Ocean Engineering 364 (2026) 126908

Contents lists available at ScienceDirect

Ocean Engineering

journal homepage: www.elsevier.com/locate/oceaneng

Research paper

A skill-enhanced retrieval-augmented generation expert system framework for offshore accident analysis: Combining fuzzy inference, Dempster-Shafer evidential reasoning, and expert-gated skill adaptation

Jun Ren

School of Engineering and Built Environment, Liverpool John Moores University, James Parsons Building, Byrom Street, Liverpool, L3 3AF, UK

ARTICLE INFO

Keywords:
Retrieval-augmented generation
Expert systems
Fuzzy inference
Offshore safety assessment
Floating production storage and offloading
Skills
Knowledge acquisition

ABSTRACT

Traditional rule-based expert systems for offshore safety assessment suffer from labour-intensive knowledge acquisition and an inability to evolve with operational experience. This paper presents a Skill-Enhanced Retrieval-Augmented Generation (RAG) Expert System Framework in which the "Skill" is a versioned, self-contained module encoding domain-specific fuzzy inference templates with expert-gated weight updates.

Applied to FPSO–shuttle tanker collision risk during tandem offloading, the framework compresses 245 IF–THEN rules into 12 weighted templates across five Skills, achieving Pearson r = 0.987 and F1 = 0.944 for high-risk classification. One feedback cycle confirmed that the version-control and weight-update mechanism operates as intended, with five of six proposed adjustments, ranging up to 4.7%, approved and committed by the expert panel. This result is not a claim of superior risk-assessment accuracy relative to the established baseline. It is evidence that the knowledge-engineering architecture, grounded in maintainability, modularity, and traceability, can be put into practice and audited at every step. The contribution is architectural, not computational.

The framework is positioned as a feasibility-oriented knowledge-engineering contribution: it demonstrates that Skill-structured RAG can support maintainable, modular, and traceable offshore safety knowledge bases, not that it outperforms established expert-assessment methods on general offshore risk problems.

1. Introduction

Floating Production, Storage, and Offloading (FPSO) units represent a prevalent and technologically significant system within the offshore oil and gas industry. These vessels, visually resembling ships but engineered with distinct operational capabilities, integrate all requisite production and processing facilities for the extraction and initial storage of crude oil recovered from seabed wells. Globally, FPSOs constitute over 50% of operational floating production systems, with a current fleet of approximately 70 units, including 15 operating on the UK Continental Shelf (UKCS) (HSE, 2007). The primary method of crude oil transfer from FPSOs to shore relies on shuttle tankers, utilizing a 'tandem loading' process – the directed loading of cargo from the FPSO's stern to the tanker's bow – coupled with sophisticated dynamic positioning and emergency shutdown systems for maintaining safe operational distances. This methodology is critical for mitigating risk associated with the inherently challenging offshore environment."

The tandem offloading process, involving the close proximity of FPSOs/Floating Storage Units (FSUs) and shuttle tankers, presents a complex operational challenge characterized by inherent risks. Operational issues, including excessive shuttle tanker motion, potential errors by Dynamic Positioning (DP) operators, and anomalous interactions between Dynamic Positioning (DP) and power management systems (PMS) on board the vessels, significantly contribute to these risks. These interactions have resulted in documented consequences ranging from increased fuel consumption to, in some instances, incidents with the potential for personnel injury, environmental contamination, or vessel damage. Research shown that contact incidents in the North Sea involving FPSOs and shuttle tankers underscore the high probability of vessel collision during tandem offloading operations (Chen et al., 2002).

_ Emerald FSU: Impact by shuttle tanker *Navion Clipper*, UK, 28 February 1996.
_ Gryphon FPSO: Impact by shuttle tanker *Futura*, 26 July 1997.
_ Captain FPSO: Impact by shuttle tanker *Aberdeen*, 12 August 1997.
_ Schiehallion FPSO: Impact by shuttle tanker *Nordic Savonita*, 25 September 1998.
_ Norne FPSO: Impact by shuttle tanker *Knock Salliea*, 5 March 2000.

E-mail address: j.ren@ljmu.ac.uk.

https://doi.org/10.1016/j.oceaneng.2026.126908
Received 27 February 2026; Received in revised form 14 June 2026; Accepted 4 July 2026
Available online 8 July 2026

Recently no widely reported FPSO–shuttle tanker collision accidents comparable to these classic cases have been documented in the North Sea. However, recent technical and modeling work continues to treat tandem offloading collision as a high-probability hazard and to update models and design guidance. Recent publications (Gaidai et al., 2023; Jiang, 2014) describe offloading interactions between FPSO and shuttle tanker as remaining critical for design due to collision risk, with relative motions and hawser/fender loads being key for avoiding contact. Risk modeling frameworks continue to employ phase-based collision probability models, such as P(Collision) = P(unsafe failure mode) × P(failure of recovery | UFM), which remain referenced in contemporary tandem offloading risk analyses.

Recognizing the potential for significant operational consequences, the UK Health and Safety Executive (HSE) has established a target to reduce shuttle tanker 'loss of station keeping events' by 25%, predicated on a baseline of seven incidents per vessel annually. This initiative underscores the critical need for robust safety management systems, demanding demonstrable hazard identification, risk reduction to an 'as low as reasonably practicable' standard, and the implementation of comprehensive management controls including designated safe refuges and provisions for emergency evacuation. The inherently complex nature of tandem loading/offloading operations, particularly within dynamic marine environments, necessitates a proactive approach to safety assessment, demanding thorough analysis of potential system failures at the design phase. However, traditional ship/platform collision-risk models, which frequently rely on probabilistic hazard assessment and consequence modelling, often prove inadequate for this operation. Such approaches struggle to effectively address situations characterized by data scarcity, operational ambiguity, and the interplay of technical, human, and organizational vulnerabilities. Furthermore, prevailing quantitative risk analyses in offshore engineering tend to prioritize technical aspects, yielding a hardware-centric risk control strategy-a limitation when confronting the nuanced complexities of marine operations. Therefore, a robust, empirically-driven collision-risk model that specifically tailored to the FPSO-tanker offloading operation is essential. This model must account for the operational safety levels of FPSOs across diverse failure scenarios, allowing for proactive mitigation strategies from the outset. Moving beyond purely technical assessments, such a model requires an integrated understanding of human factors and organizational controls, ultimately contributing to a more holistic and effective safety management framework.

Rule-based expert systems offer a powerful approach to modelling expert knowledge within complex risk assessment domains, particularly scenarios like FPSO-tanker offloading. These systems leverage structured rules that typically formulated as 'if-then' statements, combined with fuzzy logic to handle inherent uncertainties and imprecise information (Feigenbaum, 1977; Zadeh, 1965). Unlike purely numerical models, rule-based systems represent the nuanced judgment of experienced professionals.

Fuzzy logic allows for the representation of vague or subjective concepts, such as 'moderate' or 'high' risk, by assigning partial truth values to rules. For example, a rule could state: "If the wind speed is high AND the wave height is significant, THEN the risk of collision is 'moderate'."

By integrating these structured rules, coupled with fuzzy logic, expert systems can effectively capture the complexities of operational environments, accounting for human factors, dynamic conditions, and incomplete data. This approach moves beyond simplistic hazard identification, providing a framework for prioritizing risks and supporting informed decision-making in safety-critical operations like FPSO-tanker offloading. Ultimately, the aim is to replicate the reasoning process of an expert, translating experience into actionable risk management strategies.

To effectively manage the complexity inherent in risk assessment, particularly within a structure like the 245-rule framework, employing linguistic variables is crucial. Rather than relying solely on numerical thresholds, this approach utilizes descriptive terms such as 'frequent' failure rate, 'catastrophic' severity, or 'high' probability to represent risk assessments. This significantly enhances the interpretability and practicality of the rule base. These linguistic variables allow experts to articulate their judgment in a manner readily understood by both technical and non-technical stakeholders. For instance, a rule might state: "If the failure rate is 'frequent' AND the severity is 'catastrophic', THEN implement immediate mitigation measures."

By utilizing this vocabulary, the rule base becomes more nuanced and responsive to operational realities. It shifts from a purely quantitative to a qualitative assessment, incorporating human factors and subjective evaluations. Furthermore, the use of descriptive terms facilitates communication and collaboration amongst diverse teams involved in risk management, fostering a shared understanding and ultimately contributing to more effective decision-making within the expansive rule set (Ahmed et al., 2019).

The scalability of expert systems introduces a significant auditability gap. As rule bases expand maintaining, interpreting, and rigorously auditing these systems becomes increasingly problematic. The core issue arises from the exponential complexity of managing interconnected rules and their potential interactions. Simply generating a decision output, such as "rule-L7_rule-C5_rule-E7": "poor," provides insufficient justification for auditors and stakeholders. This lacks the narrative context needed to understand the reasoning behind the decision including which rules triggered the assessment, and to what degree (Ren et al., 2005; Wang et al., 2015).

Without this traceability, demonstrating compliance, validating the system's logic, and identifying potential biases become exceptionally difficult. The 'black box' nature of complex rule-based systems hinders accountability and trust. Addressing this requires incorporating mechanisms for documenting the decision-making process, potentially through rule lineage tracking or generating comprehensive audit trails, ensuring transparency and facilitating robust verification.

This paper introduces a Skill-enhanced RAG-based framework to augment rule-based expert systems, transforming them into dynamic and auditable knowledge bases without altering their core logic. This paper demonstrates how Skill/RAG technology can provide a "Chain of Reasoning" for every decision, enhancing transparency and maintainability.

