MariATE: Automatic Term Extraction Using Large Language Models in the Maritime Domain

Shijie Liu[1,2], Els Lefever[2] and Véronique Hoste[2]

[1]College of Foreign Languages, Shanghai Maritime University, Shanghai, 201306, China
[2]Language and Translation Technology Team, Ghent University, Ghent, 9000, Belgium

henryliushijie@163.com, {firstname.lastname}@ugent.be

Abstract

This study presents a comprehensive evaluation of Large Language Models (LLMs) for automatic term extraction in the maritime safety domain. The research examines the zero-shot performance of seven state-of-the-art LLMs, including both open-source and closed-source models, and investigates terminology annotation strategies for optimal coverage. Nested annotation captures both complete technical expressions and their constituent components, while full-term annotation focuses exclusively on maximal-length terms. Experimental results demonstrate Claude-3.5-Sonnet's superior performance (F1-score of 0.80) in maritime safety terminology extraction, particularly in boundary detection capabilities. Error analysis reveals three primary challenges: distinguishing contextual descriptions from legitimate terminology, handling complex multi-word expressions, and identifying maritime safety operational and navigational terms. Analysis of annotation strategies reveals that the full-term annotation approach achieves 95.24% coverage of unique terms compared to the nested annotation approach. The additional 4.76% of terms identified through nested annotation represents subcomponents of larger technical expressions. These findings advance the understanding of LLMs' capabilities in specialized terminology extraction and provide empirical evidence supporting the sufficiency of full-term annotation for comprehensive terminology coverage in domain-specific applications.

1 Introduction

Maritime safety represents a critical framework governing the world's most international industry - shipping, where preventing loss of life, vessel damage, and marine environmental harm remains paramount. Defined by the International Maritime Organization, it encompasses comprehensive technical regulations spanning vessel construction, navigation systems, crew training, and emergency procedures, all regulated through international conventions like SOLAS[1] and STCW[2]. The automated extraction of this safety-critical terminology presents unique challenges in natural language processing (NLP), particularly due to the domain's distinctive characteristics: multi-word technical expressions, nested terms, context-dependent terms, and hierarchical terminology structures reflecting complex maritime systems. While automatic terminology extraction has advanced in fields such as biomedicine (Kim et al., 2003), computational linguistics (QasemiZadeh and Schumann, 2016), education (Banerjee et al., 2022), power engineering (Ivanović et al., 2022), as well as studies covering corruption, dressage, heart failure and wind energy (Rigouts Terryn et al., 2020), the maritime safety domain remains underexplored, despite its critical role in ensuring global maritime operations.

Automatic term extraction (ATE) is a fundamental natural language processing task that aims to automatically identify domain-specific terms from text corpora (Vivaldi and Rodríguez, 2007). Traditional ATE approaches can be categorized into three main types: linguistic rule-based methods (Frantzi Katerina et al., 2000), statistical methods (Nakagawa and Mori, 2002), and hybrid approaches (Maynard and Ananiadou, 2000; Macken et al., 2013; Cram and Daille, 2016). Recent years have witnessed the emergence of more sophisticated approaches to address the limitations of traditional methods, including topic modeling (Bolshakova et al., 2013), machine learning techniques (Hazem et al., 2020; Rigouts Terryn et al., 2021,

[1]The International Convention for the Safety of Life at Sea, first adopted in 1914 in response to the Titanic disaster, establishes comprehensive standards for merchant ship safety.

[2]The International Convention on Standards of Training, Certification and Watchkeeping for Seafarers, adopted in 1978 and revised in 1995/2010, sets minimum seafarer competency standards to ensure safe and efficient vessel operations.

Proceedings of Recent Advances in Natural Language Processing,pages 663–673
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-077

2022b), and cross-domain and/or cross-lingual transfer learning methods (Tran et al., 2022a,b; Hazem et al., 2022). These approaches have been widely applied to support various downstream tasks, including information retrieval (Peñas et al., 2001), machine translation (Wolf et al., 2011), aspect-based sentiment analysis (De Clercq et al., 2015), ontology building (Iqbal et al., 2017), and translation quality estimation (Yuan et al., 2018).

The advent of Large Language Models (LLMs) has recently revolutionized natural language processing through their remarkable zero-shot and few-shot learning capabilities without extensive training (Agrawal et al., 2022; Kojima et al., 2022; Banerjee et al., 2024). LLMs have demonstrated their superior performance across various specialized tasks, including text classification (Chae and Davidson, 2023), named entity recognition (Wang et al., 2023; Ashok and Lipton, 2023), relation extraction (Zavarella et al., 2024), text augmentation (Yoo et al., 2021), aspect-based sentiment analysis (Simmering and Huoviala, 2023), and information extraction (Wei et al., 2024; Dagdelen et al., 2024).

