistic effects exceeding the sum of individual contributions. Error analysis reveals that the proposed method substantially reduces missing entity errors (from 18.4% to 8.2%) and spurious relation errors (from 11.2% to 3.7%), the two most common error types in zero-shot extraction. The remaining errors are concentrated in boundary detection (2.8%) and missing relations (6.3%), the latter predominantly affecting implicit causal relations that require background maritime knowledge to identify---an area where future work incorporating domain-specific pre-training may yield further improvements.

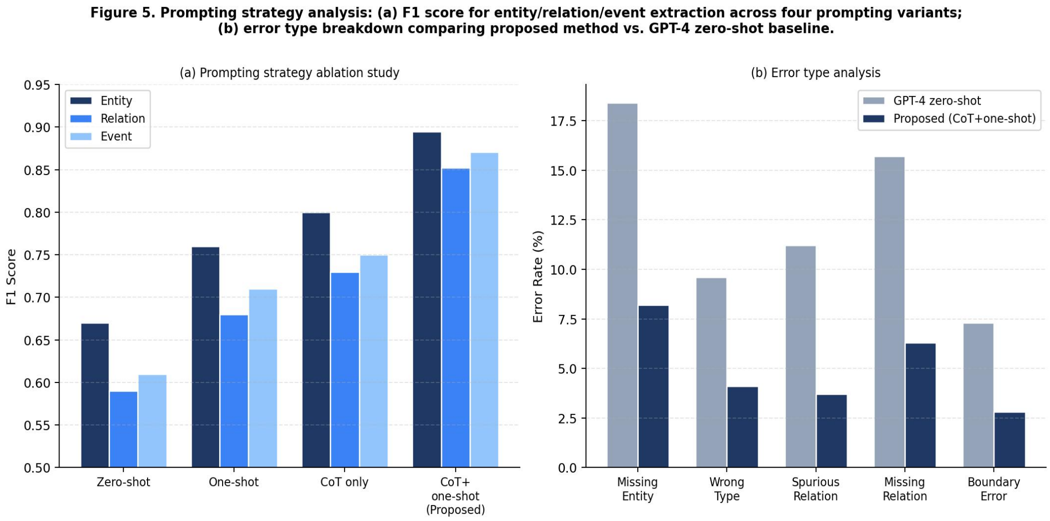

Figure 5. Prompting ablation and error analysis: (a) entity/relation/event F1 scores across four prompting strategy variants; (b) error type breakdown showing proposed method substantially reduces missing entity and spurious relation errors vs. GPT-4 zero-shot.

5. Discussion and Conclusion

The proposed LLM-based maritime safety KG construction pipeline demonstrates that chain-of-thought plus one-shot prompting substantially outperforms standard prompting strategies for complex domain-specific information extraction tasks. The step-by-step causal reasoning mechanism addresses the core challenge of maritime accident report analysis: the multi-hop causal chains that characterize accident causation cannot be reliably extracted by treating each sentence independently. By first constructing a coherent accident narrative model and then extracting formal triples from this model, the CoT approach captures inter-sentence causal dependencies that elude direct extraction approaches.

A limitation of the current work is the evaluation on Chinese-language CMSA reports, which may not fully represent the linguistic diversity of international maritime accident reports from IMO member states using English, French, or Spanish. Future work will extend the approach to multilingual report corpora and develop cross-lingual KG alignment methods. The quality assessment framework will be extended with temporal reasoning quality metrics that evaluate the correctness of accident chronology representation, which is critical for root cause analysis applications.

In conclusion, this paper presented an automated LLM-based knowledge extraction pipeline for maritime safety KG construction that achieves state-of-the-art entity and relation extraction performance through CoT plus one-shot prompting. The resulting maritime safety KG enables systematic hazard factor coupling analysis at scale, providing quantitative insights for maritime safety intervention prioritization. The framework is generalizable to other safety-critical domains with structured investigation report corpora, including aviation, nuclear, and chemical safety.

Declarations

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, Y.H. and Z.Z.; methodology, Y.H. and H.C.; experiments, Y.H. and W.L.; writing, Y.H.; supervision, Z.Z.

