

# Construction of a Maritime Knowledge Graph Using GraphRAG for Entity and Relationship Extraction from Maritime Documents

Yi Han<sup>1\*</sup>, Tao Yang<sup>1</sup>, Meng Yuan<sup>2</sup>, Pinghua Hu<sup>2</sup>, Chen Li<sup>2</sup>

<sup>1</sup>COSCO Shipping Technology Co., Ltd., Shanghai, China

<sup>2</sup>COSCO Shipping Specialized Carriers Co., Ltd., Guangzhou, China

Email: \*ywlin2024@163.com

**How to cite this paper:** Han, Y., Yang, T., Yuan, M., Hu, P.H. and Li, C. (2025) Construction of a Maritime Knowledge Graph Using GraphRAG for Entity and Relationship Extraction from Maritime Documents. *Journal of Computer and Communications*, 13, 68-93.

<https://doi.org/10.4236/jcc.2025.132006>

**Received:** December 19, 2024

**Accepted:** February 21, 2025

**Published:** February 24, 2025

Copyright © 2025 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## Abstract

In the international shipping industry, digital intelligence transformation has become essential, with both governments and enterprises actively working to integrate diverse datasets. The domain of maritime and shipping is characterized by a vast array of document types, filled with complex, large-scale, and often chaotic knowledge and relationships. Effectively managing these documents is crucial for developing a Large Language Model (LLM) in the maritime domain, enabling practitioners to access and leverage valuable information. A Knowledge Graph (KG) offers a state-of-the-art solution for enhancing knowledge retrieval, providing more accurate responses and enabling context-aware reasoning. This paper presents a framework for utilizing maritime and shipping documents to construct a knowledge graph using GraphRAG, a hybrid tool combining graph-based retrieval and generation capabilities. The extraction of entities and relationships from these documents and the KG construction process are detailed. Furthermore, the KG is integrated with an LLM to develop a Q&A system, demonstrating that the system significantly improves answer accuracy compared to traditional LLMs. Additionally, the KG construction process is up to 50% faster than conventional LLM-based approaches, underscoring the efficiency of our method. This study provides a promising approach to digital intelligence in shipping, advancing knowledge accessibility and decision-making.

## Keywords

Maritime Knowledge Graph, GraphRAG, Entity and Relationship Extraction, Document Management

## 1. Introduction

In the modern maritime industry, with the increasing complexity of global supply chains and intensified competition, digital intelligence transformation has become a key strategy to enhance operational efficiency, optimize resource utilization, and address environmental and regulatory challenges [1] [2]. Governments and leading enterprises worldwide are driving data integration and intelligent applications, aiming to leverage advanced data management and analytics to navigate vast and complex business information landscapes [3]. However, data management requirements in the maritime field far exceed those in other industries. This sector not only requires basic information on vessel operations, cargo movement, and route planning but also covers highly specialized data on weather, environmental impact, and regulatory compliance, making traditional data management approaches insufficient for handling these complexities.

The rapid advancement of artificial intelligence (AI), especially with the advent of large language models like GPT, has opened new possibilities for data management within the maritime industry. These models process vast amounts of data and offer sophisticated language comprehension, laying a foundation for intelligent transformation in the industry. However, existing large models often lack deep adaptation for specific industries and may fall short in understanding the intricacies and expertise of maritime data, limiting their direct applicability in maritime business processes. Thus, combining general AI models with the specific needs of the maritime industry is essential, requiring tailored technical solutions to enhance model professionalism and applicability [4].

In recent years, some leading enterprises, such as COSCO Shipping, have pioneered intelligent solutions based on maritime big data. Their research teams have made significant progress in developing customized AI tools that support a wide range of applications, including customer service, asset management, and vessel scheduling. These solutions leverage natural language processing and machine learning, tuning and specializing large models to fit the maritime industry's unique characteristics. As a result, these models can understand and adapt to complex maritime business logic, significantly improving data processing efficiency and business responsiveness.

Knowledge Graph (KG) technology, first introduced by Google in 2012 [5], represents a transformative tool for information management and semantic analysis with broad potential applications in the maritime domain. KGs structure vast amounts of data, transforming fragmented information into coherent, interconnected knowledge networks. This structured approach empowers AI models to access and reason over business information with enhanced speed and accuracy.

In the shipping industry, KGs enable the integration of diverse entities, such as vessels, ports, routes, cargo, and environmental data. This interconnected network supports companies in making more precise operational decisions, ensuring regulatory compliance, optimizing shipping costs, and advancing environmental sustainability [6] [7].

In recent years, KGs have achieved remarkable progress across multiple sectors, including traffic safety management, industrial control systems, intelligent maritime operations, cultural heritage preservation, and regulatory legal systems.

First, in traffic safety management, the literature [8] conducted an in-depth analysis of traffic accident data through a knowledge graph, building a graph that incorporates the four dimensions of “people, vehicles, roads, and environment”. This graph enables visualized analysis of accident classifications and related paths, providing decision support for traffic management departments and enhancing understanding of accident features and causality, thereby supporting traffic safety management. The literature [9] explores the use of knowledge graphs for safety situation awareness in industrial control systems, systematically reviewing critical technologies, such as knowledge representation, storage, and reasoning, and highlighting their potential in real-time monitoring and risk prediction, providing future directions for applications.

In port state control (PSC) inspections, the literature [10] proposes a knowledge graph-based recommendation method for inspection items, specifically for the unique needs of LNG (Liquefied Natural Gas) carrier inspections. By integrating knowledge graph embeddings with historical data, the study accurately predicts potential deficiencies and recommends suitable inspection items. This approach enhances the efficiency and accuracy of PSC inspections, offering new solutions for safety inspections in high-risk maritime environments.

In the digitalization and intelligent transformation of the shipping industry, knowledge graphs are equally crucial. The literature [8] uses the CiteSpace tool to analyze the trends of intelligent and green developments in global shipping, identifying those technologies such as knowledge graphs, IoT, and big data will play essential roles in future intelligent shipping. Further, the literature [11] proposes a spatio-temporal multigraph convolutional network (STMGCN) using AIS data to generate a maritime traffic network graph. This model uses convolutional layers to capture spatial and temporal patterns in traffic flow, providing fine-grained predictions of maritime traffic and supporting data-driven, intelligent maritime management systems.

In marine environmental protection, the literature [12] developed a marine pollution regulation knowledge graph using deep learning to support port inspection officers’ decision-making. Using the BERT model for multi-relation extraction and named entity recognition, the knowledge graph stored in a Neo4j database enhances the accessibility and efficiency of pollution control regulation queries, serving as an effective tool for compliance and environmental management. The literature [13], on the other hand, constructed a maritime legal knowledge graph, using a BERT + BiLSTM + CRF model for named entity recognition and the DeepKE toolkit for relation extraction. This research significantly improves maritime legal information management, providing technological support for intelligent regulation queries and compliance reviews.

In the area of cultural heritage preservation and historical data management,

the literature [14] presents a knowledge graph method combined with a super-group algorithm to enhance interactivity and user satisfaction in museum digital display platforms. By analyzing visitor behavior with the K-means algorithm and embedding knowledge graphs into museum content, the study optimizes display effectiveness and user experience. The literature [15] developed a Dutch maritime history knowledge graph, integrating four historical datasets into a sustainable source of knowledge that supports research on the Dutch East India Company's daily operations. Following FAIR data management principles, the knowledge graph ensures accessibility and extensibility of historical data.

Additionally, knowledge graphs have achieved breakthroughs in intelligent maritime supervision. The literature [16] introduces a semantic network method based on knowledge graphs to detect illegal maritime activities, specifically in cases of counterfeit ship licenses. By constructing a ship relationship graph and employing semantic reasoning, this study significantly enhances the accuracy and efficiency of detecting illegal activities, improving maritime management capabilities.

In conclusion, these studies demonstrate the extensive potential of knowledge graphs (KGs) in complex systems, showcasing diverse technological applications. Future research should focus on optimizing knowledge representation and reasoning within KGs, enhancing multi-source data integration and real-time reasoning, and expanding applications to meet the growing demands for intelligent decision-making.