This raises a compelling question: Can LLMs effectively extract maritime safety terminology in a zero-shot setting, given the domain's complex multi-component terms, nested terms, specialized acronyms, and context-dependent expressions? To answer this question, the present study evaluates the zero-shot performance of state-of-the-art (SOTA) LLMs in maritime safety terminology extraction, focusing on their ability to handle these domain-specific challenges.

2 Related Work

Advances in pre-trained language models have transformed ATE tasks. Early studies focused on fine-tuning BERT and its variants for terminology extraction. For example, Hazem et al. (2020) evaluated BERT-based models in the TermEval 2020 shared task[3], achieving an F1 score of 46.66% for English and 48.15% for French on the heart failure test set with named entities. Similarly, Hazem et al. (2022) investigated cross-lingual and cross-domain transfer learning for ATE tasks with BERT-based models, showing promising improvements in multilingual settings. Domain-specific applications further validated the potential of pre-trained models. Jerdhaf et al. (2022) compared focused terminology extraction models based on KB-BERT (a generalist Swedish pre-trained model) and SweDeClin-BERT (a domain-specific clinical Swedish model) to identify implant terms in medical records, while Tran et al. (2024b) demonstrated that multilingual models using XLMR improved performance when incorporating less resourced languages such as Slovenian into training data.

LLMs' advanced language understanding has revolutionized NLP tasks, including terminology extraction. Giguere and Iankovskaia (2023) provided one of the first systematic evaluations of GPT-4 for domain-specific terminology extraction, demonstrating impressive accuracy scores across legal (0.84), medical (0.78), and technical (0.73) domains, highlighting LLMs' strong generalization capabilities without task-specific training. In a separate investigation, Banerjee et al. (2024) compared GPT-3.5-Turbo with fine-tuned XLM-RoBERTa in few-shot settings, finding GPT-3.5-Turbo outperformed traditional models with as few as five examples, though performance varied by domain.

Research has also explored LLMs' capabilities in multilingual and cross-lingual scenarios. Tran et al. (2025) introduced LlamATE, a prompting-based framework for automatic term extraction with Llama-2-Chat. Their study shows that LlamATE transfers term extraction knowledge across languages, performing on par with monolingual training, especially in related languages like English, French, and Dutch. Additionally, Tran et al. (2024a) evaluated prompting strategies with open- and closed-source LLMs on the ACTER dataset (Rigouts Terryn et al., 2020). While LLMs achieved high recall (up to 79.6% for Dutch), they struggled with precision, particularly in sequence-labeling task, whereas text-extractive and generative formats improved the recall-precision balance.

The evolution of terminology extraction approaches from BERT-based models to LLMs has demonstrated significant advances in cross-domain and multilingual capabilities. This study advances the SOTA in two key aspects: evaluating LLMs' zero-shot performance in maritime safety terminology extraction and analyzing annotation strategies for optimal terminology coverage in this specialized domain. The empirical findings establish benchmarks for maritime safety terminology extraction while offering insights into terminology extraction approaches and annotation practices for domain-specific applications.

[3]The TermEval 2020 shared task: https://aclanthology.org/2020.computerm-1.12/

3 Corpus Creation

3.1 Annotation scheme

Our annotation scheme followed the ACTER Terminology Annotation Guidelines (Rigouts Terryn, 2021) with domain-specific modifications tailored to maritime safety terminology. The annotation employed the BIO (Beginning, Inside, Outside) (Ramshaw and Marcus, 1995) tagging scheme for sequence labeling, enabling precise identification of term boundaries and multi-word expressions.

Following ACTER's domain- and language-independent classification scheme, terms were categorized into four types based on their lexicon-specificity and domain-specificity characteristics. *Specific Terms* are both lexicon- and domain-specific, representing the strictest definition of maritime safety terminology that requires domain expertise to understand (e.g., *AIS transceiver*, *listening watch*). *Common Terms*, while strongly domain-related, are not highly lexicon-specific, indicating terms that are crucial to maritime safety but may be understood by non-experts (e.g., *lookout*, *crew*). *Out-of-Domain Terms* (OOD Terms) are lexicon-specific but not domain-specific, encompassing technical terminology from related fields that appears in maritime safety documentation (e.g., *differential GPS*, *cognitive tunnelling*). *Named Entities* represent proper names of vessels, organizations, locations, and other domain-relevant proper nouns, which are frequently encountered in the maritime safety domain (e.g., *Port Adelaide*, *Australian Maritime Safety Authority*).

3.2 Gold standard dataset

The gold standard dataset derives from a systematically annotated corpus based on a maritime accident investigation report[4], comprising 10,690 tokens. Manual annotation of this corpus yielded 1,558 term instances, which resolve to 424 unique terms distributed across four categories based on the established annotation scheme. As detailed in Table 1, Common Terms constitute the largest portion at 46.60% of all instances, followed by Specific Terms (31.51%), Named Entities (20.47%), and OOD Terms (1.41%), reflecting the terminological composition of maritime safety documentation. The notably high repetition rates for the primary maritime safety terminology categories (ranging from 64.77% to 79.62%) reflect the standardized nature of maritime safety communication, while the markedly lower repetition of Out-of-Domain Terms (18.18%) suggests their peripheral role in maritime accident reporting.

