# On Leveraging AI for Term Structure Understanding in Maritime Asset-Backed Deals

Narayanan Arvind[1], Yetirajan**
[1]Intain Technologies Pvt Ltd, Karapakkam, Chennai, India 600097

**Abstract:** In the maritime finance sector, structured deal documents play a critical role in governing capital deployment for shipbuilding, leasing, and offshore infrastructure projects. These documents—akin to Residential Mortgage-Backed Securities (RMBS) agreements—contain highly specialized term definitions, often buried deep within complex legal texts. Accurate and scalable extraction of these definitions is essential for automation, compliance, and risk evaluation in maritime asset-backed financing.

This work presents an AI-driven pipeline for robust term definition extraction from maritime deal documents, drawing parallels with RMBS processing frameworks. Our solution handles both digitally readable and scanned (non-readable) PDFs using a hybrid stack: pdfplumber for text-based documents and Google OCR with multithreaded parsing for image-based inputs. We classify 1,500-token chunks using large language models (LLMs) to identify glossary sections containing formal term definitions. These identified pages are clustered to isolate the definition block, preventing contamination from unrelated sections and ensuring full coverage.

We apply an overlapped chunking strategy (2400-token size with 800-token overlap) to ensure contextual continuity. Extracted definitions are stored efficiently using DuckDB, with retrieval latencies of 0.02s and an average accuracy of 90% over 20 domain-specific queries across two real-world deals.

The proposed framework offers a scalable foundation for semantic modeling and intelligent querying of financial instruments in the maritime domain, supporting audit, automation, and contract interpretation across complex offshore financing structures.

**Key words:** Maritime finance, Structured finance, Term definition extraction, Maritime deal automation, AI in contract analysis, Intelligent document understanding

---

[a] Corresponding author, Email: arvind.narayanan@intainft.com Tel: +91-8349232657
- PGP in AI/ML, Texas McCombs School of Business at the University of Texas at Austin and the Great Lakes Institute of Management
- Adv Gen AI at upGrad
- M.Tech. Software Engineering (WILP), Birla Institute of Technology and Science, Pilani
- Fellow AI/ML, GradValley DataScience, Coimbatore
- B.Tech. / M.Tech. Ocean Engineering and Naval Architecture, IIT Kharagpur
** The author's brand name

## 1 Introduction

The maritime sector has witnessed growing reliance on structured financing instruments such as asset-backed securities (ABS), secured by vessels, freight receivables, or charter contracts. These transactions demand rigorous documentation, typically spanning 100–400 pages, with extensive term definitions, cash-flow waterfalls, and credit enhancement provisions.

Despite their importance, current industry practice remains heavily manual: analysts must locate, interpret, and reconcile definitions across multiple deals. This slows due diligence, increases the risk of oversight, and hinders secondary market efficiency. Recent advances in natural language processing (NLP), embeddings, and generative AI provide an opportunity to address these limitations by enabling **automated term structure understanding (TSU)**.

Our contribution is twofold:

1. We design a modular architecture for preprocessing, extraction, and comparison of term definitions in ABS documentation.

2. We validate the approach on real-world datasets (including residential mortgage-backed securities, RMBS, as a proxy) and demonstrate that it generalizes to maritime ABS.

The proposed solution achieves high accuracy and low latency, supporting applications such as definition alignment, cross-deal comparison, and generative question answering.

## 2 Related Work

#### 2.1 Document Classification in Maritime and Shipping

OCR-driven classification has been explored in shipping documentation (invoices, customs declarations, bills of lading). A semi-automatic method achieved 92% accuracy using keyword vectors on “Contract-Breach” law documents. While effective for classification, these approaches lack the granularity needed for structured term extraction.

#### 2.2 Term Definition Extraction and Comparison

Recent architectural decision records (ADR) emphasize scalable extraction and precomputation strategies. One ADR proposed precomputing top-k similarity metrics across documents, achieving 96% accuracy in comparing definitions. This aligns closely with our requirement of cross-deal reconciliation.