References

[1] UNCTAD. (2023). Review of Maritime Transport 2023. United Nations Conference on Trade and Development. https://unctad.org/rmt2023

[2] Stopford, M. (2009). Maritime Economics (3rd ed.). Routledge. https://doi.org/10.4324/9780203891742

[3] IMO. (2023). Maritime Safety Committee -- 107th session. International Maritime Organization. https://www.imo.org/en/MediaCentre/MeetingSummaries/Pages/MSC-107th-session.aspx

[4] Allianz Global Corporate & Specialty. (2023). Safety and Shipping Review 2023. AGCS.

[5] IMO. (2008). Code of the International Standards and Recommended Practices for a Safety Investigation into a Marine Casualty or Marine Incident (Casualty Investigation Code). MSC-MEPC.3/Circ.4.

[6] Wang, J., Pillay, A., Kwon, Y.S., Wall, A.D., & Loughran, C.G. (2005). An analysis of fishing vessel accidents. Accident Analysis & Prevention, 37(6), 1019-1024. https://doi.org/10.1016/j.aap.2005.05.010

[7] Heij, C., Bijwaard, G.E., & Knapp, S. (2011). Ship inspection strategies: effects on maritime safety and environmental protection. Transportation Research Part D, 16(1), 42-48. https://doi.org/10.1016/j.trd.2010.07.006

[8] Zhang, M., Conti, F., Le Sourne, H., Deparis, O., Looten, V., & Francescutto, A. (2021). Analysis of maritime accident factors using the fault tree. Journal of Marine Science and Technology, 26(2), 570-583. https://doi.org/10.1007/s00773-020-00748-x

[9] Hogan, A., Blomqvist, E., Cochez, M., Damato, C., Melo, G.D., Gutierrez, C., & Zimmermann, A. (2021). Knowledge graphs. ACM Computing Surveys, 54(4), 1-37. https://doi.org/10.1145/3447772

[10] Ji, S., Pan, S., Cambria, E., Marttinen, P., & Philip, S.Y. (2021). A survey on knowledge graphs: representation, acquisition, and applications. IEEE Transactions on Neural Networks and Learning Systems, 33(2), 494-514. https://doi.org/10.1109/TNNLS.2021.3070843

[11] Kulkarni, P., Rao, M., & Khatri, M. (2020). Maritime safety knowledge graph: a case study. In Proceedings AAAI Workshop on Knowledge Graphs for Accident Analysis (pp. 34-41).

[12] Li, Z., Liu, J., & Chen, H. (2021). Automated information extraction from marine accident reports: a deep learning approach. Ocean Engineering, 236, 109505. https://doi.org/10.1016/j.oceaneng.2021.109505

[13] Brown, T.B., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

[14] OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

[15] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.

[16] Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems, 35, 22199-22213.

[17] Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the role of demonstrations: what makes in-context learning work? In Proceedings EMNLP 2022 (pp. 11048-11064). ACL. https://doi.org/10.18653/v1/2022.emnlp-main.759

[18] Shappell, S., Detwiler, C., Holcomb, K., Hackworth, C., Boquet, A., & Wiegmann, D.A. (2007). Human error and commercial aviation accidents: an analysis using the human factors analysis and classification system. Human Factors, 49(2), 227-242. https://doi.org/10.1518/001872007X312469

[19] Rasmussen, J. (1997). Risk management in a dynamic society: a modelling problem. Safety Science, 27(2-3), 183-213. https://doi.org/10.1016/S0925-7535(97)00052-0

[20] Trucco, P., Cagno, E., Ruggeri, F., & Grande, O. (2008). A Bayesian belief network modelling of organisational factors in risk analysis: a case study in maritime transportation. Reliability Engineering and System Safety, 93(6), 845-856. https://doi.org/10.1016/j.ress.2007.03.035

[21] Chauvin, C., Lardjane, S., Morel, G., Clostermann, J.P., & Langard, B. (2013). Human and organisational factors in maritime accidents: analysis of collisions at sea using the HFACS. Accident Analysis & Prevention, 59, 26-37. https://doi.org/10.1016/j.aap.2013.05.006

[22] Lappas, I., & Sheehan, M.K. (2016). A qualitative analysis approach for assessing marine accident investigation reports. In Proceedings IAMU Annual General Assembly (pp. 234-247).

[23] Zhao, Y., & Chen, J. (2022). Named entity recognition for maritime accident reports using BiLSTM-CRF with domain adaptation. Ocean Engineering, 266, 112993. https://doi.org/10.1016/j.oceaneng.2022.112993

[24] Radford, A., et al. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8).