Term Type	Total	Unique	Repetition (%)
Common Terms	726	168	76.86
Specific Terms	491	173	64.77
Named Entities	319	65	79.62
OOD Terms	22	18	18.18
Total	1,558	424	72.79

Table 1: Term distribution of the gold standard.

The annotation process employed a two-phase approach. The initial phase involved nested annotation, which identifies hierarchical relationships within complex maritime safety terms. For instance, in the term *bridge navigational watch alarm system*, both the complete term and its constituent components *bridge*, *watch*, and *watch alarm system* were annotated as valid terms, indicating their hierarchical relationship. The subsequent phase transformed nested annotation into a full-term format, focusing exclusively on the maximal-length terms while excluding shorter terms nested within them. For example, given *starboard bridge wing*, the full-term annotation retains only this complete term, excluding nested terms such as *starboard* and *bridge* within the larger term.

3.3 Inter-annotator agreement evaluation

The inter-annotator agreement evaluation was conducted between two annotators: a maritime language specialist and an annotation-trained linguistics researcher who performed their annotations independently. The assessment of both annotation strategies evaluated agreement using Cohen's Kappa (Cohen, 1960) for measuring categorical agreement between two raters, and Krippendorff's Alpha (Krippendorff, 2018) for assessing reliability with multiple coders, as presented in Table 2.

The results demonstrate consistently strong agreement across both annotation approaches, with the full-term approach showing marginally higher agreement across all metrics. The confusion matrix

[4]Unlabeled text selected from a marine accident investigation report by the Australian Transport Safety Bureau (ATSB). This type of report was selected because marine accident investigation reports document safety incidents and provide comprehensive descriptions of maritime safety systems, operational procedures, and regulatory frameworks governed by international conventions such as SOLAS and STCW. https://www.atsb.gov.au/sites/default/files/media/5780376/mo-2020-001-final.pdf

Metric	Nested	Full-term	Difference
Cohen's Kappa	0.7207	0.7250	+0.0043
Krippendorff's Alpha	0.7463	0.7466	+0.0003

Table 2: Inter-annotator agreement comparison between nested and full-term annotation approaches

visualization (Figure 1) provides detailed insights into annotation patterns for the full-term approach, revealing strong agreement in term identification with 1,154 matching B-tag annotations and 441 matching I-tag annotations.

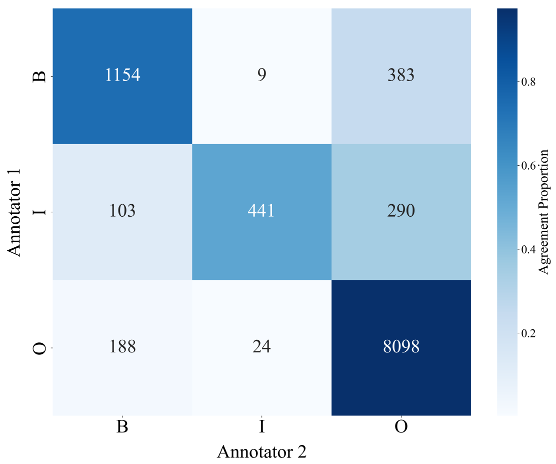

Figure 1: Inter-annotator confusion matrix.

The tag-wise performance analysis revealed strong boundary detection capabilities across annotators, with the maritime language specialist severing as the reference annotator. Analysis showed balanced precision (0.7986) and recall (0.7464) for the B-tag, while the I-tag demonstrated high precision (0.9304). These metrics indicate that annotators, particularly the annotation-trained linguistics researcher when compared against the reference, were conservative in marking term continuations, prioritizing precision over recall. These results validate the annotation guideline effectiveness and establish a solid foundation for the subsequent terminology extraction experiments.

4 Experimental Setup

4.1 Baseline

The baseline was established through BERT model fine-tuning on ACTER's English BIO-annotated data (over 200,000 tokens) (Rigouts Terryn et al., 2020). The fine-tuning experiments with BERT-base and BERT-large variants revealed BERT-base's superior performance, leading to its selection as the baseline for subsequent comparisons. See Appendix A.1 for experimental details.

4.2 Model selection

For the ATE experiments, state-of-the-art LLMs, comprising both open-source and closed-source variants, were selected based on their recent developments and varying specifications in parameter sizes (Param) and context lengths (CtxL), as detailed in Table 3. Models were chosen for their potential to perform well in zero-shot settings, enabling the evaluation of their inherent capabilities without domain-specific fine-tuning.

Type	Model	Param	CtxL
Open-source	Qwen2.5-7B-Instruct	7.61B	128K
	GLM-4-9B-Chat	9B	128K
	Llama-3.1-8B-Instruct	8B	128K
	gemma-2-9b-it	9B	8K
Closed-source	Claude-3.5-Sonnet[5]	Undisclosed	200K
	Gemini-1.5-Pro	Undisclosed	2,000K[6]
	GPT-4o-latest[7]	Undisclosed	128K

Table 3: Specifications of the evaluated models.