#### 2.3 Current State of Doc-as-Data Pipelines

The *Doc-as-Data* framework integrates OCR, embeddings, and vector DBs for term extraction. Evaluations reported 97% accuracy for extraction and 95% accuracy for comparisons using DuckDB, with latencies under 0.02s. While promising, prior work focused on mortgage securities; our work extends these principles to maritime ABS.

---

## 3 Requirements Analysis

Our requirements analysis identified several key challenges:

- **Document diversity**: Deals may be in readable PDF or scanned (non-readable) formats.
- **Scale**: Each document ranges 100–400 pages; institutional portfolios may include 1,000+ deals.
- **Accuracy**: Analysts require ≥95% precision in term definition extraction and comparison.
- **Latency**: Query responses must be interactive (<1s preferred).
- **Interpretability**: Outputs must classify definitions as *same, minor variation, materially different*.
- **Virtual assistant integration**: A conversational interface should support queries such as *"Compare 'Eligible Vessel' between Deal A and Deal B"*.

Sample analysis across 34 RMBS documents highlighted glossary sections spanning up to 60 pages, multipage definitions, quoted and underlined terms, and significant variation across readable and scanned formats.

---

## 4 System Architecture

Our architecture consists of three layers: preprocessing, search, and generative reasoning.

#### 4.1 Storage and Ingestion

Raw deal PDFs are stored in Azure Blob Storage. Upload triggers invoke extraction pipelines (via Azure Functions).

#### 4.2 Preprocessing Layer

As shown in *Figure 1* (Chroma DB pipeline), the preprocessing module:

- Converts PDFs to text (PyMuPDF).
- Applies chunking (with and without overlap).
- Classifies pages as containing term definitions.
- Groups clusters into “Term-Definition blocks.”
- Embeds and stores these in Chroma DB.

#### 4.3 Search and Re-ranking Layer

Queries are matched against the vector database, top-k results are re-ranked, and the best three are selected (*Figure 1, Step 2*).