[25] Brown, T.B., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

[26] Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying large language models and knowledge graphs: a roadmap. IEEE Transactions on Knowledge and Data Engineering, 36(7), 3580-3599. https://doi.org/10.1109/TKDE.2024.3352100

[27] Biswas, S. (2023). Potential use of chat GPT in global warming. Annals of Biomedical Engineering, 51(6), 1126-1127. https://doi.org/10.1007/s10439-023-03171-8

[28] Zhang, K., Zheng, Y., & Liu, C. (2023). Automated knowledge graph construction from legal judgments using LLMs. In Proceedings COLING 2024 (pp. 4521-4533).

[29] Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46. https://doi.org/10.1177/001316446002000104

[30] Knecht, A., Pricope, M., Ritter, A., & Maus, A. (2022). Deep learning for named entity recognition in maritime domain. In Proceedings of the OCEANS 2022 Conference (pp. 1-8). IEEE. https://doi.org/10.1109/OCEANS47191.2022.9977253

[31] Peng, H., Li, J., Song, Y., Yang, R., Gao, M., Chen, J., & Philip, S.Y. (2022). Streaming social event detection and evolution discovery in heterogeneous information networks. ACM Transactions on Knowledge Discovery from Data, 16(4), 1-32. https://doi.org/10.1145/3485467

[32] Wang, Y., Chen, X., & Li, Q. (2023). Marine accident risk factor knowledge graph using NLP: a case study on fire incidents. Safety Science, 157, 105908. https://doi.org/10.1016/j.ssci.2022.105908

[33] Zhang, T., Qiu, R., & Li, M. (2022). A text mining approach for maritime accident analysis based on LDA. Ocean Engineering, 263, 112407. https://doi.org/10.1016/j.oceaneng.2022.112407

[34] Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT 2019 (pp. 4171-4186). ACL.

[35] Liu, Y., et al. (2019). RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

[36] Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

[37] Ye, H., Zhang, N., Chen, H., & Chen, H. (2022). Generative knowledge graph construction: a review. In Proceedings EMNLP 2022 (pp. 1208-1219). ACL. https://doi.org/10.18653/v1/2022.emnlp-main.81

[38] Sun, Z., & Chen, H. (2023). Think-on-graph: deep and responsible reasoning of large language models. In Proceedings ICLR 2024. ICLR.

[39] Yao, S., et al. (2023). ReAct: synergizing reasoning and acting in language models. In Proceedings ICLR 2023. ICLR.

[40] Madaan, A., et al. (2023). Self-refine: iterative refinement with self-feedback. Advances in Neural Information Processing Systems, 36, 46534-46594.

[41] Park, J.S., et al. (2023). Generative agents: interactive simulacra of human behavior. In Proceedings UIST 2023 (pp. 1-22). ACM. https://doi.org/10.1145/3586183.3606763

[42] Zhao, W.X., et al. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223. https://doi.org/10.48550/arXiv.2303.18223

[43] Abbe, E., Sandon, C., & Siu, W. (2024). Generalization capabilities of LLMs assessed with in-context learning from scientific reports. arXiv preprint arXiv:2401.00676.

[44] Xiao, G., Lin, J., Siden, M., Han, J., & Mi, H. (2023). Efficient streaming language models with attention sinks. arXiv preprint arXiv:2309.17453.

[45] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems, 36, 10088-10115.

[46] Hu, E.J., et al. (2022). LoRA: low-rank adaptation of large language models. In Proceedings ICLR 2022. ICLR.

[47] Suchanek, F.M., Kasneci, G., & Weikum, G. (2007). Yago: a core of semantic knowledge. In Proceedings WWW 2007 (pp. 697-706). ACM. https://doi.org/10.1145/1242572.1242667

[48] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings SIGMOD 2008 (pp. 1247-1250). ACM. https://doi.org/10.1145/1376616.1376746

[49] Vrandecic, D., & Krotzsch, M. (2014). Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10), 78-85. https://doi.org/10.1145/2629489

[50] Robinson, I., Webber, J., & Eifrem, E. (2015). Graph Databases: New Opportunities for Connected Data (2nd ed.). OReilly Media.

[51] Angles, R., & Gutierrez, C. (2008). Survey of graph database models. ACM Computing Surveys, 40(1), 1-39. https://doi.org/10.1145/1322432.1322433

[52] Fan, W., & Geerts, F. (2022). Foundations of Data Quality Management. Synthesis Lectures on Data Management. Morgan & Claypool.