4.3 Zero-shot prompt design

The zero-shot prompt for maritime safety terminology extraction was initially developed based on the ReAct prompting framework (Yao et al., 2023) and Chain-of-Thought (CoT) principles (Wei et al., 2022). Its design followed the domain-specific adaptations outlined in Section 3.1.

The prompt was then iteratively refined through pilot studies across multiple models, focusing on evaluating their instruction-following capabilities and optimizing the design accordingly. This process established clear input–output formatting for consistent sequence labeling evaluation. The final prompt demonstrated reliable performance in guiding model behavior for maritime safety ATE. See Appendix A.2 for the full prompt specification.

[5] Model version: Claude-3.5-Sonnet (2024-10-22)

[6] Gemini-1.5-Pro supports a maximum context length of 2M tokens. However, Google API offers lower pricing for prompts under 128K tokens, which are more commonly used.

[7] Model version: GPT-4o-latest (2024-11-20)

Category	Model	Term-level			Label B			Label I		
		F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall
Baseline	BERT-base Fine-tuning	0.1683	0.3484	0.1110	0.1972	0.3267	0.1412	0.1174	0.4342	0.0679
Open-source	Qwen2.5-7B-Instruct	**0.5206**	0.6364	0.4404	0.5801	0.7326	0.4801	0.3912	0.4470	0.3478
	GLM-4-9B-Chat	**0.5455**	0.4491	0.6944	0.5982	0.4937	0.7587	0.4238	0.3470	0.5442
	Llama-3.1-8B-Instruct	0.2818	0.3072	0.2602	0.3256	0.3271	0.3241	0.1468	0.2170	0.1109
	gemma-2-9b-it	0.4602	0.4010	0.5398	0.5060	0.4398	0.5956	0.3518	0.3085	0.4093
Closed-source	Claude-3.5-Sonnet	**0.8002**	0.7171	0.9049	**0.8260**	0.7398	0.9350	**0.7395**	0.6639	0.8346
	Gemini-1.5-Pro	0.4986	0.5648	0.4463	0.5735	0.5603	0.5873	0.1970	0.6240	0.1169
	GPT-4o-latest	0.4374	0.5192	0.3779	0.4830	0.5946	0.4066	0.3369	0.3706	0.3087

Table 4: Comparison of overall model performance in maritime safety terminology extraction. Term-level F1-score measures overall extraction accuracy, while Label B and Label I F1-scores evaluate term boundary detection.

5 Results and Discussion

5.1 Overall model performance

To evaluate the overall effectiveness of different LLMs in maritime safety terminology extraction, we compare their term-level Precision, Recall, and F1-score, as well as their performance in boundary detection (separate Label B & Label I F1-score). Table 4 summarizes the results.

The results indicate that Claude-3.5-Sonnet outperforms all evaluated models, with its F1-score (0.8002) significantly exceeding the BERT-base fine-tuning baseline (0.1683) and demonstrating the advantages of zero-shot learning in maritime safety terminology extraction. Claude-3.5-Sonnet's superior performance may be attributed to its advanced reasoning capabilities and extensive pre-training on diverse technical domains, enabling better understanding of specialized terminology patterns. The model shows particularly strong performance in identifying term beginnings (Label B F1-score: 0.8260) compared to term continuations (Label I F1-score: 0.7395).

Among open-source models, GLM-4-9B-Chat shows relatively higher recall (0.6944) but at the cost of lower precision (0.4491), whereas Qwen2.5-7B-Instruct achieves higher precision (0.6364) but struggles with recall (0.4404). Llama-3.1-8B-Instruct performs the worst among LLMs (Term-level F1-score: 0.2818), approaching the baseline performance level, with a particular weakness in term continuation detection (Label I F1-score: 0.1468).

To examine the generalization capabilities of the top-performing models, additional experiments were conducted on ACTER's *Wind Energy* subset (57,766 tokens) (Rigouts Terryn et al., 2022a). The three models with the highest F1-scores in maritime safety terminology extraction - Claude-3.5-Sonnet, Qwen2.5-7B-Instruct, and GLM-4-9B-Chat - were evaluated. Table 5 presents the comparative results.

Model	F1-score	Precision	Recall
Claude-3.5-Sonnet	0.6775	0.6215	0.7446
Qwen2.5-7B-Instruct	0.5038	0.5425	0.4703
GLM-4-9B-Chat	0.4685	0.3979	0.5694

Table 5: Term-level performance (F1-score) on ACTER's *Wind Energy* domain subset.

The results on this standardized benchmark align with the performance patterns observed in the maritime domain. The relative ranking of models remains consistent across domains, suggesting their capabilities in specialized terminology extraction generalize beyond maritime safety. This cross-domain validation provides additional evidence for the effectiveness of these models in sequence labeling tasks across different technical domains.