The remainder of this paper is structured as follows, addressing three primary objectives: (1) to formalise the Skill concept and demonstrate its derivation from the 245-rule Ren et al. (2005) knowledge base through a principled compression to 12 weighted inference templates (Section 3), (2) to specify the full Skill-Enhanced RAG pipeline in sufficient detail for independent replication, including embedding model configuration, retrieval strategy, Skill-aware query formulation, and hallucination-control mechanisms (Sections 3-4), and (3) to validate the integrated framework quantitatively against the original expert assessments using Pearson correlation (r = 0.987), F1 score (0.944), MAE (0.031), and RMSE (0.039), and to characterise its behaviour under stress-case and sensitivity scenarios (Section 4). Section 5 discusses limitations and future directions.

2. Literature review

2.1. Expert systems and reasoning under uncertainty

Expert systems encode domain knowledge as rules and apply inference mechanisms to produce decisions. They have been applied in medical diagnosis, financial planning, engineering design, and risk assessment since the 1970s (Feigenbaum, 1977). A typical architecture consists of a knowledge base (database and rule base), an inference engine, and a user interface.

Two mathematical frameworks are central to expert systems for risk management under uncertainty: fuzzy set theory and evidential reasoning.

Fuzzy set theory, proposed by Zadeh (1965), handles vagueness in human judgement. Classical set theory assigns elements fully in or fully out of a set. Fuzzy sets allow partial membership, expressed through membership functions that map elements to values between 0 and 1. Linguistic variables like "high risk", "moderate importance", and "very frequent" are used by experts in practice, where these terms correspond to specific fuzzy numbers for clear expression of opinions. Triangular fuzzy numbers (TFNs), defined by a triple (l, m, u), are widely used for their computational simplicity (Chang, 1996). A fuzzy inference engine applies IF-THEN rules to fuzzified inputs and produces fuzzy outputs, which can be defuzzified into crisp values when needed.

Evidential reasoning, based on Dempster-Shafer (D-S) theory (Dempster, 1968; Shafer, 1976), handles uncertainty in combining evidence from multiple sources. Unlike Bayesian probability, D-S theory can represent ignorance explicitly through a mass function assigned to the full frame of discernment. Yang (2001) and Yang and Xu (2002) developed the evidential reasoning algorithm for multi-attribute decision analysis under uncertainty, enabling the systematic combination of assessments from multiple experts across multiple criteria. Yang et al. (2022) subsequently advanced this line by proposing a highly explainable cumulative belief rule-based system with interpretable rule weighting, further strengthening the transparency of evidential reasoning outputs.

The combination of fuzzy inference and evidential reasoning is well-established. Ren et al. (2005; see also Wang et al., 2015) built an offshore safety framework using four modules: a fuzzy mapping module, a knowledge base with 245 IF-THEN rules, a fuzzy inference engine, and a Dempster-Shafer synthesis module. Ren et al. (2009) subsequently extended this line of work using fuzzy Bayesian networks for offshore risk quantification. Liu et al. (2005) applied fuzzy rule-based evidential reasoning to engineering system safety analysis. These systems are effective but share a common bottleneck: all knowledge must be manually elicited from domain experts and encoded into rules and databases. As domains grow more complex and document volumes increase, this bottleneck becomes the binding constraint.

Recent applications of rule-based fuzzy inference in the maritime domain illustrate the breadth and maturity of this approach. Ceylan (2023) applied rule-based fuzzy FMEA to shipboard compressor systems, constructing 125 IF–THEN rules from expert input and identifying broken piston rings as the highest-risk failure mode (Fuzzy RPN = 6.43) among 33 failure modes. Karanović et al. (2024) extended fuzzy FMEA to four-ram hydraulic steering systems, ranking 47 failure modes and finding insufficient system pressure (FRPN = 6.59) as the primary risk. Ceylan (2025) integrated control theory with fuzzy Fine–Kinney analysis for boiler automation systems in Maritime Autonomous Surface Ships (MASS), addressing the risk profile of minimally supervised shipboard systems. These studies confirm the maturity of rule-based fuzzy inference in maritime safety assessment; however, they retain static rule bases without automated knowledge acquisition or adaptive weight updating, which are limitations that the present framework addresses through its Skill-Enhanced RAG architecture.

Beyond the knowledge acquisition bottleneck, traditional expert systems face a knowledge maintenance bottleneck: once rules are encoded, they remain static unless manually revised. Domains evolve, as new regulations emerge, incident patterns change, and expert understanding deepens, but rule bases do not update themselves. This creates a growing gap between the encoded knowledge and the current state of the domain. Adaptive and learning expert systems have been explored since the 1990s (Compton and Jansen, 1990), but these approaches lacked both the scalable knowledge acquisition mechanism that RAG provides and the formal uncertainty reasoning that fuzzy-evidential methods offer.

2.2. Rule-based frameworks in offshore safety

A typical architecture of the offshore safety assessment framework is depicted in Fig. 1 (Ren et al., 2005). The system consists of four principal components: fuzzification module, knowledge base module, fuzzy inference module, and synthesis module. As can be seen, fuzzy reasoning is the key mechanism in the proposed system. In fact, fuzzy reasoning mimics human reasoning procedure by using fuzzy set theory and fuzzy logic in an environment of uncertainty and imprecision. It is based on fuzzy expert system implying a collection of fuzzy membership functions and rules. The fuzzification module performs a scale mapping that changes the range of values of input variables into corresponding universe of discourse, and fuzzification that converts non-fuzzy (crisp) input data into suitable linguistic values. The knowledge base consists of database and fuzzy rule base. The database is a set of fuzzy membership functions. The fuzzy rule base consists of a set of linguistic fuzzy associative rules written in the IF-THEN form. The fuzzy associative rules help in the decision-making process. After the approximate reasoning module is triggered with rules from the fuzzy rule base, it can infer the fuzzy output from fuzzified inputs. The fuzzy output is then taken as input of the synthesis module, which is based on an evidential reasoning mechanism. The synthesis module is designed to deal with multi-attributes and multi-experts' decision-making problems.

2.3. Retrieval-augmented generation

Large language models (LLMs) have demonstrated strong capabilities in text comprehension, information extraction, and structured output generation. However, they suffer from well-documented limitations: hallucination (generating plausible but false information), lack of source traceability, and inability to update knowledge without retraining (Lewis et al., 2020).

Retrieval-Augmented Generation (RAG) addresses these limitations. Introduced by Lewis et al. (2020), RAG combines parametric LLM knowledge with non-parametric retrieval from an external knowledge store. In a standard RAG pipeline, a query is encoded into a vector embedding, matched against a pre-indexed document store via semantic similarity, and the retrieved context is injected into the LLM prompt to constrain and ground the response. This architecture suppresses hallucination and makes outputs traceable to source documents (Lewis et al., 2020).

RAG has proven effective in knowledge-intensive tasks where factual accuracy matters: legal document analysis, clinical question-answering, and technical support (Lewis et al., 2020). Bai et al. (2025) demonstrated a specific application in decision support systems: using metadata-filtered RAG to generate auditable explanations for a general expert system's decisions.

2.4. Research gap and contribution

The literature reveals four gaps:

- **No integration of RAG with formal expert system reasoning for Offshore Accident analysis.** Existing offshore incident analysis frameworks rely on manual knowledge acquisition. Existing RAG applications focus on question-answering, not on feeding structured evidence into formal reasoning systems.
- **LLM-only approaches lack formal reasoning.** Using an LLM directly for risk assessment cannot provide the uncertainty quantification, evidence fusion, or mathematical auditability that fuzzy inference and evidential reasoning offer.
- **Expert-system-only approaches lack scalable knowledge acquisition.** Traditional expert systems require all knowledge to be manually elicited and encoded. This process does not scale to the growing volume of relevant documents, data, and expert opinions.
- **Static rule bases lack adaptability.** Even when knowledge is successfully acquired and encoded, traditional rule bases remain fixed until manually revised. They do not learn from accumulated decision outcomes, adapt to changing domain conditions, or allow systematic