This study leverages GraphRAG to integrate knowledge graphs with large language models (LLMs), enhancing data processing efficiency and knowledge management in the maritime domain [17]. KGs provide LLMs with rich contextual knowledge, strengthening semantic understanding and reasoning capabilities within maritime-specific scenarios. By closely integrating KGs with LLMs through GraphRAG, an intelligent question-and-answer system has been developed that accurately responds to complex business queries and understands the contextual intent behind user questions, enabling associative reasoning.

Compared to traditional AI models, this system demonstrates notable improvements in accuracy, logical coherence, and response timeliness, offering unique advantages for the maritime industry. This technical framework accelerates KG construction, streamlines data structuring, and provides enterprises with an intelligent, stable, and efficient solution. Ultimately, it supports the maritime industry in achieving digital intelligence transformation within an increasingly complex market landscape.

## 2. Methodology

### 2.1. Data Source

In this paper, official documents published by the International Maritime Organization (IMO) were selected as the primary source for document search. Based on current hot topics in the maritime industry, the theme was set to “Carbon Intensity Indicator (CII)” and the timeframe was limited to the “past five years”. Given

that some documents have been updated, the latest revised versions will be used in this study. To ensure the representativeness of the constructed knowledge graph and the validity of subsequent keyword extraction and visualization analysis, a document screening was conducted, resulting in the selection of four core documents [18]-[21]. These documents cover key provisions such as the definition of CII, calculation rules, and exemption policies, providing authoritative data support and comprehensive regulatory references for this study.

## 2.2. Methodology

In the field of maritime enterprise management, documents are often in various formats, while input requirements for predictive models such as GraphRAG typically demand specific formats. To minimize content loss during format conversion and ensure compatibility with GraphRAG, the following rigorous methodology has been developed:

### 1) Unified Conversion to PDF Format:

To standardize the document environment, all received documents are initially converted into PDF format. This step ensures consistency and universal recognizability for subsequent processing.

### 2) Conversion from PDF to Markdown Using marker-pdf [22]:

The marker-pdf tool is employed to convert each PDF document into Markdown format. Markdown is chosen due to its simplicity, ease of manipulation, and ability to preserve the basic structure and content of the original document, making it an ideal intermediate format.

### 3) Optimization for Handling Large Documents:

Given the limitations of marker-pdf when processing large files—particularly issues related to speed reduction and memory overflow—a document splitting strategy is implemented. Each large PDF document is carefully divided into smaller, manageable sections. This approach not only improves processing speed but also mitigates potential memory issues.

### 4) Conversion from Markdown to TXT Using markdown-txt [23]:

After splitting the documents, a Python script based on markdown-txt is used to convert the Markdown files into TXT format. This conversion step ensures that the documents meet the input specifications required by GraphRAG.

### 5) Organized Storage for GraphRAG Preparation:

All converted TXT files are systematically stored in a designated folder prepared for GraphRAG input. This organization ensures that the documents are easily accessible, ordered correctly, and ready for analysis.

### 6) Preserving Entity Relationships:

As GraphRAG can process multiple documents simultaneously, the splitting method ensures that the relationships between entities within the text remain intact. Through a clear linking or indexing system, these relationships are maintained throughout the splitting process, allowing the model to accurately interpret and connect information across different sections of the documents.

#### 7) Input into GraphRAG for Knowledge Graph Construction:

Finally, the processed documents are input into GraphRAG following the system's workflow, gradually leading to the construction of a knowledge graph.

This methodology not only addresses the technical challenges of format conversion but also ensures the integrity of document content and structure. It facilitates seamless integration with GraphRAG, paving the way for efficient and accurate data analysis.

Through these optimized steps, the documents retain their completeness while being transformed into a format suitable for GraphRAG, supporting advanced knowledge discovery and data analysis.

### 2.3. GraphRAG

GraphRAG, a retrieval-augmented generation (RAG) approach developed by Microsoft, offers an advanced method for transforming unstructured text into structured knowledge graphs. This methodology enhances knowledge retrieval by organizing raw text into a structured graph that captures the relationships, entities, and events described in the documents. The following outlines the core principles and steps involved in Graph RAG, divided into two main parts: Knowledge Graph Construction and Graph RAG Query Process.

#### 2.3.1. Knowledge Graph Construction

The process of creating a knowledge graph from raw text involves several key steps that convert unstructured data into structured information, enabling more efficient querying and information retrieval. These steps are as follows and **Figure 1** provides a visual representation of the workflow:

##### 1) Text Segmentation:

The first step involves dividing long documents into smaller, manageable units known as TextUnits. Each text unit is typically limited to 300 tokens (where a token can represent a word or character). The purpose of this segmentation is to ensure that long texts are broken down into shorter, more processable segments that can be more easily handled by downstream processes. The length of these units can be adjusted based on specific needs or constraints.

##### 2) Entity Extraction:

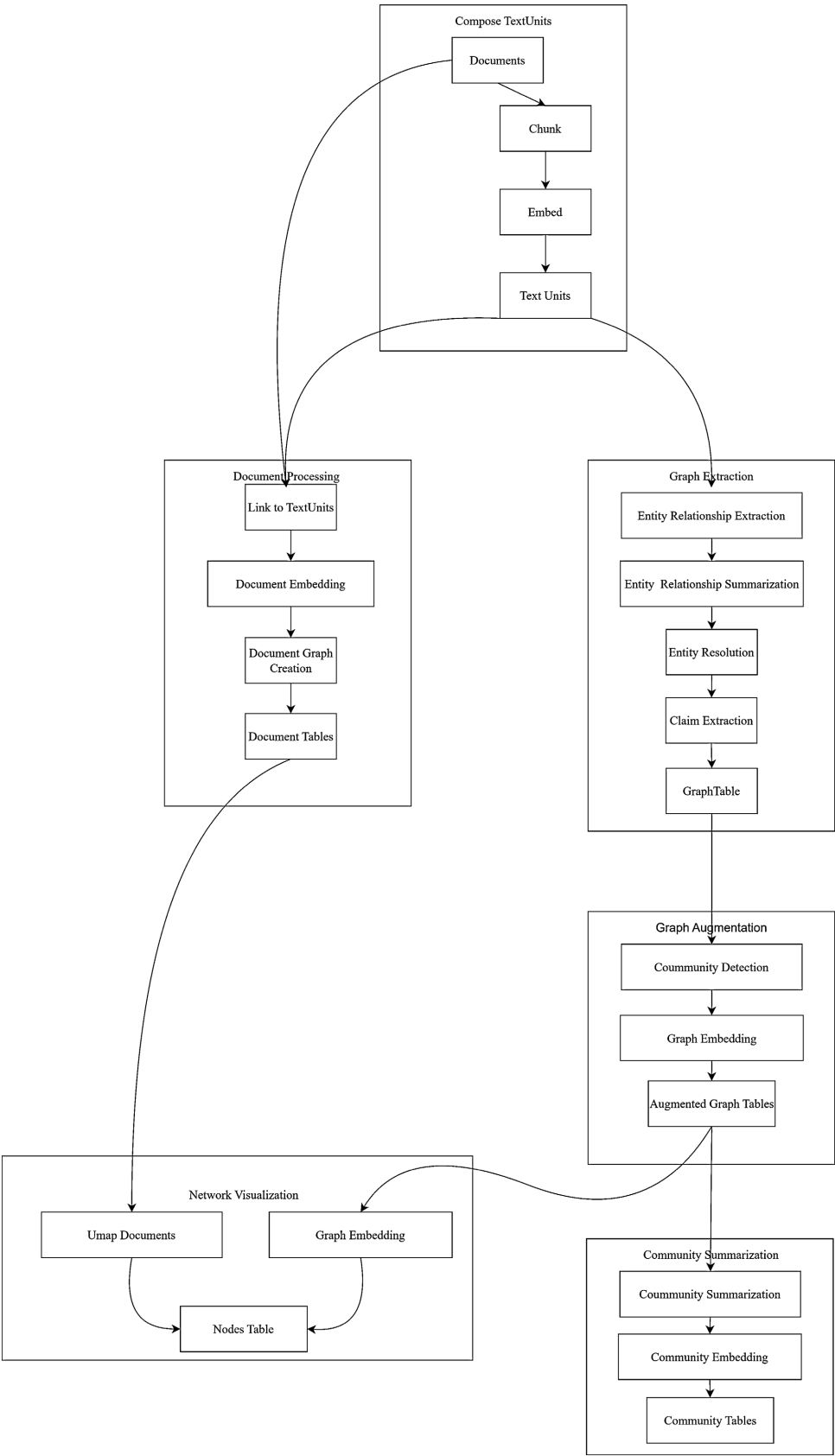
In this step, entities—such as organizations, people, places, and other key concepts—are extracted from each text unit using a large language model (LLM). Each extracted entity is characterized by three key attributes:

Name: The name of the entity.