5.2 Performance across different term types

Table 6 presents the F1-scores across four term categories, with category-specific performance calculated by mapping the BIO-tagged sequences to their corresponding term types using the gold standard classification. Claude-3.5-Sonnet maintains its superior performance across Common Terms (F1-score of 0.7149) and Specific Terms (F1-score of 0.6102), aligning with its strong overall term-level metrics. Among open-source models, Qwen2.5-7B-

Category	Model	Common Terms	Specific Terms	Named Entities	OOD Terms
Open-source	Qwen2.5-7B-Instruct	**0.5058**	0.4466	**0.5418**	0.0310
	GLM-4-9B-Chat	0.3983	0.2997	0.2456	0.0142
	Llama-3.1-8B-Instruct	0.2961	0.1454	0.1152	0.0071
	gemma-2-9b-it	0.3680	0.2232	0.2360	0.0106
Closed-source	Claude-3.5-Sonnet	**0.7149**	**0.6102**	**0.5304**	0.0708
	Gemini-1.5-Pro	**0.5028**	0.3108	0.3170	0.0099
	GPT-4o-latest	0.4055	0.3483	0.3628	0.0142

Table 6: Performance by term category (F1-score)

Instruct demonstrates balanced performance across Common Terms (F1-score of 0.5058) and named entities (F1-score of 0.5418), though notably lower than Claude-3.5-Sonnet's results.

All models show significantly reduced performance in OOD Terms, with even the best-performing Claude-3.5-Sonnet achieving only 0.0708. This consistent pattern across models suggests inherent challenges in distinguishing domain-adjacent technical terminology through zero-shot learning, particularly given the limited OOD Terms in the gold standard (18 unique terms) and the complexity of applying domain-specificity criteria through prompt-based instructions.

5.3 Terminology coverage analysis

A key consideration in terminology extraction is the choice between full-term annotation and nested annotation approaches. The full-term annotation approach captures entire domain-specific terms as single units, whereas the nested annotation approach also identifies sub-components within multi-word expressions. To quantify the impact of these strategies, Table 7 provides a comparison of extracted terms by annotation approach. Full-term annotation identified 420 unique terms, while the nested annotation approach extracted an additional 21 unique terms, resulting in a total of 441 distinct terms. This corresponds to a coverage rate of 95.24% for full-term annotation approach, with the additional 4.76% of unique terms contributed by nested annotation approach.

Compared to the full-term annotation approach, the additional terms identified through the nested annotation approach consist primarily of sub-components from larger technical expressions (e.g., *pilotage* from *pilotage exemption certificate*, *mooring station* from *forward mooring station*, and *AIS stations* from *shore-based AIS stations*) rather than new conceptual entities. This suggests that nested annotation primarily decomposes existing multi-word terms rather than expanding the terminological inventory of the domain.

Term Type	Nested	Full-term	Exclusive (Nested)
Common Terms	175	164	11
Specific Terms	181	173	8
Named Entities	66	65	1
OOD Terms	19	18	1
Total	441	420	21

Table 7: Comparison of extracted terms by annotation approach

This modest gain in terminology coverage, particularly in structured domains such as maritime safety, indicates that the full-term annotation approach adequately captures domain knowledge. Building on these findings, the full-term annotation approach can serve as a viable choice for large-scale ATE tasks in well-defined domains, where efficiency and practicality are crucial considerations. This approach offers comprehensive coverage without significant information loss while also reducing the additional complexity introduced by the nested annotation approach. The complete list of additional terms identified through nested annotation approach is provided in Appendix A.3.

5.4 Error analysis

To better understand the challenges faced by different models in maritime safety terminology extraction, a detailed error analysis was conducted by examining false positives (FPs) and false negatives (FNs) across all models against the gold standard of 1,558 term instances. FPs refer to terms extracted by the model that are not in the gold standard, while FNs indicate valid terms from the gold standard that the model failed to detect. Table 8 summarizes the FPs and FNs counts for each model.

Model	FPs	FNs
Qwen2.5-7B-Instruct	361	351
GLM-4-9B-Chat	1,380	182
Llama-3.1-8B-Instruct	1,104	484
gemma-2-9b-it	1,299	251
Claude-3.5-Sonnet	522	95
Gemini-1.5-Pro	781	579
GPT-4o-latest	403	421

Table 8: False positives and false negatives per model

5.4.1 Quantitative analysis of errors

The performed error analysis reveals distinct error patterns across models. There is considerable variation in performance among open-source models, with Qwen2.5-7B-Instruct demonstrating controlled performance while GLM-4-9B-Chat and gemma-2-9b-it show a significant tendency towards overgeneration with a high number of FPs (1,380 and 1,299 respectively), indicating challenges in precise term boundary detection. Among closed-source models, Claude-3.5-Sonnet notably minimizes FNs (95) despite generating more FPs (522), suggesting a stronger capability in identifying valid maritime safety terms, though with some tendency to over-include general domain language. In contrast, both Gemini-1.5-Pro and GPT-4o-latest exhibit higher error rates across both categories, reflecting greater difficulty in accurate maritime safety terminology extraction.