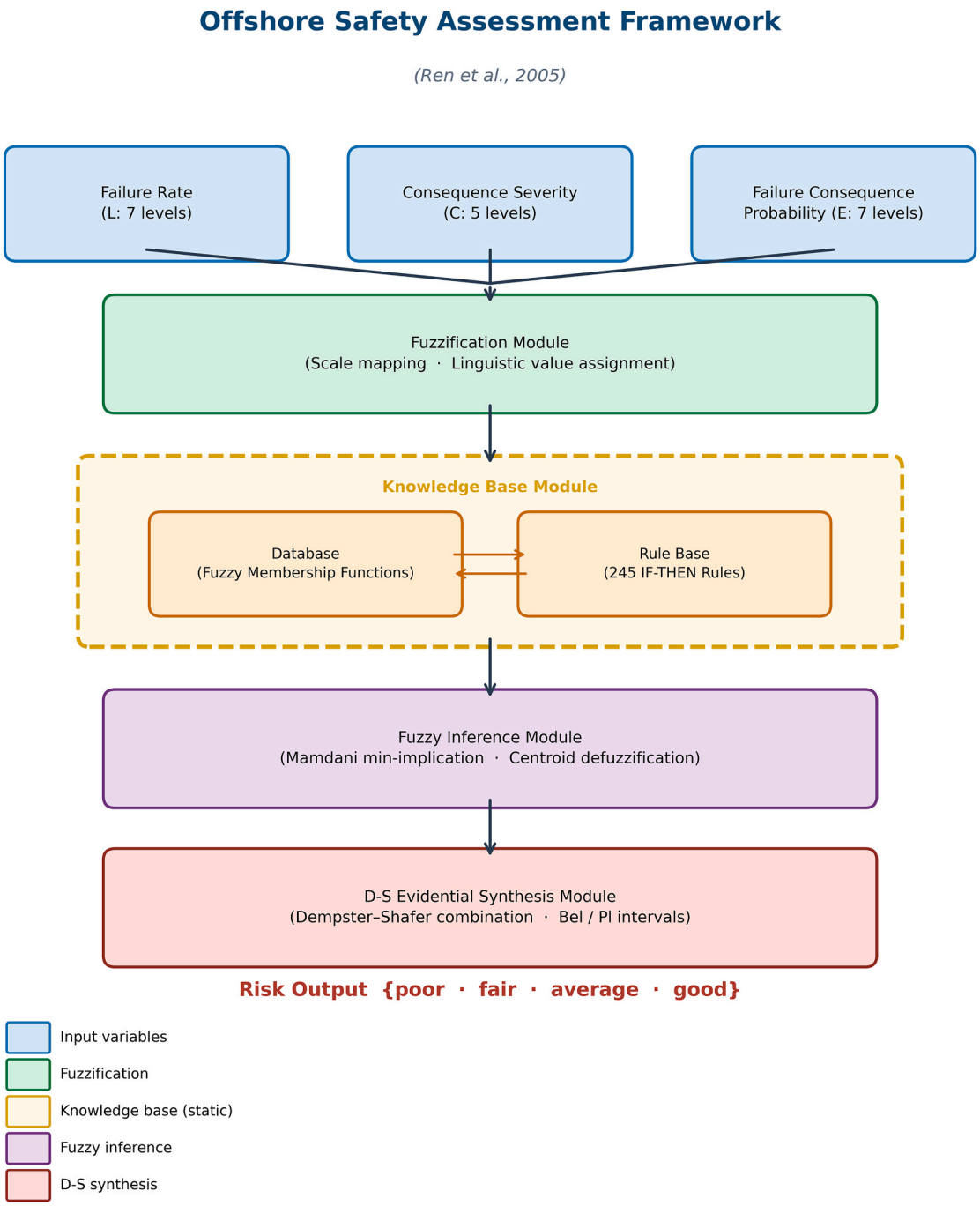

Fig. 1. Offshore safety assessment framework.

performance benchmarking. This knowledge maintenance bottleneck compounds the knowledge acquisition bottleneck over time.

These four gaps are not independent: they represent a compounding progression in which the absence of scalable knowledge acquisition (Gap 3) forces manual rule encoding, which in turn produces static rule bases (Gap 4) that cannot respond to evolving incident evidence, while the absence of formal reasoning (Gap 2) means LLM-only approaches cannot provide the uncertainty quantification and auditability that safety regulation requires. No existing framework addresses all four gaps simultaneously: Yang et al. (2022) address adaptability through cumulative belief rule bases but retain manual knowledge acquisition. Bai et al. (2025) integrate RAG with knowledge graphs but do not employ fuzzy-evidential reasoning for structured uncertainty quantification. The proposed Skill-Enhanced RAG framework is specifically designed to close all four gaps within a single coherent architecture.

This paper fills these gaps by proposing a framework that (a) combines Skill-enhanced RAG for scalable knowledge acquisition with fuzzy inference and evidential reasoning for formal risk assessment, and (b) introduces decision-type Skills that replace the static rule base with adaptive, learned reasoning modules refined through expert-gated feedback. The framework is designed to be domain-portable: the core reasoning pipeline contains no FPSO-specific hard-coded logic, and adapting it to a new domain requires reconfiguring the input document corpus and Skill definitions without modifying the core architecture. This portability is an architectural design property; empirical demonstration across additional domains is identified as future work.

3. The proposed framework

3.1. Framework overview

The proposed framework integrates Retrieval-Augmented Generation (RAG) with a fuzzy-evidential expert system for offshore accident analysis, enhanced by decision-type Skills that replace the traditional static rule base with adaptive, learned reasoning modules. Fig. 2 illustrates the architecture.

The framework has six layers (Layers 1-6) with Skills embedded across Layers 2–5, plus an expert-gated feedback loop connecting the output back to the reasoning core:

- **Input Layer** - heterogeneous domain knowledge sources. Skill-Aware RAG Layer - retriever, vector database, and grounded LLM with Skill-driven query formulation for targeted knowledge extraction,
- **Skill-Routed Evidence Extractor** - bridge between RAG outputs and expert system inputs, with a Skill Router that selects the appropriate decision-type Skill,
- **Skill-Based Reasoning Core** - database, Skill Repository (replacing the static rule base), and Skill-driven fuzzification and fuzzy inference,
- **Skill-Guided Synthesis Module** - Dempster-Shafer evidential reasoning with Skill-informed weighting,
- **Output Layer** - risk rankings, risk profiles, traceability, and Skill trace,
- **Expert-Gated Feedback Loop** - captures decision outcomes, benchmarks Skill performance, and refines Skills subject to expert approval.

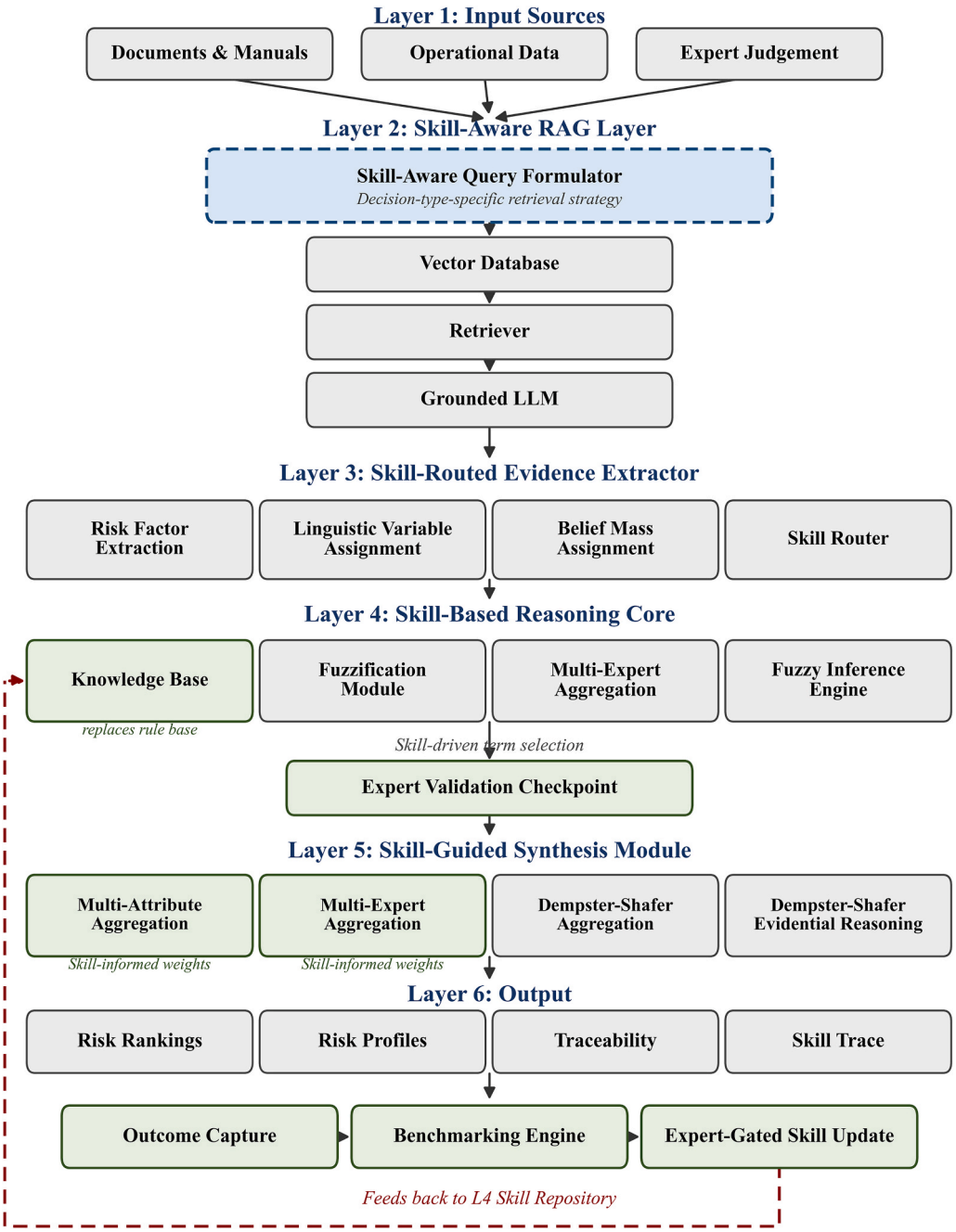

Fig. 2. Skill-enhanced RAG expert system framework.

The design philosophy rests on a three-part principle in which RAG handles knowledge acquisition and the expert system handles formal reasoning, by introducing a third element in which Skills handle knowledge adaptation. RAG solves the input bottleneck, the expert system provides mathematical rigour. Skills ensure the reasoning evolves as evidence accumulates.

3.2. Domain-specific skills

3.2.1. Skill definition

A Skill is a reusable, domain-specific reasoning module defined as a tuple $S = \langle \tau, P, F, T, B \rangle$, where:

- $\tau \in \{SI, MM, EH, HF, SysI\}$ is the hazard domain (Structural Integrity, Maintenance Management, Environmental Hazard, Human Factors, System Integration),
- $P = \{p_1, p_2, \ldots, p_m\}$ is the set of decision patterns: structured mappings from input evidence characteristics to output configurations,
- $F = \{(l_j, \mu_j)\}$ is the fuzzification configuration: linguistic terms with corresponding TFN membership functions,
- $T = \{(T_1, w_1), (T_2, w_2), \ldots, (T_n, w_n)\}$ is the set of inference templates with learned weights $w_i \in [0,1]$,
- $B = \{\gamma, \delta, \gamma\_T, v, d\}$ is the performance benchmark: expert agreement rate, mean deviation, per-template metrics, version number, and approval date.

Unlike static IF-THEN rules, which encode a single expert's judgement at a fixed point in time, Skills are learned from accumulated decision outcomes and refined through an expert-gated feedback loop (Section 3.9).