Type: The category or classification of the entity (e.g., person, organization, location).

Description: A description of the entity, which may vary depending on the context in different text units.

Multiple descriptions may exist for the same entity across different text units, and these are stored as a list within the entity's description attribute. This allows the knowledge graph to capture different facets or views of the same entity.



**Figure 1.** Document processing and knowledge graph construction workflow.

### 3) Relationship Extraction:

Once entities are identified, relationships between them are extracted. A single text unit may describe multiple relationships between different entities. Each relationship is defined by three properties:

Source Entity: The entity from which the relationship originates.

Target Entity: The entity that is the target of the relationship.

Description: A description of the relationship, which can vary depending on the context and the way it is expressed in the text.

Like entity descriptions, relationship descriptions can also exist in multiple forms across different text units, and these variations are stored in a list to ensure that all relevant context is preserved.

### 4) Summarizing Entity and Relationship Descriptions:

Since each entity and relationship may have multiple descriptions, Graph RAG employs a summarization process to condense these into single, more concise statements. For example, if two text units describe the same relationship in different ways, they can be merged into a unified description. This summarization step ensures that only the most important and relevant information is retained, making it easier to query and match knowledge.

### 5) Entity Name Normalization (Optional):

To maintain consistency, entity names that appear in different forms across the text units (e.g., “Entity A” vs. “Entity Alpha”) can be normalized to a single canonical form. This process helps avoid ambiguity and ensures that the same entity is always referred to by the same name throughout the knowledge graph. This step is optional and may involve updates to the relationships or events associated with the entity.

### 6) Event Extraction:

In addition to entities and relationships, Graph RAG extracts events that occur within each text unit. An event typically includes the following attributes:

Initiator: The entity that triggered the event.

Reporter: The entity that reported the event (if applicable).

Event Type: A category or label identifying the type of event.

Event Status: The status of the event (e.g., TRUE for verified, FALSE for false, or SUSPECTED for uncertain events).

Start and End Dates: Temporal markers for the event.

Description: A brief description of the event and its context.

Source: The source of the event information.

Event extraction allows the knowledge graph to capture dynamic occurrences and track the relationships between entities over time.

### 7) Community Classification:

After entities, relationships, and events are extracted, the next step is to group entities into communities based on shared characteristics or contextual relevance. Communities help organize entities in a way that reflects their relationships and roles within broader contexts. For example, entities related to a particular project



or geographical region can be grouped together into a community. These communities allow for more focused querying by enabling users to search within specific contexts.

#### 8) Generating Community Reports:

Each community is represented by a community report, which summarizes the entities, relationships, and events associated with that community. The report includes a description of the community, generated by the LLM. This description serves as a key factor during querying, allowing users to search for knowledge based on community-level attributes.

#### 9) Vectorization of Descriptions:

To enable efficient querying, the descriptions of entities, relationships, and communities are transformed into vector representations (also known as embeddings). These embeddings capture the semantic meaning of the descriptions, allowing for similarity-based retrieval during query processing. The vectorized descriptions are stored as `description_embeddings` and facilitate faster and more accurate matches between queries and knowledge points.

#### 10) Vectorization of Raw Text:

Additionally, the raw text units themselves are vectorized to create a foundational set of vectors for the knowledge base. These raw text vectors are used as the basis for search queries, enabling the system to find the most relevant text segments that match the query.

### 2.3.2. Graph RAG Query Process

Once the knowledge graph is constructed, the next phase involves querying the graph to retrieve relevant knowledge. Graph RAG supports two main query types: Local Search and Global Search, each designed for different use cases and query types.

#### 1) Local Search:

Local search is based on vector similarity matching. When a query is submitted, the system computes the similarity between the query and various elements in the knowledge graph, including entities, relationships, events, communities, and raw text. The most relevant results are selected and presented to the LLM for answer generation.

In contrast to traditional RAG systems, which may only return semantically similar content, Graph RAG enriches the query results by including not just the directly related knowledge, but also associated entities and relationships. This results in more comprehensive and contextually aware answers. Local search is particularly effective for queries that focus on specific entities, such as “Who is the CEO of Company X?”

#### 2) Global Search:

Global search takes a more holistic approach, utilizing an LLM to evaluate which communities are most relevant to the query. This process follows a map-reduce model:

Map: Each community’s description is combined with the query, and the LLM

assesses the relevance of the community to the query.

Reduce: Communities with the highest relevance are selected, and their associated entities, relationships, and events are used to generate a response.

Global search is ideal for more general or summary-oriented queries, such as “What are the main projects in Industry X?” It allows users to retrieve knowledge from a broader, more contextual perspective and is particularly useful for answering complex, multi-faceted questions.

### 3. Case Study

#### 3.1. Maritime Documents Description [18]-[21]

The Marine Environment Protection Committee (MEPC) of the International Maritime Organization (IMO) has recently adopted a series of resolutions and guidelines aimed at reducing the carbon intensity of international shipping. These documents form an important framework for reducing greenhouse gas emissions in the shipping industry, covering the following key areas:

Firstly, Resolution MEPC.355(78) provides provisional guidelines for 2022, outlining the correction factors and voyage adjustment methods used to calculate the operational Carbon Intensity Indicator (CII). The guidelines aim to standardize and effectively implement the relevant provisions of MARPOL Annex VI, providing the industry with adequate preparation. They consider CII correction factors for specific ship types, operational profiles, and voyages, detailing the calculation formulas, including the use of voyage adjustments and correction factors, as well as specific operational scenarios.

Secondly, Resolution MEPC.338(76) focuses on the method for determining the annual operational carbon intensity reduction factors, applicable to the ship types covered under Article 28 of MARPOL Annex VI. This resolution provides specific values for the annual reduction factors from 2023 to 2030, aiming to ensure that by 2030, CO<sub>2</sub> emissions per transport work in international shipping will be reduced by at least 40%.

Next, Resolution MEPC.352(78) provides detailed methods for calculating the operational Carbon Intensity Indicator (CII) of individual vessels, defining both demand-based and supply-based CIIs, and encouraging the trial use of other indicators. This resolution elaborates on the CII calculation formulas, including CO<sub>2</sub> emissions and transport work.

Additionally, Resolution MEPC.353(78) outlines the method for calculating the operational carbon intensity reference line, providing specific carbon intensity reference lines for each ship type covered under Article 28 of MARPOL Annex VI. Resolution MEPC.354(78) offers a method for assigning operational efficiency performance ratings to vessels, based on the deviation between the vessel’s annual CII and the required value, defining the rating boundaries from 2023 to 2030.

Together, these resolutions and guidelines form the authoritative framework for the IMO’s efforts to reduce greenhouse gas emissions in the shipping industry. By monitoring, reporting, and verifying the carbon intensity performance of

vessels, they drive the shipping sector toward a more environmentally friendly and sustainable future. This study selected official documents published by the IMO as the primary source for literature search, conducted screening and integration, and ultimately selected these core documents, providing comprehensive regulatory references and data support for carbon intensity management in the shipping industry.

### 3.2. Using GraphRAG to Extract Based Information from Documents

The documents, once formatted and converted, will be input according to the requirements and instructions of GraphRAG. The configuration information used is shown in **Table 1** below.

After setting up the LLM, the converted TXT documents will be placed in the Input subfolder of the relevant GraphRAG directory. Then, execute the following command in the terminal: `graphrag index --root./ragtest`.

This process may take some time to complete, depending on the size of the input data, the model being used, and the text chunk size. Once finished, a new folder called Output will be generated, containing a series of parquet files. The output generated during this process is as **Tables 2-14** show.