#### 4.4 Generative Search Layer

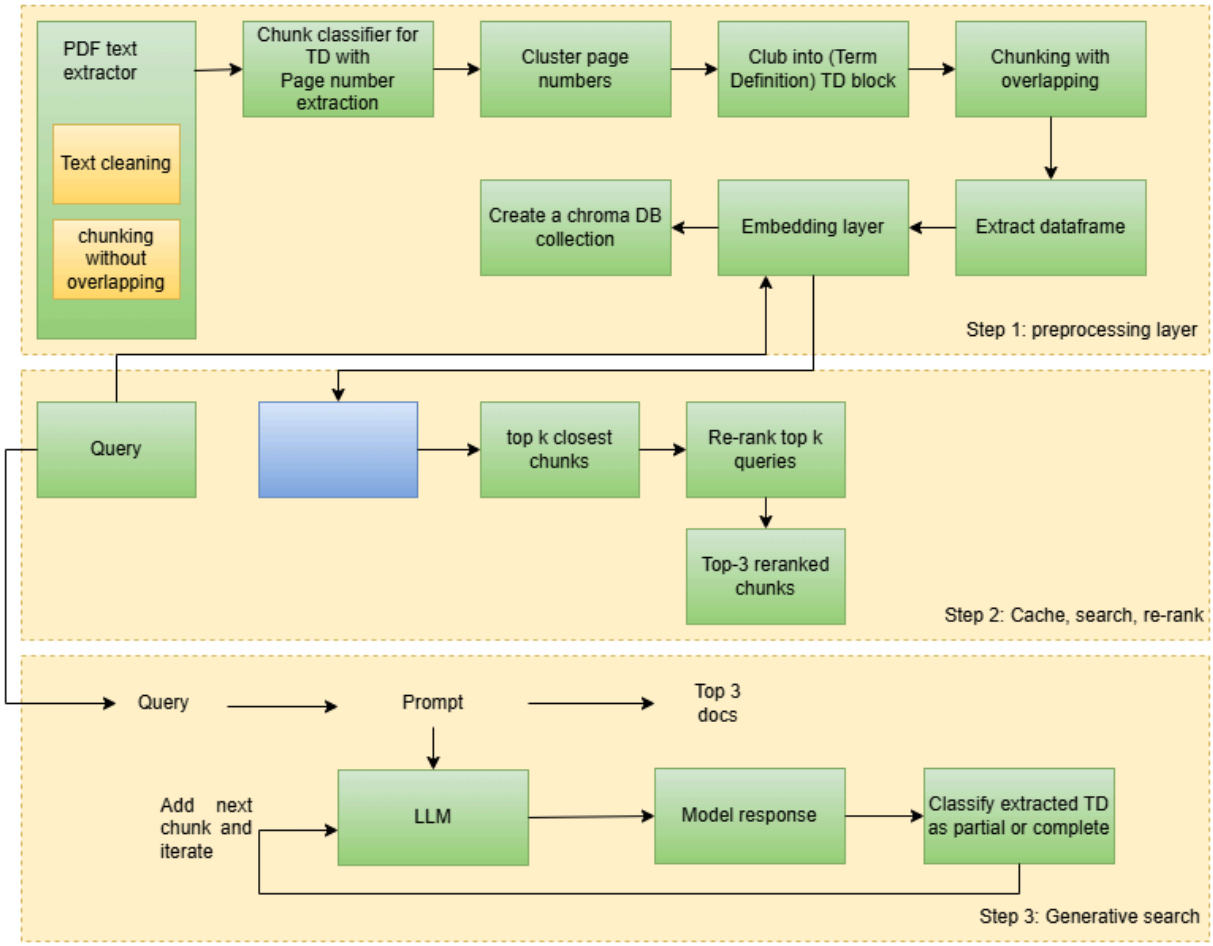
An LLM ingests the top-k chunks, generates candidate responses, and classifies outputs as complete or partial term definitions (*Figure 1, Step 3*).

#### 4.5 Analytics and Comparison Layer

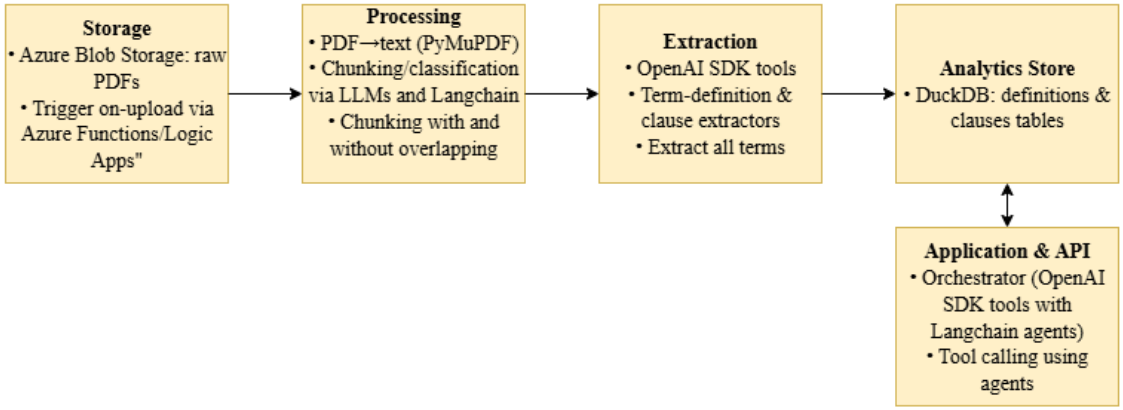
For high-speed comparison, pre-extracted definitions are stored in DuckDB. Precomputed similarity metrics (semantic + lexical + differential) enable sub-second comparison across thousands of definitions (*Figure 2*).

---

- **Figure 1**: Chroma DB pipeline (green boxes: preprocessing, cache, re-rank, generative search).