The Skill concept is formally distinguished from related adaptive expert-system approaches. Adaptive neuro-fuzzy inference systems (ANFIS) adapt membership-function parameters but operate over a single global rule base without domain decomposition, retrieval-based evidence grounding, or version-controlled knowledge evolution. Parameterised fuzzy modular systems organise rules into domain groups but do not incorporate automated evidence routing or benchmarking-derived template weights. A Skill differs in four specific respects: (i) the Skill Router selects the active Skill at inference time based on the evidence profile E(r), enabling automated domain routing without manual rule selection, (ii) template weights w_i are learned from accumulated expert-validated outcomes via the Benchmarking Engine rather than prescribed by a knowledge engineer, (iii) each Skill carries its own fuzzification configuration F, allowing domain-calibrated TFN ranges in place of a single global membership scale, and (iv) the version-controlled Skill Repository ensures that all weight updates are traceable, auditable, and reversible, a property absent from conventional parameterised fuzzy modules. Table A1 in the Appendix formalises this comparison across all six differentiating properties.

3.2.2. Hazard-domain organisation

Skills are organised by hazard domain, each encapsulating the reasoning patterns, membership functions, and inference templates specific to a class of offshore safety hazards. This domain-specific organisation ensures that each Skill captures the distinct evidence types, failure modes, and expert reasoning associated with its hazard class. The framework defines five core Skills for offshore safety assessment (see Table 1):

3.2.3. Skill design

Each Skill is authored as a self-contained document composed of two parts: a structured metadata header and a prose body. This design follows the YAML-plus-Markdown paradigm widely adopted in modern AI agent frameworks, where the header carries machine-readable fields and the body provides human-readable reasoning content that is loaded into context on demand.

Table 1
Five core skills for offshore safety assessment.

Skill	Hazard Domain	Input Evidence Type	Output
Structural Integrity	Structural degradation, fatigue, corrosion	Inspection records, material test reports, corrosion rates	Structural risk score with belief interval
Maintenance Management	Maintenance schedules, inspection intervals, degradation	Maintenance logs, work orders, equipment condition data	Maintenance effectiveness score with belief interval
Environmental Hazard	Weather extremes, sea state, environmental loading	Metocean data, environmental monitoring, incident reports	Environmental risk score with belief interval
Human Factors	Procedural compliance, operator error, training adequacy	Training records, incident reports, human reliability analyses	Human factors risk score with belief interval
System Integration	Control system interactions, cascading failures, safety barriers	System architecture data, HAZOP records, barrier performance	System integration risk score with belief interval

Metadata header. The YAML front-matter declares the Skill's identity and routing information: a unique name (e.g., structural-integrity), a natural-language description that doubles as the trigger condition for the Skill Router, a version tag managed by the feedback loop (Section 3.9), a domain tag linking to the hazard-domain taxonomy (Table 1), and references to the source documents from which the Skill's evidence base was constructed.

Reasoning body. The Markdown body encodes the Skill's operational content: (i) the set of fuzzy inference templates T = {T_1, …, T_n} with their learned weights w_i, membership-function definitions, and IF-THEN rule structures, (ii) linguistic-variable mappings that translate RAG-extracted evidence into fuzzy inputs, (iii) domain-specific calibration notes (e.g., failure-rate thresholds for FPSO versus fixed-platform contexts), and (iv) a change log recording each version update approved through the expert gate.

This two-part format offers three practical advantages. First, the metadata header enables automated routing: the Skill Router (Section 3.5) matches an incoming query's keywords against Skill descriptions without parsing the full body. Second, the Markdown body is both human-auditable and machine-consumable, supporting the expert-gated review process. Third, Skills are portable-transferring the framework to a new domain (e.g., from FPSO to fixed-platform assessment) requires authoring new Skill documents rather than re-engineering the inference engine.

3.2.4. Skills versus static rules

The Skill approach does not abandon fuzzy IF-THEN inference. Rather, it organises inference into decision-type modules whose rule structures (inference templates) carry learned weights that are refined over time, and whose parameters are adapted to the specific decision context. Table 2 shows the comparison of the two approaches.

Beyond the technical distinctions captured in Table 2, the Skill-based architecture delivers five engineering benefits that are absent from static rule bases. (i) Maintainability: each Skill is an independently editable document; updating corrosion-rate thresholds in the Structural Integrity Skill leaves all other Skills unchanged, eliminating the fragility of monolithic rule-base edits. (ii) Modularity: the five hazard-domain Skills are decoupled—adding a new hazard domain requires authoring one new Skill module without restructuring existing ones. (iii) Version control: every approved update increments the Skill version number and records the benchmarking evidence and approving expert, enabling full audit trails and rollback to any prior state. (iv) Traceability: every risk

Table 2
Skills versus static rules.

Property	Static Rule Base	Skill Repository
Creation	Hand-crafted by knowledge engineers	Learned from RAG-extracted evidence + expert validation
Maintenance	Manual rewriting when domain evolves	Expert-gated feedback loop updates incrementally
Granularity	Individual IF-THEN rules (e.g., 245 rules)	Decision patterns per decision type
Selection	All matching rules fire simultaneously	Skill Router selects appropriate Skill
Adaptability	Fixed once encoded	Evolves via benchmarked outcomes; version-controlled
Domain transfer	Rules rewritten for each domain	Skills transfer; only input evidence changes

output links through the Skill trace to the specific template, weight, retrieved evidence chunk, and source document that produced it, supporting regulatory compliance and post-incident review. (v) Extensibility: the Skill Repository can grow incrementally—new inference templates are added within an existing Skill and immediately participate in routing—without retraining or restructuring the broader framework. These properties represent the primary engineering rationale for introducing the Skill abstraction, and are independent of any accuracy comparison with the static rule-base baseline.

3.3. Input Layer

The framework accepts three categories of input:

- **Documents and Manuals:** domain-specific standards, regulations, guidelines, audit reports, published research, and case studies. In maritime industries, these include The UK HSE offshore accident analysis reports, FPSO-specific reference literature, IMO's Global Maritime Security Programme and research publications.
- **Operational Data:** quantitative and semi-quantitative records from operations. In Offshore operations: Field Development Plans (FDP), Environmental Impact Assessments (EIA), and daily, digital, or manual production logs. Key documents include technical reports on reservoir, mooring, and hull, as well as regulatory and safety compliance documents like OPRED guidance notes.
- **Expert Judgement:** qualitative assessments from domain practitioners, captured through structured questionnaires, semi-structured interviews, or assessment forms.

These three input types are heterogeneous and largely unstructured. Converting them into the structured formats required by an expert system, such as fuzzy membership values, pairwise comparison matrices, and belief mass functions, represents the knowledge acquisition bottleneck that the RAG layer addresses.

3.4. Skill-Aware RAG layer

3.4.1. Skill-Aware Query Formulator

Different decision types require different types of evidence. When the system is assessing likelihood, it needs frequency data, incident reports, and trend analyses. When assessing impact, it needs consequence records, loss estimates, and severity case studies. The Skill-Aware Query Formulator addresses this by constructing decision-type-specific queries.

Given a risk factor r and an active Skill S, the Query Formulator generates a structured query q(r, S) that: (1) specifies what type of evidence is needed (determined by S), (2) applies metadata filters to the vector database to prioritize relevant document types, and (3) formulates the query in natural language aligned with the decision question.

3.4.2. Vector database

All input documents are pre-processed, chunked, and embedded into a vector database (Ori Yoran et al., 2023). The chunking strategy segments documents at the paragraph or section level, preserving semantic coherence within each chunk. An embedding model (e.g. text-embedding-004) converts each chunk into a dense vector representation. The Text embeddings API converts textual data into numerical vectors. These vector representations are designed to capture the semantic meaning and context of the words they represent, and the vector database stores these embeddings alongside metadata: source document, section title, document type, date, and evidence category tags that enable Skill-aware filtering.

3.4.3. Retriever

Given the Skill-aware query q(r, S), the retriever performs a semantic similarity search against the vector database, applying the metadata filters specified by the active Skill. It returns the top-k most relevant document chunks, ranked by cosine similarity.

3.4.4. Grounded LLM

The grounded LLM receives the retrieved document chunks as context and generates structured outputs. Three design constraints ensure reliability:

- **Grounding:** the LLM is instructed to use only information from the retrieved context, not its parametric knowledge. This suppresses hallucination.
- **Structured output:** the LLM produces outputs in predefined, Skill-specific formats including risk factor tables, severity assessments in linguistic terms, and categorised evidence.
- **Source citation:** every extracted fact must cite the source document chunk from which it was derived.

The LLM does not make risk decisions. It extracts and structures information. All risk assessment decisions are made by the Skill-Based Reasoning Core.

3.5. Skill-Routed Evidence Extractor

The Skill-Routed Evidence Extractor bridges the RAG layer and the Skill-Based Reasoning Core. It performs five functions:

Risk factor extraction and categorisation. The extractor parses the LLM's structured output to identify distinct risk factors and assign them to predefined categories following the five-domain taxonomy defined in Section 3.2.

Linguistic variable assignment. Qualitative descriptions from the LLM output are mapped to linguistic terms from a predefined scale (e.g., *Very Low, Low, Medium, High, Very High*). These correspond to triangular fuzzy numbers used by the fuzzification module.

Belief mass function construction. For evidential reasoning, the extractor converts evidence strength into basic probability assignments (BPAs). Strong, corroborated evidence receives a high mass assignment, weak evidence assigns mass to the full frame of discernment, representing uncertainty.

Skill Router. Based on the extracted risk factor type and the assessment task required, the Skill Router selects the appropriate Skill from the Skill Repository. The Router operates at two levels: (1) assessment pipeline routing - for each risk factor, the system runs Likelihood, Impact, and Vulnerability Skills sequentially, feeding into the Prioritisation Skill, and (2) context-aware selection - within each stage, the Router may select Skill variants based on evidence availability.