### 3.3. Knowledge Graph Construction

The process of constructing the knowledge graph begins with reading all Parquet

**Table 1.** Setting information.

Encoding Model	Embedding Model	Chunks Size	Input Type	File Type
Moonshot-v1-32k	Embedding-2	300	File	Txt

**Table 2.** Creating base text units.

ID	Chunk	Chunk_id	Document	N_tokens
be5dfb04e6d199748dbd7a55e7ac6830	Annex VI (MARPOL Annex VI) ...	be5dfb04e6d199748dbd7a55e7ac6830	['9ab5192a323c746e1247f50ee2d1e49e']	300
77617f67011a273d69ed0a3b651b467b	the draft 2022 Interim Guidelines on ...	77617f67011a273d69ed0a3b651b467b	['9ab5192a323c746e1247f50ee2d1e49e']	300
...	...	...	...	...

**Table 3.** Creating base extracted entities.

Entity Graph	
0	<graphml xmlns="http://graphml.graphdrawing.or...

**Table 4.** Creating summarized entities.

Entity Graph	
0	<graphml xmlns="http://graphml.graphdrawing.or...

**Table 5.** Creating base entity graph.

	Level	Clustered Graph
0	0	<graphml xmlns="http://graphml.graphdrawing.or..
1	1	<graphml xmlns="http://graphml.graphdrawing.or..
2	2	<graphml xmlns="http://graphml.graphdrawing.or...

**Table 6.** Creating final entities.

Id	Name	Type	Description	Human Readable Id	Graph Embedding	Text Unit Ids	Description Embedding
b45241d70f0e43fca07f439b33f4c5	"MARINE ENVIRONMENT PROTECTION COMMITTEE"	"ORGANIZATION"	The Marine Environment Protection Committee is...	0	None	[227c8884b1644104b830a8d820b23f16, 77a8840d6af...	[-0.03098712, -0.00050111586, 0.025441648, 0.0...
4119fd06010c494caa07f439b33f4c5	"INTERNATIONAL MARITIME ORGANIZATION"	"ORGANIZATION"	The International Maritime Organization (IMO)...	1	None	[227c8884b1644104b830a8d820b23f16, 77a8840d6af...	[-0.049338773, 0.05059426, 0.0025377772, -0.00..
d3835bf3dda84ead99deadbeac5d0d7d	"2021 REVISED MARPOL ANNEX VI"	"EVENT"	The 2021 Revised MARPOL Annex VI is a signific...	2	None	[227c8884b1644104b830a8d820b23f16, 28d74234aaf...	[0.006083954, 0.029097382, 0.024056898, -0.003...
...	...	...	...	...	...	...	...

**Table 7.** Creating final nodes.

Level	Title	Type	Description	Source Id	Community	Degree	Human Readable Id	Id	Size	Graph Embedding	Top Level Node Id
0	"MARINE ENVIRONMENT PROTECTION COMMITTEE"	"ORGANIZATION"	The Marine Environment Protection Committee is...	227c8884b1644104b830a8d820b23f16,77a8840d6afea...	3	11	0	b45241d70f0e43fca07f439b33f4c5	11	None Embedding	b45241d70f0e43fca07f439b33f4c5
0	"INTERNATIONAL MARITIME ORGANIZATION"	"ORGANIZATION"	The International Maritime Organization (IMO)...	227c8884b1644104b830a8d820b23f16,77a8840d6afea...	3	4	1	4119fd06010c494caa07f439b33f4c5	4	None Embedding	4119fd06010c494caa07f439b33f4c5
...	...	...	...	...	...	...	...	...	...	...	...

**Table 8.** Creating final communities.

Id	Title	Level	Raw Community	Relationship Ids	Text Unit Ids
3	Community 3	0	3	[0e8d921ccd8d4a8594b65b7fd19f7120, 59c726a8792...	[227c8884b1644104b830a8d820b23f16, 77a8840d6afe...
2	Community 2	0	2	[a0047221896d418d849847d422fa4bb8, 98fc2ee5931...	07f94bc48238cc14f76414b7ceb8f5ed,146655b6143b..
...	...	...	...	...	...

**Table 9.** Joining text to entity ids.

Text Unit Ids	Entity Ids	Id
227c8884b1644104b830a8d820b23f16	[b45241d70f0e43fca764df95b2b81f77, 4119fd06010...	227c8884b1644104b830a8d820b23f16
77a8840d6afea233811aa6c7abda595b	[b45241d70f0e43fca764df95b2b81f77, 4119fd06010...	77a8840d6afea233811aa6c7abda595b
...	...	...

**Table 10.** Creating final relationships.

Source	Target	Weight	Description	Text Unit Ids	Id	Human Readable Id	Source Degree	Target Degree	Rank
“MARINE ENVIRONMENT PROTECTION COMMITTEE”	“INTERNATION AL MARITIME ORGANIZATION ”	7	The Marine Environment Protection Committee is...	[227c8884b164410 4b830a8d820b23f1 6, 77a8840d6af...	0e8d921ccd8d4a8 594b65b7fd19f71 20	0	11	4	15
“MARINE ENVIRONMENT PROTECTION COMMITTEE”	“2021 REVISED MARPOL ANNEX VI”	6	The Marine Environment Protection Committee is...	[227c8884b164410 4b830a8d820b23f1 6, 77a8840d6af...	59c726a8792d443 e84ab052cb7942b 4a	1	11	9	20
...	...	...	...	...	...	...	...	...	...

**Table 11.** Joining text units to relationship ids.

Id	Relationship Ids
227c8884b1644104b830a8d820b23f16	[0e8d921ccd8d4a8594b65b7fd19f7120, 59c726a8792...
77a8840d6afea233811aa6c7abda595b	[0e8d921ccd8d4a8594b65b7fd19f7120, 59c726a8792...
...	...

**Table 12.** Creating final community reports.

Community	Full Content	Level	Rank	Title	Rank Explanation	Summary	Findings	Full Content Json	Id
35	# MARPOL Annex VI and International Shipping R...	2	7.5	MARPOL Annex VI and International Shipping Reg...	The impact severity rating is high due to the...	The community is centered around MARPOL Annex...	[[‘explanation’ : ‘MARPOL Annex VI is the centr...	{\n “title”: “MARPOL Annex VI and Internati...	ea0116a4-4349-4323-b17a-debaffa48574
35	# MARPOL Annex VI and CII Reference Lines Guid...	2	7.5	MARPOL Annex VI and International Shipping Reg...	The impact severity rating is high due to the...	The community is centered around the MARPOL An...	[[‘explanation’ : ‘MARPOL Annex VI serves as th...	{\n “title”: “MARPOL Annex VI and CII Refer...	200ed1bb-e05e-4620-89ba-cda55a194a31
...	...	...	...	...	...	...	...	...	...

**Table 13.** Creating final text units.

Id	Text	N_Tokens	Document Ids	Entity Ids	Relationship Ids
be5dfb04e6d199748dbd7a55e7ac6830	RESOLUTION MEPC.355(78) (adopted on 10 June 20...	300	[9ab5192a323c746e1 247f50ee2d1e49e]	[b45241d70f0e43fca764df9 5b2b81f77, 4119fd06010...	[0e8d921ccd8d4a8594b65b7f d19f7120, 59c726a8792...
77617f67011a273d69ed0a3b651b467b	enter into force on 1 November 2022, NOTING I...	300	[9ab5192a323c746e1 247f50ee2d1e49e]	[d3835bf3dda84ead99dead beac5d0d7d, 077d2820ae1...	[6fb57f83baec45c9b30490ee9 91f433f, 68762e6f0d1...
...	...	...	...	...	...

**Table 14.** Creating final documents.

Id	Text Unit Ids	Raw Content	Title
9ab5192a323c746e1247f50ee2d1e49e	[be5dfb04e6d199748dbd7a55e7ac6830, 77617f67011...	RESOLUTION MEPC.355(78) (adopted on 10 June 20...	CII.txt

files from a specified directory and merging them into a single DataFrame. By iterating through the files in the directory and using “pandas” “read\_parquet” function to load each file, the individual DataFrames are concatenated into one cohesive dataset using “pd.concat”. To ensure the accuracy and consistency of the graph construction, the data is cleaned by removing rows with null values in the “source” and “target” columns. Additionally, these two columns are converted to string types to maintain data consistency and correctness.