- **Figure 2**: DuckDB architecture (storage → processing → extraction → analytics



## 5 Methodology

### 5.1 Chunking Strategies

- **Without overlap**: For lightweight storage and classification.
- **With overlap**: To preserve context in multi-page definitions.

### 5.2 Extraction Pipeline

- OCR for scanned PDFs; pdfplumber for readable PDFs.
- LLM classifier for detecting definition pages.
- Embedding (OpenAI ada-002 / text-embedding-3-large).
- Chroma DB collection for retrieval.

### 5.3 Comparison Pipeline

- Pre-extraction of all terms.
- Storage in DuckDB.
- Precomputation of similarity scores (cosine, lexical, differential).
- Query-time lookup with <0.02s latency.

### 5.4 Generative Search

- LLM (GPT-4o mini in tests) processes top-k retrieved chunks.
- Iterative prompting until complete definitions are reconstructed.
- Classification into *partial/complete*.

---

## 6 Experimental Results

#### 6.1 Document Classification

- Semi-automatic shipping classifier achieved **92% accuracy**.
- PDF readability classifier: **95% accuracy, 0.52s latency**.

#### 6.2 Term Definition Extraction

- On 34 RMBS deal documents: **97% accuracy, 13.39s latency**.
- Handles multipage, nested, and long definitions.

#### 6.3 Term Definition Comparison

- Embedding + similarity pipeline: **96% accuracy, 0.5s latency**.
- DuckDB precomputation: **>95% accuracy, 0.02s latency**.

#### 6.4 Test Results from Spreadsheets

- **Extraction Tests**: Consistent accuracy across both readable and non-readable samples, including Ameriquest, Long Beach, and Morgan Stanley series.
- **Comparison Tests**: Across HE, CES, and AQ series, results aligned with ADR claims (classification as *same, minor, materially different*).

#### 6.5 Scalability

Storage feasibility confirmed: top-k comparisons for **1,000 deals fit within 1 GB**.

---

## 7 Discussion

The experiments confirm that TSU is feasible and accurate for maritime ABS, with three key insights:

1. **Latency–accuracy trade-off**: Direct retrieval yields higher latency, while DuckDB precomputation enables near-instant lookup.

2. **OCR dependency**: Accuracy is highest for readable PDFs (≥97%), slightly lower for scanned OCR (~90%).

3. **Generative augmentation**: LLMs provide interpretability by reassembling partial definitions and handling ambiguous cases.

These findings highlight a practical pathway for deploying TSU in production finance environments, bridging the gap between structured legal prose and data-driven analytics.

---

**8 Conclusion and Future Work**

We presented a layered architecture for **AI-driven term structure understanding in maritime ABS**. By combining OCR, embeddings, vector DB retrieval, and precomputed similarity analysis, the system achieves:

- 95% accuracy for extraction and comparison.
- Query latencies under 0.02s with DuckDB.
- Interpretability via generative reasoning.

**Future directions** include:

- Multimodal fusion (text + table + image embeddings).
- Domain-specific fine-tuning for maritime contracts.
- Integration with LayoutParser for structured element detection.
- Expansion to 1,000+ deal portfolios with automated refresh pipelines.

This research demonstrates that AI-based TSU can transform maritime ABS analysis, enabling scalable due diligence and market transparency.

---

**References**

[1] Narayanan, A. et al. “A semi-automatic method for document classification in the shipping industry.” Internal study, IN-D, Proceedings of Neptune's conference 2023, Samudramanthan, IIT Kharagpur

[2] Narayanan, A. et al.. ADR: “Term Definition Comparison Module.” Internal architecture decision record, June 2025, Intain Technologies Pvt Ltd

[3] Narayanan, A. et al. “Doc as Data: Current State and Future Enhancements.” Internal report, June 2025, Intain Technologies Pvt. Ltd

[4] Kanakasabai, N., Manoharan, R., et al. “Requirements analysis for term definition extraction.” Project report, April 2025, Intain Technologies Pvt Ltd.