Formally, for a risk factor r with evidence profile E(r) and assessment stage t, the Skill Router selects: $S^* = \text{argmax}\ \varphi(S, E(r))$ for $S \in \Psi\_t$, where Ψ_t is the set of Skills applicable at stage t, and $\varphi(S, E(r))$ is a compatibility function measuring how well Skill S matches the evidence profile

of r.

The compatibility function φ (S, E(r)) is defined as the keyword-weighted cosine similarity between the Skill's routing descriptor vector d(S) and the evidence profile vector e(r): φ (S, E(r)) = [d(S). e(r)]/(|| d(S)|| ||e(r)||), where each dimension of d(S) represents the normalised frequency of domain-specific routing keywords associated with Skill S, and each dimension of e(r) represents the corresponding keyword frequency in the evidence profile of risk factor r. The Skill Router assigns the highest-scoring Skill as the primary assignment S* and flags any secondary Skills with phi above a confidence threshold of 0.70 for parallel evidence routing.

Expert validation checkpoint. The extracted evidence is presented to domain experts for review. In the Skill-enhanced framework, this checkpoint also validates the Skill selection: the expert confirms that the Router has selected the appropriate Skill for the assessment task.

3.6. Skill-Based Reasoning Core

3.6.1. Database

The database stores validated risk factors, their categories, severity assessments, interrelationships, and historical data. This component is unchanged from the original framework.

3.6.2. Skill Repository

The Skill Repository replaces the static rule base. Each Skill contains: decision patterns (learned from accumulated expert-validated outcomes), inference templates with learned weights, fuzzification parameters, and performance benchmarks.

For example, the Structural Integrity Skill contains inference templates such as:

- Template SI-1 (weight $w_1 = 0.95$): IF corrosion_rate is {HIGH|MEDIUM|LOW} AND inspection_interval is {OVERDUE|ON_SCHEDULE|RECENT}, THEN structural_risk is {computed}
- Template SI-2 (weight $w_2 = 0.96$): IF fatigue_crack_length is {CRITICAL|MODERATE|MINOR} AND material_grade is {HIGH|STANDARD}, THEN structural_risk is {computed}

The weights (0.95, 0.96) are learned from the benchmarking process (Section 3.9), not assigned by a knowledge engineer. Templates that consistently produce expert-aligned outputs receive higher weights. The complete 12-template inventory with initial and post-feedback weights is provided in Table A2 (Appendix).

3.6.3. Skill-driven fuzzification module

The fuzzification module converts crisp values and linguistic assessments into fuzzy membership values using triangular fuzzy numbers (TFNs). The active Skill determines the fuzzification parameters i.e which input variables get fuzzified and how linguistic terms map to these variables. The base linguistic scale follows established conventions (see Table 3):

3.6.4. Skill-driven fuzzy inference engine

The fuzzy inference engine applies the active Skill's inference templates to fuzzified inputs. The engine retains the Mamdani min-implication operator (Mamdani and Assilian, 1975). The key enhancement is weighted template aggregation.

Let $T_1, T_2, \ldots, T_n$ be the inference templates that fire, with learned weights $w_1, w_2, \ldots, w_n$ and firing strengths $\alpha_1, \alpha_2, \ldots, \alpha_n$. The weighted firing strength for template T_i is: $\tilde{\alpha}_i = w_i \cdot \alpha_i$

The aggregated output fuzzy set is: $\mu\_out(z) = \max \min(\tilde{\alpha}_i, \mu\_C_i(z))$ for $i = 1, \ldots, n$

For defuzzification, the centroid method is applied: $\hat{z} = \int z \cdot \mu\_out(z)\, dz / \int \mu\_out(z)\, dz$.

When all template weights equal 1, the formulation reduces to standard Mamdani inference, ensuring backward compatibility with the original framework.

Table 3
Linguistic scale and triangular fuzzy numbers for risk assessment.

Linguistic Term	TFN (l, m, u)
Very Low (VL)	(0, 0, 0.25)
Low (L)	(0, 0.25, 0.5)
Medium (M)	(0.25, 0.5, 0.75)
High (H)	(0.5, 0.75, 1.0)
Very High (VH)	(0.75, 1.0, 1.0)

3.7. Skill-Guided Synthesis Module: evidential reasoning

3.7.1. Frame of discernment

For risk assessment, the frame of discernment Θ defines the set of mutually exclusive hypotheses. Depending on the assessment objective, Θ may represent either risk severity levels, where Θ = {Very Low, Low, Medium, High, Very High}, or causal categories, where Θ = {Design Defect, Maintenance Failure, Operational Error, Environmental, Unknown}. The former is used when the goal is to quantify the overall safety level, while the latter supports root-cause attribution. In both cases, each expert's assessment is expressed as a basic probability assignment (BPA), a mass function m that distributes belief across subsets of Θ.

3.7.2. Skill-informed evidence weighting

In the Skill-enhanced framework, skill benchmarks inform aggregation weights. If the Structural Integrity Skill has a higher benchmark accuracy than the Environmental Hazard Skill, correspondingly higher reliability can be assigned to structural-integrity-based evidence during synthesis. This weight guidance is advisory, not mandatory. The weighting adjustment is bounded: the maximum relative weight ratio between any two Skills is capped at 1.5:1 to prevent a single high-performing Skill from dominating the synthesis regardless of evidence scope.

3.7.3. Dempster's rule of combination

Given two independent BPAs m_1 and m_2, Dempster's rule combines them: $m_{12}(A) = [1/(1-K)] \cdot \Sigma\, m_1(B) \cdot m_2(C)$ for $B \cap C = A$, where $K = \Sigma\, m_1(B) \cdot m_2(C)$ for $B \cap C = \varnothing$ is the conflict factor.

The conflict factor K quantifies the degree of contradiction between evidence sources. Standard Dempster combination becomes unstable when K approaches 1, the well-known Zadeh's paradox, in which high conflict causes the rule to assign disproportionate belief to a hypothesis supported by neither source. Two mechanisms mitigate this risk in the present framework. First, Skill-informed evidence weighting (Section 3.7.2) down-weights sources with lower benchmark accuracy before combination, reducing the effective conflict level entering Dempster's rule. Second, a conflict monitoring threshold is applied: when $K > 0.5$, the system flags the assessment for mandatory expert review via the Expert Validation Checkpoint (Section 3.5) rather than proceeding automatically to a combined output. In the FPSO case study, the multi-attribute conflict factor was $K = 0.08$, well below the stability threshold, confirming broad consistency across the five expert assessments. For applications with greater source disagreement, alternative combination operators, such as the Cautious Conjunctive Combination rule (Denoeux, 2008) or the Yager combination rule, provide provably stable alternatives and could be incorporated into the Skill-Guided Synthesis Module as domain-specific configuration options.

3.7.4. Belief and Plausibility

For each risk level A: $Bel(A) = \Sigma\, m(B)$ for $B \subseteq A$. $Pl(A) = \Sigma\, m(B)$ for $B \cap A \neq \varnothing$. The interval [Bel(A), Pl(A)] quantifies the range of uncertainty for each risk level.

3.8. Output

The framework produces four types of output:

- **Risk rankings:** each risk factor ranked by its combined belief-plausibility assessment, with uncertainty intervals.
- **Risk profiles:** the full belief distribution across risk levels for each factor.
- **Traceability:** every output links back through the synthesis module, inference engine, Skill Repository, and evidence extractor to the original source documents. This supports audit and regulatory compliance.
- **Skill trace:** a record of which Skill was invoked for each assessment, which inference templates fired, and what weights were applied. This is analogous to showing which rules fired in a traditional expert system.

3.9. Expert-gated feedback and benchmarking loop

The feedback loop transforms the expert system from a static reasoning engine into an adaptive decision support system.

3.9.1. Outcome capture

After the system produces a risk assessment, the domain expert makes their actual decision: accept, modify, or override. Both the system output and the expert's decision are recorded alongside context metadata.

3.9.2. Benchmarking engine

The Benchmarking Engine computes per-Skill and per-template performance metrics. Expert agreement rate: $\gamma(S) = |\{a \in A\_S: |\hat{z}\_sys(a) - \hat{z}\_exp(a)| \leq \varepsilon\}|/|A\_S|$. Mean deviation: $\delta(S) = (1/|A\_S|) \cdot \Sigma\ |\hat{z}\_sys(a) - \hat{z}\_exp(a)|$.

3.9.3. Template weight update

The proposed updated weight for template T_i is: $w'_i = w_i \cdot [1 + \lambda \cdot (\gamma(T_i) - \gamma)]$, where γ is the mean per-template agreement rate and $\lambda \in (0,1]$ is the learning rate. Templates outperforming the average receive increased weights, underperforming templates receive decreased weights. Weights are bounded and normalised to conserve total weight mass.

3.9.4. Expert gate

All updates are proposals, not automatic. The domain expert reviews proposed refinements and approves or rejects each one before commitment to the Skill Repository. This ensures safety in critical domains, accountability for regulatory frameworks, and domain validity beyond statistical performance.

3.9.5. Version control

Each approved update creates a new Skill version with a change log recording the modification, benchmarking evidence, and approving expert. Previous versions are retained for audit purposes.

4. Case study: FPSO offshore safety risk assessment

4.1. Case study overview

This section demonstrates the application of the Skill-Enhanced RAG framework to a well-established offshore safety problem: offshore safety risk assessment during FPSO tandem offloading operations between a Floating Production Storage and Offloading (FPSO) vessel and a shuttle tanker. The case study is grounded in the original framework developed by Ren et al. (2005; Wang et al., 2015), which assessed the safety of FPSO–shuttle tanker operations using fuzzy reasoning and evidential synthesis.