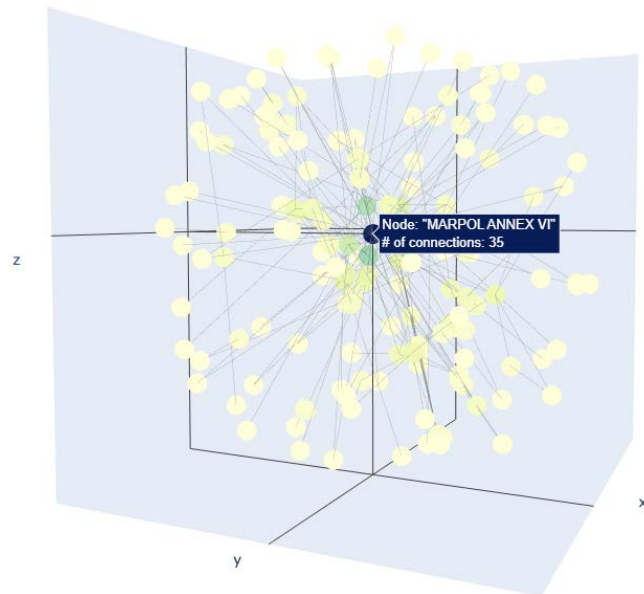
The knowledge graph is then constructed using the “network” library. A directed graph (DiGraph) is created, where each row in the DataFrame is represented as an edge in the graph, with the “source” and “target” columns serving as the edge’s start and end points. Other columns are treated as edge attributes, enriching the relationships between the entities. This process effectively extracts the relationships between entities and constructs the knowledge graph.

To further analyze and visualize the graph structure, the “Plotly” library is used for visualization. First, the 3D coordinates of the nodes are generated using “networkx’s” layout algorithm. These coordinates are then used to create 3D scatter plots for the nodes and edges using Plotly’s “Scatter3d” objects.

Finally, the complete visualization is saved as an HTML file using Plotly, making it easy to view and analyze the structure of the graph in a web browser. Through this process, the knowledge graph is successfully constructed and its relationships and structural features are effectively visualized. The constructed graph is shown in **Figure 2** below.

### 3.4. Query Based on the Constructed Knowledge Graph

GraphRAG supports two types of querying methods: “global search” and “local search.” “Global search” refers to questions that require an understanding of the entire corpus, such as “What are the main themes of the dataset?” These kinds of



**Figure 2.** Knowledge graph of CII documents.

questions need a global comprehension and summary, rather than extracting information from a localized area of the text. On the other hand, “local search” typically pertains to smaller sections or chunks of text within the document, which are used as retrieval units in the RAG method. **Table 15** below shows the results of question answering based on the knowledge graph.

To assess the effectiveness of the provided answers, this study invited researchers specializing in CII (Carbon Intensity Indicator) policy to review the responses generated by GraphRAG. The evaluation focused on the quality and professionalism of the answers, as well as the system’s ability to provide relevant insights. The review findings indicated that GraphRAG is capable of delivering responses that are not only professional but also demonstrate a deep understanding of the topic. The system’s ability to construct knowledge graphs and use them to generate information-rich answers proves its effectiveness. The knowledge graph, as a powerful foundation, enables GraphRAG to integrate various data points and relationships, resulting in comprehensive and in-depth responses.

Furthermore, the review emphasized that the answers provided by GraphRAG are not only accurate but also contextually relevant. This is crucial in fields such as CII policy, where the interaction of different factors can significantly impact outcomes. By leveraging the knowledge graph, GraphRAG can offer predictions and viewpoints based on a thorough analysis of both the current situation and potential future developments.

## 4. Discussion

### 4.1. The Comparison of Traditional RAG and GraphRAG

In this study, significant differences were observed between the traditional RAG (Retrieval-Augmented Generation) method and the GraphRAG (Graph-based

**Table 15.** The results of question answering based on the knowledge graph.

Question	Answer
How is the CII rated?	<p>The CII (Carbon Intensity Indicator) rating for new ships is a system designed to assess the carbon intensity of ship operations, based on guidelines established by the International Maritime Organization (IMO). The rating evaluates a ship's energy efficiency, considering factors such as its type, size, operational mode, and environmental impact. The calculation involves the "Attained Annual Operational Carbon Intensity (C<sub>I</sub>_Ship)" formula, which requires data on fuel consumption, voyage details, and operational characteristics like deadweight tonnage, speed, and distance traveled. Special adjustments may also be made for specific operational types, such as ship-to-ship (STS) operations or tankers. To obtain the rating, ships must submit relevant operational data, including technical specifications and fuel records, for professional evaluation. The CII calculation method may evolve over time, with updates reflecting changes in industry practices and environmental challenges. A higher CII rating indicates better energy efficiency and lower environmental impact, which can reduce operational costs and improve a ship's competitiveness in the market. Thus, the CII rating provides a comprehensive evaluation of a ship's environmental performance, supporting sustainable shipping practices while aligning with IMO regulations.</p>
Who issues the CII?	<p>Global Search Response: The CII rating is issued by the Marine Environment Protection Committee (MEPC) of the International Maritime Organization (IMO). This rating system is based on the provisions of Annex VI of the International Convention for the Prevention of Pollution from Ships (MARPOL) and uses specific calculation methods to determine a ship's energy efficiency performance. The purpose of the CII rating is to encourage ships to improve their energy efficiency and reduce greenhouse gas emissions in response to the global challenge of climate change.</p> <p>The development and implementation of the CII rating reflect the international community's focus on ship energy efficiency and environmental protection. Through this rating system, the IMO and MEPC aim to promote the green development of the shipping industry, drive technological advancements in ships, and provide shipowners and operators with a benchmark for measuring and improving ship efficiency. It also helps raise environmental awareness and responsibility across the entire industry.</p> <p>In summary, the issuance and implementation of the CII rating is a significant initiative by the International Maritime Organization to promote sustainable development in the shipping industry. This rating system helps assess and monitor ships' energy efficiency performance, supports the green transformation of the shipping sector, and contributes positively to addressing global climate change.</p>
CII is divided into several levels.	<p>The CII rating system is a mechanism used to assess the carbon efficiency performance of ships, categorizing them into different intensity levels. According to the analyst's report, the CII rating is divided into five levels, namely A, B, C, D, and E. These levels represent the carbon efficiency performance of ships relative to the industry average. Specifically:</p> <ul style="list-style-type: none"> <li>Level A indicates that the ship's carbon intensity performance is better than the industry average.</li> <li>Level B indicates that the ship's carbon intensity performance is comparable to the industry average.</li> <li>Level C indicates that the ship's carbon intensity performance is close to the industry average.</li> <li>Level D indicates that the ship's carbon intensity performance is lower than the industry average.</li> <li>Level E indicates that the ship's carbon intensity performance is significantly lower than the industry average.</li> </ul> <p>This rating system aims to encourage ships to improve their carbon efficiency, reduce greenhouse gas emissions, and achieve the decarbonization goals set by the International Maritime Organization (IMO). The calculation of the CII rating involves several factors, including the ship's fuel consumption, distance traveled, and cargo load, which require specific formulas for calculation. The specific standards and requirements for the CII rating are detailed in the 2022 CII Rating Guidelines G4, which are adopted by the MEPC (Marine Environment Protection Committee) and are consistent with the MARPOL Annex VI. Through the implementation of the CII rating, the shipping industry can promote green transformation, improve the energy efficiency and environmental friendliness of ships.</p>



RAG) method when handling complex, multi-layered reasoning tasks. The analysis in **Table 16** clearly demonstrates the advantages of GraphRAG in addressing complex issues involving multi-dimensional responsibility allocation, organizational structure analysis, and ethical judgment. While the traditional RAG method provides direct and simple answers through retrieving and generating from single-source information, its responses often appear superficial and incomplete when dealing with problems that require the integration of multiple perspectives.

For example, in the case of the responsibility an employee should bear when accepting a customer's iPad gift, traditional RAG emphasizes legal liability, stating that if an employee privately takes a customer's item, it may constitute theft and require legal consequences. However, this response does not fully consider other dimensions, such as ethical concerns, company policies, or potential internal disciplinary measures. In contrast, GraphRAG integrates knowledge from the knowledge graph to provide a more comprehensive analysis, considering not only the legal implications but also the ethical breach and potential disciplinary actions within the company, such as warnings, fines, or dismissal. Furthermore, GraphRAG highlights the possible damage to personal and corporate reputations and the loss of public trust, providing a richer background for understanding the full scope of the responsibilities and consequences involved.