5.4.2 False positives analysis

The analysis of false positives reveals two primary causes of errors: (1) contextual descriptions tagged as terms, and (2) boundary detection errors.

Misidentification of contextual descriptions as terminology Models with higher FPs extract general descriptive phrases that lack terminological status, leading to outputs that contain contextual metadata rather than precise domain-specific terms. For example, GLM-4-9B-Chat extracts operational descriptions such as *all ships between 10 and 24 m in length*, *vessels over 150 gross tonnage*, and *collisions between trading ships and small vessels*. Similarly, gemma-2-9b-it misidentifies navigational contexts like *passage within the harbour in darkness*, *events and conditions that increase risk*, *requirement to record hours of rest*, and *risk of the collision posed by the other*. Qwen2.5-7B-Instruct also demonstrates similar errors, such as *ATSB investigation report findings focus on safety factors*. These cases highlight the models' difficulty in distinguishing between contextual information and specialized maritime safety terminology.

Boundary detection failures in term extraction A key challenge across models is the inability to establish precise term boundaries, resulting in two common failure modes: overextension (inclusion of extraneous contextual elements) and truncation (omission of critical components).

Overextension occurs when models extract phrases that incorporate unnecessary surrounding contextual elements, rather than isolating the precise technical term. This differs from the previously discussed contextual misidentification issue, as the extracted text still contains a valid term but includes additional, non-essential elements. Examples include GLM-4-9B-Chat producing errors such as *Accolade II for the previous*, *chief mate 's*, *helmsman changed*, *acquiring AIS icons*, and *ATSB 's investigation*. Similarly, gemma-2-9b-it generates erroneous phrases like *an automatic identification system*, *anchorage within the port*, and *port 's working channel*, *skipper altered*. Llama-3.1-8B-Instruct exhibits overextended terms such as *on VHF*, *skipper called*, and *chief mate was*, while GPT-4o-latest demonstrates this issue with *Adelaide 's rules*, *and starboard*, and *'s starboard*.

Truncation, though less frequent, results in partial extraction of technical terms, leading to ambiguous or incorrect outputs. Examples include GLM-4-9B-Chat extracting *Australian Maritime Safety* instead of the correct *Australian Maritime Safety Authority*, and *automatic* instead of the complete terms *Automatic Identification System* or *Automatic Radar Plotting Aid*. Similarly, gemma-2-9b-it produces errors like *class 3B* (instead of *class 3B DCV*, *class 3B domestic commercial vessel*, or *class 3B vessels*) and *shuttle belt* instead of *shuttle belt conveyor*. Gemini-1.5-Pro also exhibits truncation errors, such as extracting *pre-departure* instead of *bridge pre-departure checklist* or *bridge pre-departure checks*.

5.4.3 False negatives analysis

The analysis of false negatives also reveals two fundamental challenges in maritime safety terminology extraction: (1) the extraction of multi-word terms, and (2) terms requiring domain-specific professional knowledge.

Difficulties in recognizing multi-component maritime safety terminology Models exhibit significant deficiencies in extracting terms that integrate multiple interrelated maritime concepts, often failing to capture complex multi-word expressions while correctly identifying their shorter variants. For example, all models struggle with full terms like *Standards of Training, Certification and Watchkeeping for Seafarers*, and *International Convention for the Safety of Life at Sea*, despite correctly recognizing their corresponding acronyms (STCW and SOLAS).

Beyond regulatory terms, similar challenges arise in navigation safety systems. Llama-3.1-8B-Instruct frequently misses complex multi-component terms such as *vessel traffic service*, *vessel monitoring system*, and *bridge navigational watch alarm system*. Likewise, gemma-2-9b-it fails to extract *safety management system*, *fatigue management system*, and *vessel monitoring system*. These errors indicate difficulty in capturing domain-specific multi-word expressions where foundational maritime terms (e.g., *vessel*, *bridge*, *system*) are embedded within functional descriptors (e.g., *traffic service*, *monitoring*).

The issue extends to maritime documentation and regulatory terminology, where models frequently omit terms reflecting multi-level administrative structures. GPT-4o-latest fails to extract *minimum safe manning document* and *AMSA-issued certificate of operation*, while Qwen2.5-7B-Instruct misses *certificate of competency* and *certificate of recognition*. Similarly, Gemini-1.5-Pro fails to extract regulatory terms such as *pilotage exemption certificate* and *national standard for commercial vessels*, and Llama-3.1-8B-Instruct struggles with operational communication terms like *bridge resource management practices* and *bridge procedures guide*. These cases illustrate broader limitations in handling structured terminology within maritime regulatory and operational contexts.