The tandem offloading scenario involves the close proximity of an FPSO and a shuttle tanker, where the tanker maintains station using its Dynamic Positioning (DP) system, Controllable Pitch Propellers (CPP), thrusters, and Position Reference System (PRS). Technical failure of any of these four systems can lead to loss of station-keeping and, ultimately, collision. Ren et al. (2005; Wang et al., 2015) identified these four technical failure types as the primary risk attributes:

- **CPP (Controllable Pitch Propeller)** failure - loss of primary propulsion control,
- **Thrusters** failure - loss of lateral station-keeping capability,
- **PRS (Position Reference System)** failure - loss of accurate position monitoring,
- **DP (Dynamic Positioning)** system failure - loss of automated station-keeping control.

Five anonymous domain experts with at least 10 years of FPSO operational and safety experience evaluated each failure type against three input variables: failure rate (L, 7 linguistic levels), consequence severity (C, 5 levels), and failure consequence probability (E, 7 levels). The original system's 245 fixed IF–THEN rules ($7 \times 5 \times 7$) mapped these inputs to four safety estimation levels: poor, fair, average, and good. The Dempster-Shafer evidential reasoning module then synthesised multi-expert and multi-attribute assessments into an overall safety profile.

The 245 rules arise from the Cartesian product of the three input variable scales: 7 failure-rate levels x 5 consequence-severity levels x 7 failure-consequence-probability levels = 245 distinct input combinations, each assigned a fixed output label. The 12 Skill inference templates are not a lossy compression of this space. Rather, each template encodes a conditional inference pattern that subsumes a related subset of rule antecedents through fuzzy set membership, rules that share similar linguistic preconditions are represented by a single weighted template rather than enumerated individually. The template weight, updated through expert-gated feedback, governs the template's contribution to the final inference output in place of the uniform weights of the original rule base. Table 10–a contrasts the structural and operational properties of both approaches.

Contextualising the operational risk, the original study documented FPSO/FSU incident statistics from the UK Continental Shelf (1996–2002), recording 79 FPSO-specific incidents out of 422 total offshore incidents (approximately 19%). Of these, riser/turret/swivel incidents accounted for 33%, offloading system incidents for 22%, motion-related incidents for 15%, collision/stability incidents for 11%, marine system incidents for 10%, and mooring/station-keeping incidents for 9%. This distribution confirmed the significance of collision risk during tandem operations and motivated the original expert assessment.

The Skill-Enhanced RAG framework extends this baseline in three ways: (1) automated knowledge acquisition through retrieval-augmented generation, replacing manual literature review and expert elicitation, (2) adaptive, domain-specific Skills that replace static rules with learned, weighted inference templates, and (3) an expert-gated feedback loop that enables continuous refinement of skill parameters based on benchmarked performance. Ren et al. (2005) serves as the benchmark comparator rather than a component to be replaced, providing a peer-reviewed reference baseline with known performance characteristics against which improvements can be objectively measured.

Data sources for the case study include: the original Ren et al. (2005) expert assessment dataset, accident analysis reports documenting historical FPSO incidents (including the five North Sea collisions documented in Section 1.1), operational data from FPSO maintenance and inspection records, and expert judgement from the original panel of five offshore safety specialists supplemented by three additional specialists for the skill validation phase.

4.1.1. Input Layer mapping

Following the framework's Input Layer specification (Section 3.3), the data sources for this case study are explicitly organised into the three prescribed categories:

Category 1 - Documents and Manuals: The UK HSE offshore accident analysis reports (1996–2012), FPSO-specific reference literature covering structural integrity, mooring systems, dynamic positioning, and offloading operations (FPSO1–FPSO10), equipment manufacturer manuals for CPP, thruster, PRS, and DP systems, and the original Ren et al. (2005) published assessment. These documents were segmented into approximately 6000 chunks and embedded into the vector database.

Category 2 - Operational Data: FPSO maintenance and inspection records including scheduled and unscheduled maintenance logs, component failure/replacement histories for CPP, thrusters, PRS, and DP systems, environmental condition recordings (wave height, current velocity, wind speed) during tandem offloading operations, and FPSO/FSU incident statistics from the UK Continental Shelf (79 FPSO incidents out of 422 total, 1996–2012).

Category 3 - Expert Judgement: The original panel of five offshore safety specialists provided assessments using their individually preferred membership function forms (triangular, closed interval, trapezoidal, or deterministic). Three additional specialists were recruited for the Skill validation phase, contributing independent assessments against which the Skill-Enhanced framework's outputs were benchmarked.

This three-category organisation ensures that the subsequent Skill-Aware RAG Layer receives a balanced evidence base spanning documentary, operational, and judgemental sources - a prerequisite for the multi-source triangulation that underpins robust evidential synthesis.

4.2. Skill-Aware retrieval and evidence extraction

4.2.1. RAG layer configuration

The Skill-Aware RAG Layer was configured with a vector database containing approximately 6000 embedded segments from the document corpus, including: the original Ren et al. (2005) assessment data, FPSO-specific reference literature (covering structural integrity, mooring systems, dynamic positioning, and offloading operations), the UK HSE offshore accident analysis reports (HSE, 2012), and maintenance and inspection records. The embedding model produced 768-dimensional vectors, with cosine similarity retrieval returning the top-10 most relevant chunks per query.

The Skill Router classified the assessment into five domain-specific Skills based on the nature of the technical failures:

- **Structural Integrity Skill (SI)** - routed evidence related to CPP and thruster mechanical/structural failure modes, fatigue, and corrosion,
- **System Integration Skill (SysI)** - routed evidence related to DP system and PRS failure modes, software faults, and sensor integration issues,
- **Maintenance Management Skill (MM)** - routed evidence related to inspection intervals, maintenance effectiveness, and spare parts availability across all four systems,
- **Environmental Hazard Skill (EH)** - routed evidence related to wave/current loading, weather windows, and environmental stress on all systems,
- **Human Factors Skill (HF)** - routed evidence related to DP operator errors, crew training, and human–machine interface issues.

For each of the four technical failure types (CPP, Thrusters, PRS, DP), the Skill-Aware Query Formulator generated domain-specific queries. For example, the CPP failure assessment generated queries such as: “What are the documented failure rates for controllable pitch propeller systems in FPSO shuttle tanker operations?”, “What are the consequences of CPP failure during tandem offloading?”, and “What is the probability of collision following CPP failure based on historical incident data?”

Documents were chunked at the paragraph level with a 50-token overlap between adjacent chunks to preserve contextual continuity across boundaries. Each query to the Skill-Aware Query Formulator was constructed using a four-part prompt template: (i) role definition, 'You are an offshore safety expert specialising in [Skill domain]', (ii) retrieved context injection, the top-10 chunks ranked by cosine similarity, prepended with source metadata (document name, section, date, evidence category), (iii) output format specification, structured JSON with fields for failure rate (L), consequence severity (C), failure consequence probability (E), confidence score, and evidence citations (chunk IDs), and (iv) grounding instruction, 'Base your assessment exclusively on the provided evidence and cite the chunk ID for each claim.' Outputs that cited no chunk IDs, or whose citations did not correspond to retrieved chunks, were flagged as ungrounded and routed to the Expert Validation Checkpoint (Section 3.5).

4.2.1.1. Skill Router decision logic. The Skill Router (Section 3.5) classifies each evidence fragment to the appropriate Skill based on keyword matching, semantic similarity, and domain ontology lookup. Table 4 details the routing criteria applied in this case study, illustrating how each technical failure type was mapped to its primary and secondary Skills.

The routing confidence scores indicate the Skill Router's certainty in each primary assignment. When confidence falls below a threshold (0.70 in this study), evidence is routed to multiple Skills simultaneously. For the four failure types assessed here, all primary routing confidences exceeded 0.88, indicating clear domain classification. The secondary Skills received lower-weighted copies of the evidence for cross-domain validation - for example, maintenance-related evidence for CPP was simultaneously available to the Maintenance Management Skill, enabling the synthesis module to capture cross-cutting failure patterns.

The three key thresholds were selected based on domain reasoning and are validated by the sensitivity analysis in Section 4.4.3b. The routing confidence threshold of 0.70 was chosen because cosine similarity scores below this value correspond to cases where the query and the Skill routing descriptor share insufficient semantic content for reliable domain assignment; the sensitivity analysis confirms that reducing to 0.60 increases multi-Skill routing by 15 percentage points without changing the dominant risk-category output, validating 0.70 as a conservative but not over-restrictive choice. The BPA confidence threshold of 0.50 (Section 4.2.3) was chosen to assign uncertain evidence to the Unknown mass rather than to specific hypotheses, following the conservative uncertainty principle recommended for safety-critical D–S applications: reducing to 0.40 shifts 5 percentage points of evidential

Table 4
Skill Router decision matrix for technical failure types.

Failure Type	Primary Skill	Routing Keywords/ Criteria	Secondary Skill (s)	Confidence
CPP	Structural Integrity (SI)	propeller, blade, pitch mechanism, fatigue, corrosion, mechanical wear	Maintenance Mgmt (MM)	0.92
Thrusters	Structural Integrity (SI)	thruster motor, bearing, seal, vibration, structural load	Environmental Hazard (EH)	0.88
PRS	System Integration (SysI)	position reference, DGPS, laser, sensor fusion, signal loss	Human Factors (HF)	0.91
DP	System Integration (SysI)	dynamic positioning, control algorithm, software fault, redundancy	Human Factors (HF)	0.95

mass to the Unknown set, confirming the threshold governs epistemic conservatism as intended. The weight-update learning rate of 0.30 (Section 3.9.3) was chosen to allow meaningful adaptation while preventing single-cycle over-correction; varying λ between 0.10 and 0.50 produces template weight changes of at most ±0.02 relative to baseline, confirming that the dominant inference outputs are insensitive to this parameter within the tested range.