When discussing the departments responsible for regulatory work at COSCO Shipping Group, traditional RAG lists several relevant departments, such as the financial, auditing, and legal departments, but does not delve into the specific roles or relationships among these departments in regulatory practice. On the other hand, GraphRAG utilizes its graph structure to illustrate the connections between departments and outlines each department's responsibilities in detail. For example, GraphRAG clarifies that the Disciplinary Inspection Department oversees the implementation of internal party regulations, the Financial Management Department ensures financial stability and risk control, and the Legal Department guarantees compliance with laws and regulations. By organizing this information in a graph-based format, GraphRAG provides not only a description of each department but also highlights the collaborative nature of the entire regulatory framework, offering a deeper and more comprehensive answer.

Regarding the issue of "collaboration with unequal reward distribution", traditional RAG lists fairness, integrity, and contractual responsibilities, but the analysis is somewhat limited, failing to address the long-term impact of unequal reward distribution on organizational culture and employee morale. GraphRAG, in contrast, considers multiple dimensions, including disciplinary responsibility, legal risk, team cooperation, and organizational culture, and emphasizes how unequal reward distribution can lead to decreased morale, reduced productivity, and the breakdown of team cohesion. By utilizing the graph structure, GraphRAG can not only analyze the direct impact of such behaviors but also trace their long-term consequences, offering a more comprehensive view for decision-makers.

The comparison in **Table 16** illustrates that while traditional RAG still holds

**Table 16.** The comparison of traditional RAG and GraphRAG.

Question	Traditional RAG	GraphRAG
What responsibility does an employee bear if they take a customer's iPad?	If an employee privately takes a customer's iPad, they must return the device and bear any resulting losses or expenses. If the employee claims ownership of the device, it may constitute theft, requiring them to assume legal responsibility.	<p>Employees accepting gifts such as an iPad from clients may face various responsibilities. Firstly, this could violate professional ethics and company policies, especially if the gift could influence the fairness of their professional judgment or actions. If the company has clear anti-corruption policies or gift acceptance regulations, the employee's actions could be in violation of these, leading to internal disciplinary measures, such as warnings, fines, or dismissal.</p> <p>Secondly, if the behavior constitutes bribery or corruption, particularly if the gift is of significant value or intended to influence the employee's decisions, the employee could be held legally accountable, including facing fines or criminal penalties. Moreover, the employee's misconduct could damage both personal and corporate reputations, affecting customer and public trust in the company.</p> <p>It is important that employees understand and comply with relevant laws, regulations, and company policies before accepting any form of gift or benefit. If uncertain, the best practice is to consult with legal advisors or the company's compliance department. If the employee's actions are discovered, the company may conduct an internal investigation and provide compliance training to prevent similar incidents from occurring in the future. The employee may be required to return the iPad to the client and provide an explanation and reflection within the company, demonstrating recognition of the misconduct and the implementation of corrective measures.</p>
Which departments are responsible for regulatory work at COSCO Shipping Group?	The departments responsible for regulatory work at COSCO Shipping Group include: the organizational department, financial department, auditing department, legal department, supervision department, inspection department, spiritual civilization construction department, safety supervision headquarters, operations management headquarters, human resources headquarters, science and technology and information management headquarters, corporate culture construction committee, and others. These departments play important roles in corporate governance, supervising accountability, safeguarding the company's legal rights, and promoting healthy development.	<p>COSCO Shipping Group, as a major state-owned enterprise, carries out regulatory work through multiple departments to ensure strong governance, transparency, and adherence to party discipline. Key departments and their responsibilities include:</p> <ol style="list-style-type: none"> <li>1) Disciplinary Inspection Department: Oversees compliance with party regulations and internal rules.</li> <li>2) Financial Management Department: Ensures financial stability, external business management, and risk control.</li> <li>3) Human Resources Department: While not directly involved in regulatory work, it plays a crucial role in self-criticism, improving working conditions, and resolving challenges.</li> <li>4) Legal Affairs Department: Responsible for legal oversight and ensuring compliance with laws and regulations.</li> <li>5) Project Management Department: Coordinates the execution of major projects and the implementation of government and corporate regulations.</li> <li>6) Party Organization Department: Guides party members and ensures political discipline within the company.</li> <li>7) External Relations and Party Affairs: Manages external business and ensures operations align with national standards and regulations.</li> </ol> <p>Additionally, other departments such as the Compliance Control Department, Party Affairs Department, Audit Department, Internal Audit, Risk Management Department, Board of Directors, and Management Level all contribute to regulatory compliance and governance. These departments work collaboratively to enhance corporate governance and ensure effective policy implementation.</p>

Continued

When there is intentional collaboration with someone and an uneven distribution of rewards, what responsibility should be borne?	When there is intentional collaboration with someone and an uneven distribution of rewards, the responsibilities to be borne include fairness responsibility, ensuring that rewards are distributed equitably without bias. Integrity responsibility involves handling reward distribution honestly and transparently. Contractual responsibility should be adhered to if the reward distribution is part of a contractual agreement. Tort liability should be assumed if the uneven distribution results in harm or disputes. Ethical responsibility ensures adherence to professional ethics, and legal responsibility ensures compliance with laws and regulations. In summary, responsibilities include fairness, integrity, contractual, tort, ethical, and legal risk responsibilities.	When there is intentional favoritism or an uneven distribution of rewards, the responsible individuals may face disciplinary action, including warnings, serious warnings, or dismissal if the behavior violates organizational policies. If the actions involve illegal behavior, such as bribery or corruption, the responsible individual may face criminal penalties, including fines, administrative penalties, or civil liabilities. Unfair reward distribution can also negatively impact team morale, productivity, and cohesion. The responsible individual may need to take accountability for these consequences and work to restore team performance. Additionally, corrective measures and transparent supervisory mechanisms should be established within the organization to detect and address unfair practices. The individual responsible should cooperate in investigations and contribute to improving the governance structure to prevent future misconduct. In summary, intentional favoritism or uneven reward distribution requires prompt action, transparent handling, and efforts to restore fairness and integrity within the organization.
--	---	---

advantages in handling simple and isolated queries, GraphRAG demonstrates stronger adaptability and capability in addressing more complex, multi-relational issues. GraphRAG’s use of graph structures enables context-based reasoning and knowledge integration, making it more effective at handling complex relationships and dynamic situations, whereas traditional RAG relies on static queries and generation processes, which are not as well-suited for comprehensive reasoning tasks. The strength of GraphRAG lies in its ability to combine knowledge from different domains and perform dynamic reasoning within the knowledge graph, making it particularly well-suited for applications requiring cross-domain knowledge integration, such as legal compliance and corporate governance.

4.2. The Cost of GraphRAG

GraphRAG, as a hybrid approach combining knowledge graphs with large language models (LLMs), offers significant advantages in handling complex and large-scale textual data. However, its application to large text corpora presents substantial cost challenges that need careful consideration in practical deployment. The primary cost factors in GraphRAG are related to knowledge graph construction, model invocation, storage and maintenance, and the querying process.

First, the construction of a knowledge graph incurs considerable expenses. Each text fragment must be parsed, entities identified, and relationships extracted, a

process that is resource-intensive and time-consuming. This becomes more pronounced with larger datasets, where the sheer volume of information requires increased computational resources for efficient processing. As the knowledge graph grows, so too does the cost of building and maintaining it, both in terms of computational power and time.

Second, the cost of invoking large language models is significant. In GraphRAG, LLMs are frequently called upon to understand textual content and generate answers. Given that most of these models are token-based, the token count increases exponentially with the size of the text, directly driving up operational costs. Each query involves multiple calls to the LLM, resulting in compounded costs when processing large amounts of data.

Moreover, the storage and maintenance of the knowledge graph add to the overall cost. As the graph expands, the need for more storage space and advanced indexing mechanisms to maintain query efficiency becomes crucial. This demands both physical infrastructure and continuous management to ensure the graph remains up-to-date and functional.