Limited understanding of domain-specific professional knowledge Models also struggle with terms that require deeper professional knowledge within the maritime domain. Even Claude-3.5-Sonnet, despite its relatively strong performance, occasionally fails to extract highly specialized terms such as *dual-fuel propulsion engines* and *automatic radar plotting aid*. GPT-4o-latest frequently misses *blind sectors*, *visual obstruction*, *voyage data recorder*, and *compass bearings*. Similarly, gemma-2-9b-it fails to recognize *echo sounder*, *gyrocompass*, and *radar echoes*, while Llama-3.1-8B-Instruct overlooks critical safety terms like *close-quarters situation* and *proper lookout*. These examples suggest that models struggle with extracting technical terminology that necessitates deeper comprehension of maritime operations and navigational safety principles.

6 Conclusion

This study evaluates LLMs' capabilities in maritime safety terminology extraction in zero-shot settings. The experimental results demonstrate that while LLMs show promise in handling maritime safety terminology, their performance varies significantly between closed-source and open-source implementations. The analysis further reveals that full-term annotation achieves comprehensive terminology coverage while maintaining practical efficiency, challenging assumptions about annotation granularity requirements in domain-specific terminology extraction.

The findings suggest several promising directions for future research. Few-shot learning represents a compelling direction for maritime safety terminology extraction, where models could be prompted with a small set of carefully selected examples. This approach could explore how different prompt engineering strategies and example selection methods affect terminology identification performance, particularly for handling specialized maritime safety concepts. Fine-tuning approaches offer another valuable research direction, particularly through cross-lingual and cross-domain adaptation. This direction could leverage well-established terminology datasets such as ACTER (Rigouts Terryn et al., 2020), GENIA (Kim et al., 2003), and ACL RD-TEC 2.0 (QasemiZadeh and Schumann, 2016) for initial model adaptation. Further fine-tuning could then be performed using bilingual maritime-specific resources, potentially revealing insights into the transferability of terminology extraction capabilities across technical domains and languages.

Acknowledgments

We gratefully acknowledge support from the China Scholarship Council for the 2024 National Construction of High-Level University Graduate Research Program (Project No. 202408310292). We also thank the three anonymous reviewers for their constructive feedback.

Supplementary material

The supplementary material, including all appendices mentioned in the main text, is uploaded onto GitHub https://github.com/Ethan-Liu-Ethan/MariATE_RANLP_2025

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. *Preprint*, arXiv:2205.12689.

Dhananjay Ashok and Zachary C. Lipton. 2023. PromptNER: Prompting for named entity recognition. *Preprint*, arXiv:2305.15444.

Shubhanker Banerjee, Bharathi Raja Chakravarthi, and John Philip McCrae. 2022. A dataset for term extraction in Hindi. In *Proceedings of the Workshop on Terminology in the 21st century: many faces, many places*, pages 19–25, Marseille, France. European Language Resources Association.

Shubhanker Banerjee, Bharathi Raja Chakravarthi, and John Philip McCrae. 2024. Large language models for few-shot automatic term extraction. In Amon Rapp, Luigi Di Caro, Farid Meziane, and Vijayan Sugumaran, editors, *Natural Language Processing and Information Systems*, volume 14762, pages 137–150. Springer Nature Switzerland, Cham.

Elena Bolshakova, Natalia Loukachevitch, and Michael Nokel. 2013. Topic models can improve domain term extraction. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, Stefan Rüger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz, editors, *Advances in Information Retrieval*, volume 7814, pages 684–687. Springer Berlin Heidelberg, Berlin, Heidelberg.

Youngjin Chae and Thomas Davidson. 2023. Large language models for text classification: From zero-shot learning to fine-tuning. *Open Science Foundation*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Damien Cram and Béatrice Daille. 2016. Terminology extraction with term variant detection. In *Proceedings of ACL-2016 System Demonstrations*, pages 13–18, Berlin, Germany. Association for Computational Linguistics.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

Orphée De Clercq, Marjan Van de Kauter, Els Lefever, and Véronique Hoste. 2015. LT3: Applying Hybrid Terminology Extraction to Aspect-Based Sentiment Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 719–724, Denver, Colorado. Association for Computational Linguistics.

Frantzi Katerina, Ananiadou Sophia, and Mima Hideki. 2000. Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Julie Giguere and Anna Iankovskaia. 2023. Leveraging large language models to extract terminology. In *Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications*, pages 57–60, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Amir Hazem, Mérième Bouhandi, Florian Boudin, and Beatrice Daille. 2022. Cross-lingual and cross-domain transfer learning for automatic term extraction from low resource data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 648–662, Marseille, France. European Language Resources Association.

Amir Hazem, Mérième Bouhandi, Florian Boudin, and Béatrice Daille. 2020. TermEval 2020: TALN-LS2N system for automatic term extraction. In *Proceedings of the 6th International Workshop on Computational Terminology*, Marseille, France.