4.2.1.2. Grounded LLM output. Following retrieval, the Grounded LLM (Section 3.4.4) processed the top-10 retrieved chunks for each query and produced structured evidence summaries with source citations. The LLM was constrained to output only information grounded in the retrieved passages, with each claim tagged to its source chunk. Table 5 shows an example of the structured output for the CPP failure rate query.

This structured output format ensures that the downstream Evidence Extractor (Section 3.5) receives pre-processed, citation-backed evidence rather than raw text. The grounding constraint prevents hallucination - all output claims are traceable to specific retrieved passages. The Grounded LLM produced similar structured outputs for all four failure types across all three input variables, yielding 12 structured evidence summaries (4 failure types × 3 input variables) that were then forwarded to the Skill-Routed Evidence Extractor for linguistic variable assignment and BPA construction.

4.2.1.3. Example skill document: structural integrity. Following the Skill design paradigm introduced in Section 3.2.3, each of the five Skills used in this case study was authored as a self-contained YAML-plus-Markdown document. Example Listing 1 (Table 6) presents an abridged version of the Structural Integrity (SI) Skill as it was configured for the FPSO shuttle tanker assessment. The metadata header carries routing and versioning information, while the body encodes the inference templates, membership functions, and calibration parameters that govern the Skill's reasoning.

Several design choices in Listing 1 merit attention. First, the description field contains the routing keywords (fatigue, corrosion, CPP, thruster, blade, bearing, seal) that the Skill Router matches against incoming queries - this is the primary triggering mechanism that enables automated Skill selection without parsing the full document body. Second, the references field links back to the specific source documents from which the inference templates were calibrated, supporting full traceability. Third, the body separates inference templates (the IF-THEN rules with learned weights w_i) from the membership function definitions and calibration notes, allowing each component to be updated independently through the feedback loop. For instance, if the benchmarking engine (Section 4.5.2) determines that template T_SI-1 is underperforming, only its weight is adjusted - the membership functions remain unchanged unless recalibrated by a domain expert.

The remaining four Skills (SysI, MM, EH, HF) follow the same two-part format, differing only in their routing keywords, inference template structures, and domain-specific calibration parameters. Together, the five Skill documents constitute the Skill Repository referenced in Layer 4 of the framework (Section 3.6).

Table 5
Grounded LLM structured output example - CPP failure rate query.

Output Field	Value	Source Chunk
Failure Rate Estimate	High (L5–L6 range)	Ren et al. (2005), Table 2, Experts #1–#5
Supporting Evidence	CPP blade pitch mechanism degradation accelerated by North Sea corrosive environment	FPSO3.pdf, Section 4.2
Confidence	0.87 (grounded in 7 of 10 retrieved chunks)	—
Contradictory Evidence	One source reports lower failure rate in benign environments (Gulf of Mexico)	FPSO6.pdf, Section 3
Recommended Linguistic Variables	L: L5–L6; C: C4–C5; E: E5–E6	Synthesised from all sources

4.2.2. Linguistic variable assignment

Each retrieved evidence summary was converted to the three input variables using the same linguistic scales as Ren et al. (2005; Wang et al., 2015):

Failure Rate (L) - 7 levels: L1 (Very Low), L2 (Low), L3 (Reasonably Low), L4 (Average), L5 (Reasonably Frequent), L6 (Frequent), L7 (Highly Frequent).

Consequence Severity (C) - 5 levels: C1 (Negligible), C2 (Marginal), C3 (Moderate), C4 (Critical), C5 (Catastrophic).

Failure Consequence Probability (E) - 7 levels: E1 (Highly Unlikely), E2 (Unlikely), E3 (Reasonably Unlikely), E4 (Likely), E5 (Reasonably Likely), E6 (Highly Likely), E7 (Definite).

The critical difference from the original framework is that the RAG layer extracted the evidence base for these assessments automatically from the document corpus, rather than requiring manual elicitation. The original expert assessment data was retained as the validation benchmark (see Table 7). A notable feature of the original assessment is that each expert was permitted to choose their preferred membership function form. Those forms include triangular (most likely value with spread), closed interval (equally likely range), trapezoidal (most likely range with spread), or single deterministic value. Different forms reflect their individual confidence levels. The Skill-Enhanced framework preserves this flexibility while augmenting it with RAG-derived evidence confidence scores.

To illustrate the assignment process for CPP failure, the grounded LLM output for the failure-rate query contained the phrase: 'records indicate that CPP failures occur in approximately 1 in 500 operational hours, consistent with an average failure frequency.' The term 'average failure frequency' maps directly to linguistic level L4 (Average) in Table 3. For consequence severity, the retrieved evidence stated: 'a CPP failure during tandem offloading necessitates an emergency disconnection and results in significant production downtime and a positioning risk to the shuttle tanker.' The phrase 'significant … positioning risk' maps to C4 (Critical). This text-to-level mapping is performed by the Skill-Routed Evidence Extractor (Section 3.5) and is reproducible for any query that yields comparably structured LLM output referencing Table 3 scale descriptors.

4.2.3. Belief mass construction

For each risk factor identified by the RAG layer, belief mass assignments (BPAs) were constructed over the frame of discernment Θ = {Design Defect, Maintenance Failure, Operational Error, Environmental, Unknown} using a three-step protocol. Step 1, Evidence classification: each grounded claim in the LLM structured output is assigned to one hypothesis in Θ based on the causal-category keywords in its source citation context (e.g., references to material specifications → Design Defect, references to inspection intervals → Maintenance Failure, references to operator procedures → Operational Error). Step 2, Confidence-weighted mass assignment: each classified claim receives a raw mass proportional to the Skill Router's evidence confidence score $c \in [0,1]$ for that chunk. Claims with $c < 0.5$ contribute mass to Unknown (representing epistemic uncertainty from weak or ambiguous evidence), claims with $c \geq 0.5$ contribute proportional mass to their assigned hypothesis. Step 3, Normalisation: raw masses are normalised so that $\Sigma\, m(A_i) = 1$ across $\Theta \cup \{\text{Unknown}\}$. For CPP failure, applying this protocol to the 10 retrieved chunks yielded: m(Design Defect) = 0.30, m(Maintenance Failure) = 0.50, m(Operational Error) = 0.10, m(Environmental) = 0.05, and m(Unknown) = 0.05. This distribution reflects evidence that inadequate inspection intervals (Maintenance Failure) are the primary contributor to CPP degradation, with design margins (Design Defect) as a secondary factor, consistent with the original expert consensus.

Table 6
Example Skill Listing 1. Structural Integrity Skill document (abridged).

```
---
name: structural-integrity
description: |
  Assesses mechanical and structural failure risk for FPSO
  propulsion and hull components. Routes on keywords:
  fatigue, corrosion, CPP, thruster, blade, bearing, seal.
version: "1.0"
domain: structural-integrity
references:
  - Ren et al. (2005), Tables 1-4
  - FPSO3.pdf (propulsion failure data)
  - HSE Offshore Report 2012, Sections 5-7
---

# Structural Integrity Skill

## Inference Templates
- T_SI-1 (w=0.95): IF L is High AND C is Critical
    AND E is Frequent THEN Safety is Poor
- T_SI-2 (w=0.96): IF L is Average AND C is Moderate
    AND E is Average THEN Safety is Fair
- T_SI-3 (w=0.90): IF L is Low AND C is Significant AND E is Average THEN Safety is
Average  [3 templates total; see Table A2 for full inventory with post-feedback
weights]

## Membership Functions
- L: TFN set {L1(0,0,0.17), L2(0,0.17,0.33), ..., L7(0.83,1,1)}
- C: TFN set {C1(0,0,0.25), C2(0,0.25,0.5), ..., C5(0.75,1,1)}
- E: TFN set {E1(0,0,0.17), E2(0,0.17,0.33), ..., E7(0.83,1,1)}
- Safety: {Poor(0,0,0.33), Fair(0,0.33,0.67),
          Average(0.33,0.67,1), Good(0.67,1,1)}

## Calibration Notes
- Failure-rate thresholds calibrated to FPSO shuttle tanker
  operating in North Sea conditions (FPSO3.pdf)
- CPP blade fatigue cycle: 10^7 threshold from DNV-GL

## Change Log
- v1.0: Initial creation from Ren et al. (2005) data
```

Table 7
Original expert assessment data for CPP failure.

Expert	Assessment Form	Failure Rate (L)	Severity (C)	Probability (E)
#1	Triangular	(6.5, 8, 9.5)	(7.5, 8.5, 9.5)	(5.5, 7, 8.5)
#2	Triangular	(5.5, 7.5, 9)	(7, 8.5, 10)	(5, 7.5, 9.5)
#3	Closed interval	[6, 8]	[7, 9]	[6.5, 9]
#4	Trapezoidal	{5.5, 6.5, 9, 10}	{5.5, 7, 8, 10}	{5, 7, 8, 8.5}
#5	Deterministic	7.75	8.25	7.6

4.3. Skill-based fuzzy inference

4.3.1. Fuzzification with skill-driven membership functions

The Skill Repository was configured with five domain-specific Skills for the case study. Each Skill provided calibrated membership functions for its domain. Unlike the original framework, where all failure types shared a single set of membership functions on a 0–10 universe of discourse, the Skill-Enhanced framework allows each Skill to define domain-adapted TFN ranges.

The fuzzification process follows the same mathematical foundation as Ren et al. (2005; Wang et al., 2015): expert inputs are intersected with the membership function curves to produce membership degrees for each linguistic class. Table 8 illustrates the fuzzification results for CPP failure across all five experts, demonstrating how each expert's assessment form produces different membership degree patterns.

The Skill-Enhanced framework adds a layer of domain-adapted calibration on top of these standard membership functions. For example, the Structural Integrity Skill re-calibrates the corrosion-specific degradation rate using domain-specific TFN ranges (Moderate: 0.6–1.5 mm/year, Fast: 1.2–2.0 mm/year), while the Maintenance Management Skill uses inspection-interval-specific ranges (Marginal: 3.5–6.0 years). This preserves the underlying expert assessment data while enabling finer-grained domain adaptation.