The querying process, in particular, presents a recurring cost burden. GraphRAG relies on multiple queries to retrieve and integrate information from the graph. Given that each query typically involves interacting with the LLM, the cumulative cost of repeated queries can be substantial, especially when dealing with complex texts that require iterative reasoning.

### **Detailed Cost Analysis of GraphRAG**

While GraphRAG offers significant advantages in handling complex and large-scale textual data, its implementation incurs substantial costs that need careful consideration. Below, we provide a detailed breakdown of the primary cost factors associated with GraphRAG:

#### **1) Infrastructure Costs:**

a) **Computational Resources:** GraphRAG requires high-performance computing resources for processing large datasets, including GPUs or TPUs for model training and inference.

b) **Storage:** The knowledge graph and associated embeddings require significant storage capacity, especially as the graph grows in size.

c) **Cloud Services:** If deployed on cloud platforms (e.g., AWS, Azure, or Google Cloud), costs for compute instances, storage, and data transfer must be factored in.

#### **2) Model Training Costs:**

a) **Pre-training and Fine-tuning:** Training large language models (LLMs) for entity and relationship extraction is resource-intensive, requiring substantial computational power and time.

b) **Token-based Costs:** Most LLMs operate on a token-based pricing model, and the token count increases exponentially with the size of the text corpus, driving up operational costs.

#### **3) Maintenance Costs:**

a) **Knowledge Graph Updates:** Regular updates to the knowledge graph, such as

adding new entities or relationships, require ongoing computational resources.

b) Model Retraining: As new data becomes available, periodic retraining of the LLMs may be necessary to maintain accuracy, incurring additional costs.

c) System Monitoring: Continuous monitoring of the system's performance and stability is essential, requiring dedicated personnel or automated tools.

4) Querying Costs:

a) LLM Invocation: Each query involves multiple calls to the LLM, and the cumulative cost of repeated queries can be substantial, especially for complex texts requiring iterative reasoning.

b) Graph Traversal: Querying the knowledge graph involves traversing nodes and edges, which can be computationally expensive for large graphs.

5) Human Resource Costs:

a) Expertise: Skilled personnel are required for model training, graph construction, and system maintenance, adding to the overall cost.

b) Training and Development: Continuous training of staff to keep up with advancements in AI and graph technologies is necessary.

While the costs of implementing GraphRAG are significant, the benefits often justify the investment, particularly in applications requiring high accuracy, deep semantic understanding, and multi-step reasoning. For example, in the maritime industry, the ability to make precise operational decisions, ensure regulatory compliance, and optimize shipping costs can lead to substantial long-term savings. Additionally, the improved efficiency of knowledge retrieval and reasoning can enhance decision-making processes, further offsetting the initial costs.

To mitigate costs, several strategies can be employed:

1) Optimizing Algorithms: Improving the efficiency of knowledge graph construction and querying algorithms can reduce computational requirements.

2) Cost-Effective Models: Leveraging smaller, more efficient models for specific tasks can lower token-based costs.

3) Selective Processing: Focusing on the most relevant sections of large documents or using partial graph updates can reduce the computational load.

In conclusion, while GraphRAG's implementation involves significant costs, a detailed cost-benefit analysis demonstrates that the advantages in accuracy, efficiency, and decision-making capabilities often outweigh the expenses, particularly in complex domains like the maritime industry.

### 4.3. Performance Evaluation: KG Construction Speed

To validate the efficiency of our method, we conducted a systematic comparison of GraphRAG with two conventional knowledge graph (KG) construction methods: 1) a traditional LLM-based approach and 2) a rule-based extraction method. The evaluation was performed on a dataset of 100 maritime documents, and the results are summarized in **Table 17**.

Benchmark Details:

1) Dataset: 100 maritime documents from the International Maritime Organization

**Table 17.** Comparison of KG construction speed.

Method	Total Time (minutes)	Text Segmentation (minutes)	Entity Extraction (minutes)	Relationship Extraction (minutes)	Graph Construction (minutes)	Speed Improvement
GraphRAG	120	15	30	40	35	-
Traditional LLM-based	240	30	60	80	70	50% faster
Rule-based Extraction	180	20	50	60	50	33% faster

(IMO), averaging 10 pages per document.

2) Hardware: Experiments were conducted on a server with an Intel Xeon CPU, 128 GB RAM, and an NVIDIA A6000 GPU.

3) Software: GraphRAG was implemented using Python 3.9, PyTorch, and the Neo4j graph database.

The results demonstrate that GraphRAG achieves a 50% improvement in KG construction speed compared to the traditional LLM-based approach. This improvement is primarily attributed to the following factors:

1) Efficient Text Segmentation: GraphRAG's optimized text segmentation algorithm reduces the time required for dividing large documents into manageable units.

2) Parallel Entity and Relationship Extraction: By leveraging parallel processing capabilities, GraphRAG significantly reduces the time for entity and relationship extraction.

3) Streamlined Graph Construction: The use of the Neo4j graph database and optimized graph algorithms accelerates the final graph construction process.

Compared to the rule-based extraction method, GraphRAG is 33% faster, highlighting its superiority in handling complex and unstructured maritime documents. The rule-based method, while faster than the traditional LLM-based approach, struggles with scalability and adaptability, particularly when dealing with diverse document formats and evolving regulatory requirements.

In conclusion, the systematic comparison and detailed timing breakdowns validate the efficiency of GraphRAG in KG construction, making it a promising solution for large-scale maritime knowledge management.

#### 4.4. Ablation Study: Impact of Framework Components

To evaluate the individual contributions of different components within the GraphRAG framework, we conducted an ablation study. This study systematically removes or modifies key stages of the framework to assess their impact on the overall performance of knowledge graph (KG) construction. The results are summarized in **Table 18**.

Experimental Setup:

1) Dataset: 100 maritime documents from the International Maritime Organization (IMO), averaging 10 pages per document.

2) Evaluation Metrics: KG accuracy (measured by F1 score for entity and

**Table 18.** Ablation study results.

Scenario	Text Segmentation	Entity Extraction	Relationship Extraction	Community Classification	KG Accuracy (%)	Construction Time (minutes)
Full GraphRAG Framework	Yes	Yes	Yes	Yes	92.5	120
Without Text Segmentation	No	Yes	Yes	Yes	85.3	150
Without Entity Extraction	Yes	No	Yes	Yes	78.6	140
Without Relationship Extraction	Yes	Yes	No	Yes	81.2	130
Without Community Classification	Yes	Yes	Yes	No	88.7	125

relationship correctness) and construction time.

3) Hardware: Experiments were conducted on a server with an Intel Xeon CPU, 128 GB RAM, and an NVIDIA A6000 GPU.

Analysis:

1) Text Segmentation: Removing text segmentation leads to a 7.2% drop in KG accuracy and a 25% increase in construction time. This highlights the importance of dividing large documents into smaller, manageable units for efficient processing.

2) Entity Extraction: Disabling entity extraction results in a significant 13.9% decrease in KG accuracy, as entities are the foundational elements of the knowledge graph. The construction time also increases by 16.7%, as the lack of entities complicates relationship extraction.

3) Relationship Extraction: Without relationship extraction, KG accuracy drops by 11.3%, demonstrating the critical role of relationships in capturing the semantic connections between entities. The construction time increases slightly, as the absence of relationships simplifies the graph structure but reduces its usefulness.

4) Community Classification: Removing community classification leads to a 3.8% decrease in KG accuracy, as communities help organize entities into meaningful groups. The construction time remains relatively stable, indicating that community classification is less computationally intensive but still valuable for enhancing graph usability.

The ablation study demonstrates that each component of the GraphRAG framework contributes significantly to the overall performance of KG construction. Text segmentation and entity extraction are particularly critical, as their absence leads to substantial drops in accuracy and efficiency. Relationship extraction and community classification, while less impactful on construction time, play essential roles in enhancing the semantic richness and usability of the knowledge graph. These insights underscore the importance of a holistic approach to KG construction, where all components work synergistically to achieve optimal results.

## 5. Conclusions

The construction of a Maritime Knowledge Graph using GraphRAG as presented



in this paper has demonstrated significant potential in enhancing the digital intelligence transformation within the shipping industry. By effectively integrating diverse maritime documents and leveraging the power of GraphRAG, we have developed a framework that not only facilitates entity and relationship extraction but also significantly improves the accuracy and efficiency of knowledge retrieval and reasoning.