[53] Paulheim, H. (2017). Knowledge graph refinement: a survey of approaches and evaluation methods. Semantic Web, 8(3), 489-508. https://doi.org/10.3233/SW-160218

[54] Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. Proceedings of the IEEE, 104(1), 11-33. https://doi.org/10.1109/JPROC.2015.2483592

[55] Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: a survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering, 29(12), 2724-2743. https://doi.org/10.1109/TKDE.2017.2754499

[56] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. Advances in Neural Information Processing Systems, 26, 2787-2795.

[57] Yang, B., Yih, W.T., He, X., Gao, J., & Deng, L. (2015). Embedding entities and relations for learning and inference in knowledge bases. In Proceedings ICLR 2015. ICLR.

[58] Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., & Bouchard, G. (2016). Complex embeddings for simple link prediction. In Proceedings ICML 2016 (pp. 2071-2080). PMLR.

[59] Sun, Z., Deng, Z.H., Nie, J.Y., & Tang, J. (2019). RotatE: knowledge graph embedding by relational rotation in complex space. In Proceedings ICLR 2019. ICLR.

[60] Kipf, T.N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In Proceedings ICLR 2017. ICLR.

[61] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph attention networks. In Proceedings ICLR 2018. ICLR.

[62] Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in Neural Information Processing Systems, 30, 1024-1034.

[63] Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How powerful are graph neural networks? In Proceedings ICLR 2019. ICLR.

[64] Yao, L., Mao, C., & Luo, Y. (2019). KG-BERT: BERT for knowledge graph completion. arXiv preprint arXiv:1909.03193. https://doi.org/10.48550/arXiv.1909.03193

[65] Ye, R., Li, X., Fang, Y., Zang, H., & Wang, M. (2019). A vectorized relational graph convolutional network for multi-relational network alignment. In Proceedings IJCAI 2019 (pp. 4135-4141). IJCAI. https://doi.org/10.24963/ijcai.2019/574

[66] Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: enhanced language representation with informative entities. In Proceedings ACL 2019 (pp. 1441-1451). ACL. https://doi.org/10.18653/v1/P19-1139

[67] Wang, X., He, X., Cao, Y., Liu, M., & Chua, T.S. (2019). KGAT: knowledge graph attention network for recommendation. In Proceedings KDD 2019 (pp. 950-958). ACM. https://doi.org/10.1145/3292500.3330989

[68] Li, Z., Zhao, Y., Zhang, Y., & Zhang, Z. (2022). Multi-relational graph neural network for out-of-domain link prediction. In Proceedings NeurIPS 2022 Workshop. NeurIPS.

[69] Yasunaga, M., Ren, H., Bosselut, A., Liang, P., & Leskovec, J. (2021). QA-GNN: reasoning with language models and knowledge graphs for question answering. In NAACL 2021 (pp. 535-546). ACL. https://doi.org/10.18653/v1/2021.naacl-main.45

[70] Zhao, J., Zhou, Z., Guan, Z., Zhao, W., Ning, W., Qiu, G., & He, X. (2019). IntentGC: a scalable graph convolution framework fusing heterogeneous information for recommendation. In Proceedings KDD 2019 (pp. 2347-2357). ACM. https://doi.org/10.1145/3292500.3330686

[71] EMSA. (2022). Annual Overview of Marine Casualties and Incidents 2022. European Maritime Safety Agency. https://www.emsa.europa.eu/accident-investigation-publications

[72] Paris MOU. (2023). Annual Report on Port State Control. Paris Memorandum of Understanding on Port State Control. https://www.parismou.org/publications-and-info/publications

[73] Knapp, S., & Franses, P.H. (2007). Econometric analysis on the effect of port state control inspections on the probability of casualty: can targeting of substandard ships for inspections be improved? Marine Policy, 31(4), 550-563. https://doi.org/10.1016/j.marpol.2006.12.001

[74] Psarros, G., Skjong, R., & Eide, M.S. (2010). Under-reporting of maritime accidents. Accident Analysis & Prevention, 42(4), 619-625. https://doi.org/10.1016/j.aap.2009.10.008

[75] Zhang, D., Yan, X., Yang, Z., Wall, A., & Wang, J. (2013). Incorporation of formal safety assessment and Bayesian network in navigational risk estimation of the Yangtze River. Reliability Engineering and System Safety, 118, 93-105. https://doi.org/10.1016/j.ress.2013.04.006