Rizwan Iqbal, Masrah Azrifah Azmi Murad, Aida Mustapha, and Nurfadhlina Mohd. Sharef. 2017. An ontology development approach using concept maps driven by automatic term extraction. *International Journal of Information and Communication Technology*, 10(1):51.

Tanja Ivanović, Ranka Stanković, Branislava Šandrih Todorović, and Cvetana Krstev. 2022. Corpus-based bilingual terminology extraction in the power engineering domain. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 28(2):228–263.

Oskar Jerdhaf, Marina Santini, Peter Lundberg, Tomas Bjerner, Yosef Al-Abasse, Arne Jonsson, and Thomas Vakili. 2022. Evaluating pre-trained language models for focused terminology extraction from Swedish medical records. In *Proceedings of the Workshop on Terminology in the 21st century: many faces, many places*, pages 30–32, Marseille, France. European Language Resources Association.

J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. SAGE Publications.

Lieve Macken, Els Lefever, and Véronique Hoste. 2013. TExSIS: Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 19(1):1–30.

Diana Maynard and Sophia Ananiadou. 2000. Identifying terms by their family and friends. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Hiroshi Nakagawa and Tatsunori Mori. 2002. A Simple but powerful automatic term extraction method. In *COLING-02: COMPUTERM 2002: Second International Workshop on Computational Terminology*.

Anselmo Peñas, Felisa Verdejo, and Julio Gonzalo. 2001. Corpus-based terminology extraction applied to information access. In *Proceedings of Corpus Linguistics (2001)*.

Behrang QasemiZadeh and Anne-Kathrin Schumann. 2016. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1862–1868, Portorož, Slovenia. European Language Resources Association (ELRA).

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text Chunking using Transformation-Based Learning. *Preprint*, arXiv:9505040.

Alya Rigouts Terryn, Véronique Hoste, and Els Levefer. 2020. In no uncertain terms: A dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Language Resources and Evaluation*, 54(2):385–418.

Ayla Rigouts Terryn. 2021. ACTER terminology annotation guidelines.

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2021. Hybrid Adaptable Machine Learning approach to Extract Terminology. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 27(2):254–293.

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2022a. ACTER 1.5: Annotated Corpora for Term Extraction Research. In *CLARIN Annual Conference 2022*, Prague.

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2022b. Tagging terms in text: A supervised sequential labelling approach to automatic term extraction. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 28(1):157–189.

Paul F. Simmering and Paavo Huoviala. 2023. Large language models for aspect-based sentiment analysis. *Preprint*, arXiv:2310.18025.

Hanh Thi Hong Tran, Carlos-Emiliano González-Gallardo, Julien Delaunay, Antoine Doucet, and Senja Pollak. 2024a. Is prompting what term extraction needs? In Elmar Nöth, Aleš Horák, and Petr Sojka, editors, *Text, Speech, and Dialogue*, volume 15048, pages 17–29. Springer Nature Switzerland, Cham.

Hanh Thi Hong Tran, Carlos-Emiliano González-Gallardo, Antoine Doucet, and Senja Pollak. 2025. LlamATE: Automated term extraction using large-scale generative language models. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 31(1):5–36.

Hanh Thi Hong Tran, Matej Martinc, Andraz Pelicon, Antoine Doucet, and Senja Pollak. 2022a. Ensembling Transformers for cross-domain automatic term extraction. In *From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries*, pages 90–100, Cham. Springer International Publishing.

Hanh Thi Hong Tran, Matej Martinc, Andraz Repar, Antoine Doucet, and Senja Pollak. 2022b. A Transformer-based sequence-labeling approach to the Slovenian cross-domain automatic term extraction. In *Conference on Language Technologies & Digital Humanities*, Ljubljana, Slovenia.

Hanh Thi Hong Tran, Matej Martinc, Andraz Repar, Nikola Ljubešić, Antoine Doucet, and Senja Pollak. 2024b. Can cross-domain term extraction benefit from cross-lingual transfer and nested term labeling? *Machine Learning*, 113(7):4285–4314.

Jorge Vivaldi and Horacio Rodríguez. 2007. Evaluation of terms and term extraction systems: A practical approach. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 13(2):225–248.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. GPT-NER: Named entity recognition via large language models. *Preprint*, arXiv:2304.10428.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2024. ChatIE: Zero-shot information extraction via chatting with ChatGPT. *Preprint*, arXiv:2302.10205.

Petra Wolf, Ulrike Bernardi, Christian Federmann, and Sabine Hunsicker. 2011. From statistical term extraction to hybrid machine translation. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, Leuven, Belgium. European Association for Machine Translation.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. *Preprint*, arXiv:2104.08826.

Yu Yuan, Yuze Gao, Yue Zhang, and Serge Sharoff. 2018. Cross-lingual terminology extraction for translation qualityestimation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Vanni Zavarella, Juan Carlos Gamero-Salinas, and Sergio Consoli. 2024. A few-shot approach for relation extraction domain adaptation using large language models. *Preprint*, arXiv:2408.02377.