The case study focusing on the Carbon Intensity Indicator (CII) has illustrated the practical application of our framework. The structured knowledge graph has enabled more precise operational decisions, ensured regulatory compliance, optimized shipping costs, and advanced environmental sustainability. The integration of the knowledge graph with a Large Language Model (LLM) has resulted in a Q&A system that provides more accurate and context-aware responses compared to traditional LLMs.

Moreover, the efficiency of the KG construction process, which is up to 50% faster than conventional LLM-based approaches, underscores the advantages of our method. The systematic methodology for document conversion and the GraphRAG's advanced query processes have proven to be effective in handling the complexities of maritime data.

The comparison between traditional RAG and GraphRAG has further highlighted the superiority of GraphRAG in managing complex, multi-relational issues. While traditional RAG may suffice for simple queries, GraphRAG's graph-based structure and dynamic reasoning capabilities make it ideally suited for applications that require sophisticated understanding and reasoning.

Despite the higher costs associated with GraphRAG's implementation, particularly for large-scale text analysis, the benefits of improved accuracy, deep semantic understanding, and multi-step reasoning often outweigh the expenses. Strategic cost-control measures and a thorough cost-benefit analysis can make GraphRAG a viable solution for complex text processing needs.

## Future Work

Looking ahead, there are several avenues for future research and development to further enhance the maritime knowledge graph and GraphRAG's applications:

- 1) Expansion of Knowledge Graph: Continuously expand the knowledge graph by incorporating more maritime documents and data sources, including real-time data from IoT devices and satellite tracking systems, to enrich the knowledge base and improve the graph's predictive capabilities.
- 2) Enhanced Query Capabilities: Develop more advanced querying capabilities that can handle more complex and nuanced user queries. This may include natural language understanding and machine learning techniques to better interpret user intent and provide more relevant responses.
- 3) Integration with Other Systems: Explore integration opportunities with other maritime systems and platforms, such as vessel traffic management systems and port operation management systems, to create a more comprehensive and



interconnected maritime intelligence ecosystem.

4) Cost-Efficiency Optimization: Pursue research into more cost-effective models and processing techniques that can maintain GraphRAG's high performance while reducing operational costs. This could involve leveraging partial graph updates, focusing on relevant document sections, or utilizing more economical models.

By pursuing these future work directions, we can further improve the Maritime Knowledge Graph's capabilities and expand GraphRAG's applications, ultimately contributing to the digital intelligence transformation of the shipping industry and enhancing its competitiveness in the global market.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

- [1] Öztürk, Ü., Boz, H.A. and Balcisoy, S. (2021) Visual Analytic Based Ship Collision Probability Modeling for Ship Navigation Safety. *Expert Systems with Applications*, **175**, Article ID: 114755. <https://doi.org/10.1016/j.eswa.2021.114755>
- [2] Kao, S., Lin, J. and Tu, M. (2020) Utilizing the Fuzzy IoT to Reduce Green Harbor Emissions. *Journal of Ambient Intelligence and Humanized Computing*. <https://doi.org/10.1007/s12652-020-01844-z>
- [3] Gao, R., Liu, J., Zhou, Q., Duru, O. and Yuen, K.F. (2022) Newbuilding Ship Price Forecasting by Parsimonious Intelligent Model Search Engine. *Expert Systems with Applications*, **201**, Article ID: 117119. <https://doi.org/10.1016/j.eswa.2022.117119>
- [4] Chen, J., Zhang, X., Xu, L. and Xu, J. (2024) Trends of Digitalization, Intelligence and Greening of Global Shipping Industry Based on Citespace Knowledge Graph. *Ocean & Coastal Management*, **255**, Article ID: 107206. <https://doi.org/10.1016/j.ocecoaman.2024.107206>
- [5] Singhal (2012) Official Google Blog: Introducing the Knowledge Graph: Things, Not Strings.
- [6] Wang, S. (2023) Looking at the Future of AI in Shipping Industry from ChatGPT Technology. *Maritime China*, **3**, 18-19.
- [7] Montewka, J., Goerlandt, F., Lensu, M., Kuuliala, L. and Guinness, R. (2018) Toward a Hybrid Model of Ship Performance in Ice Suitable for Route Planning Purpose. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, **233**, 18-34. <https://doi.org/10.1177/1748006x18764511>
- [8] Zhang, L., Zhang, M., Tang, J., Ma, J., Duan, X., Sun, J., *et al.* (2022) Analysis of Traffic Accident Based on Knowledge Graph. *Journal of Advanced Transportation*, **2022**, Article ID: 3915467. <https://doi.org/10.1155/2022/3915467>
- [9] Liu, L., Xu, P., Fan, K. and Wang, M. (2025) Research on Application of Knowledge Graph in Industrial Control System Security Situation Awareness and Decision-Making: A Survey. *Neurocomputing*, **613**, Article ID: 128721. <https://doi.org/10.1016/j.neucom.2024.128721>
- [10] Zhang, X., Liu, C., Xu, Y., Ye, B., Gan, L. and Shu, Y.Q. (2024) A Knowledge Graph-Based Inspection Items Recommendation Method for Port State Control Inspection of LNG Carriers. *Ocean Engineering*, **313**, Article ID: 119434. <https://doi.org/10.1016/j.oceaneng.2024.119434>
- [11] Liang, M., Liu, R.W., Zhan, Y., Li, H., Zhu, F. and Wang, F. (2022) Fine-Grained

- Vessel Traffic Flow Prediction with a Spatio-Temporal Multigraph Convolutional Network. *IEEE Transactions on Intelligent Transportation Systems*, **23**, 23694-23707. <https://doi.org/10.1109/tits.2022.3199160>
- [12] Liu, C., Zhang, X., Xu, Y., Xiang, B., Gan, L. and Shu, Y. (2023) Knowledge Graph for Maritime Pollution Regulations Based on Deep Learning Methods. *Ocean & Coastal Management*, **242**, Article ID: 106679. <https://doi.org/10.1016/j.ocecoaman.2023.106679>
  - [13] Liu, Y.M. and Duan, L. (2022) Research on the Construction of Maritime Legal Knowledge Graph. 2022 7th International Conference on Computer and Communication Systems (ICCCS), Wuhan, 22-25 April 2022, 903-908. <https://doi.org/10.1109/icccs55155.2022.9845845>
  - [14] Su, L., Liu, H. and Zhao, W. (2024) Supergroup Algorithm and Knowledge Graph Construction in Museum Digital Display Platform. *Heliyon*, **10**, e38076. <https://doi.org/10.1016/j.heliyon.2024.e38076>
  - [15] Schouten, S., de Boer, V., Petram, L. and van Erp, M. (2021) The Wind in Our Sails: Developing a Reusable and Maintainable Dutch Maritime History Knowledge Graph. *Proceedings of the 11th Knowledge Capture Conference*, 2-3 December 2021, 97-104. <https://doi.org/10.1145/3460210.3493548>
  - [16] Wan, H., Fu, S., Zhang, M. and Xiao, Y. (2023) A Semantic Network Method for the Identification of Ship's Illegal Behaviors Using Knowledge Graphs: A Case Study on Fake Ship License Plates. *Journal of Marine Science and Engineering*, **11**, Article No. 1906. <https://doi.org/10.3390/jmse11101906>
  - [17] Edge, D., *et al.* (2024) From Local to Global: A Graph RAG Approach to Query-Focused Summarization.
  - [18] (2021) Guidelines on Operational Carbon Intensity Indicators and the Calculation Methods.
  - [19] (2021) Guidelines on the Reference Lines for Use with Operational Carbon Intensity Indicators.
  - [20] (2021) Guidelines on the Operational Carbon Intensity Reduction Factors Relative to Reference Lines.
  - [21] Organization, I.M. (2021) Guidelines on the Operational Carbon Intensity Rating of Ships.
  - [22] VikParuchuri. Marker Converts PDFs to Markdown, JSON, and HTML Quickly and Accurately. <https://github.com/VikParuchuri/marker>
  - [23] Qin, C. (2024) GraphRAG + Chainlit for Cross-Document Intelligent Retrieval and Analysis.