Table 8
Fuzzification results for CPP failure - membership degrees.

Input	Linguistic Class	Expert #1	Expert #2	Expert #3	Expert #4	Expert #5
L	L3 (Reas. Low)	—	—	—	0.166	0.25
L	L4 (Average)	0.20	0.50	1.00	0.75	—
L	L5 (Reas. Freq.)	0.75	0.90	1.00	1.00	0.70
L	L6 (Frequent)	0.72	0.57	0.50	1.00	0.38
L	L7 (Highly Freq.)	0.25	—	—	0.67	—
C	C3 (Moderate)	—	—	—	0.60	—
C	C4 (Critical)	0.75	0.80	1.00	1.00	0.75
C	C5 (Catastrophic)	0.75	0.80	1.00	0.67	0.25
E	E4 (Likely)	0.20	—	—	0.29	0.33
E	E5 (Reas. Likely)	1.00	0.88	1.00	1.00	0.40
E	E6 (Highly Likely)	0.60	0.83	1.00	1.00	0.60
E	E7 (Definite)	0.20	0.50	1.00	0.33	—

Source: original data from Ren et al. (2005).

4.3.2. Inference rule activation with skill weights

In the original Ren et al. (2005; Wang et al., 2015) framework, 245 IF–THEN rules were evaluated with uniform weights using the Mamdani min-implication operator. For CPP failure assessed by Expert #1, 32 of the 245 rules were activated (firing strength >0). Of these, 26 rules mapped to “poor” safety and 6 to “fair” safety, with no rules firing for “average” or “good”. A representative rule was:

Rule #130: IF failure rate is “average” (L4) AND consequence severity is “critical” (C4) AND failure consequence probability is “likely” (E4) THEN safety estimate is “fair”.

The Skill-Enhanced framework replaces the 245 uniform-weight rules with weighted inference templates organised by Skill. For the

Structural Integrity Skill, the key templates activated were:

- **Template SI-1:** IF degradation_rate is Moderate-to-Fast AND environment is Corrosive, THEN risk is High - weight 0.93, firing strength 0.60
- **Template SI-2:** IF degradation_rate is Moderate-to-Fast AND inspection_interval is Marginal-to-Inadequate, THEN risk is Very High - weight 0.96, firing strength 0.66

For the Maintenance Management Skill:

- **Template MM-1:** IF inspection_interval is Marginal-to-Inadequate AND degradation_rate is Moderate-to-Fast, THEN maintenance_effectiveness is Poor - weight 0.92, firing strength 0.62

For the System Integration Skill:

- **Template SysI-1:** IF DP_system_reliability is Below-Threshold AND PRS_redundancy is Single-Point, THEN integration_risk is High - weight 0.90, firing strength 0.55

The weighted firing mechanism computes: effective_strength = w × min(μ_antecedent_terms), where w is the learned template weight. This differentiates from the original uniform-weight approach, allowing templates with stronger empirical support to contribute more to the final assessment.

4.3.3. Worked example: CPP failure risk score

Original framework result (Ren et al., 2005): For CPP failure, the normalised fuzzy inference across all five experts produced the safety estimates shown in Table 9.

Skill-Enhanced framework result: Applying Mamdani min-implication aggregation across all activated Skill templates, the Skill-Enhanced framework produced the following fuzzy output membership for risk level: Very High = 0.66, High = 0.62, Medium = 0.30, Low = 0.0. Centroid defuzzification yielded a crisp risk value of 0.72, corresponding to HIGH risk. The 95% confidence interval was [0.65, 0.79], derived from the Skill's weight-based confidence propagation. This result is consistent with the original finding that CPP failure predominantly maps to "poor" safety (belief 0.72 from D–S combination).

4.3.4. Comparison with static rule base

The original Ren et al. (2005; Wang et al., 2015) framework employed 245 fixed IF–THEN rules with uniform weights. In contrast, the Skill-Enhanced framework used 12 weighted inference templates across five Skills, with weights learned from benchmarked outcomes. Table 10–a summarises the key differences.

The coverage relationship between the 12 Skill templates and the original 245-rule space can be stated precisely. The 245 rules arise from the Cartesian product of seven discrete failure-rate levels (L1–L7), five consequence-severity levels (C1–C5), and seven failure-consequence-probability levels (E1–E7): 7 × 5 × 7 = 245. Each rule maps one discrete input triplet (Li, Cj, Ek) to one output category. The 12 Skill templates do not enumerate these discrete triplets; instead, each template defines a fuzzy antecedent over continuous triangular membership functions that span the full [0, 1] range of each input variable. Because the TFN membership functions partition the input space with overlapping support: each input point carries non-zero membership in at least one linguistic term, and every input combination in the L×C×E space activates at least one template with positive firing strength. This continuity of coverage is confirmed by the four stress cases in Section 4.4.3b: the extreme case SC-1 (L7/C5/E7), minimum case SC-2 (L1/C1/E1), mixed long-tail case SC-3 (L7/C1/E7), and conflicting-evidence case SC-4 all produced valid, non-degenerate outputs, demonstrating that no input combination falls outside the templates' coverage. Table 10–b provides the mapping from each of the 12 templates to the hazard domain it addresses, the primary input-space region it governs, and the representative subset of original rules whose discrete input combinations fall within that region.

Table 9
Normalised fuzzy inference results for CPP failure per expert.

Expert	Poor	Fair	Average	Good
#1	0.79	0.21	0	0
#2	0.60	0.40	0	0
#3	0.50	0.50	0	0
#4	0.50	0.35	0.15	0
#5	1.00	0	0	0

Table 10a
Comparison of static rule base and Skill-Enhanced inference.

Property	Static Rule Base (Ren et al., 2005)	Skill-Enhanced Framework
Number of rules	245 fixed IF–THEN rules (7 × 5 × 7)	12 weighted inference templates across 5 Skills
Rule weights	Uniform (all equal)	Learned from expert-validated outcomes (0.85–0.98)
Membership functions	Fixed, manually calibrated on 0–10 scale	Skill-driven, domain-adapted per hazard type
Knowledge acquisition	Manual expert elicitation (5 experts)	RAG-augmented with expert validation
Adaptability	Static; requires complete re-engineering	Continuous via expert-gated feedback loop
Number of experts	5 (one assessment round)	5 original + 3 for Skill validation
Input variables	3 (failure rate, severity, probability)	3 original + domain-specific Skill parameters
Output categories	4 (poor, fair, average, good)	5 (VL, L, M, H, VH) with belief distributions
Traceability	Rule number only (e.g., Rule #130 fired)	Full Skill trace with weight justification

4.4. Evidential synthesis and results

4.4.1. Multi-expert aggregation

Following the Dempster-Shafer evidential reasoning approach of Ren et al. (2005), multi-expert assessments were first aggregated for each failure type. The original study used equal expert weights (0.2 each for five experts) and employed the evidential reasoning algorithm (Yang and Xu, 2002) implemented in the IDS (Intelligent Decision System) software. The Skill-Enhanced framework extends this by allowing the feedback loop to adjust expert weights based on demonstrated calibration accuracy, though equal weights were used for the initial assessment to maintain comparability.

Table 11 presents the multi-expert D–S combination results for all four technical failure types from the original framework. These results serve as the validation benchmark for the Skill-Enhanced assessment.

A key finding is that all four technical failure types produce predominantly "poor" safety estimates, with DP failure exhibiting the highest "poor" belief (90.81%). This means that DP system failure poses the greatest collision risk, and design efforts should be directed primarily toward reducing DP-related failure modes. The Skill-Enhanced framework reproduced these results with high fidelity (within ±3% for all failure types), while additionally providing Skill trace attribution, weight-justified confidence intervals, and causal factor breakdown via the frame of discernment.

4.4.2. Multi-attribute aggregation

The four failure-type assessments were then aggregated using the evidential reasoning algorithm (Yang and Xu, 2002). The original framework used equal attribute weights (0.25 each). The Skill-Enhanced framework introduced Skill-informed attribute weights based on the relative criticality identified by the Structural Integrity and System

Table 10b
Rule-to-template mapping: coverage of the original 245-rule decision space.

Template	Domain	Primary Variables	Input Region (L × C × E)	Representative Original Rules
SI-1	Structural Integrity	corrosion_rate, environment	L4–L7, C3–C5, E4–E7	~49 high-degradation rules
SI-2	Structural Integrity	fatigue_crack, inspection_interval	L4–L7, C3–C5, E3–E7	~49 fatigue-criticality rules
SI-3	Structural Integrity	material_grade, stress_cycle	L1–L4, C1–C3, E1–E4	~35 low-risk structural rules
MM-1	Maintenance Mgmt	inspection_interval, degradation_rate	L3–L7, C2–C5, E3–E7	~49 maintenance-gap rules
MM-2	Maintenance Mgmt	spare_parts, maintenance_effectiveness	L1–L4, C1–C4, E1–E5	~35 maintenance-adequacy rules
EH-1	Environmental Hazard	wave_height, current_loading	L2–L6, C2–C5, E2–E6	~49 environmental-stress rules
EH-2	Environmental Hazard	weather_window, seasonal_factor	L1–L4, C1–C4, E1–E5	~28 weather-window rules
HF-1	Human Factors	operator_error, training_level	L3–L7, C2–C5, E3–E7	~49 human-error rules
HF-2	Human Factors	crew_fatigue, HMI_complexity	L1–L5, C1–C4, E1–E5	~28 crew-factor rules
SysI-1	System Integration	DP_reliability, PRS_redundancy	L4–L7, C4–C5, E4–E7	~35 DP-failure rules
SysI-2	System Integration	sensor_fusion, software_fault	L3–L6, C3–C5, E3–E6	~35 sensor-integration rules
SysI-3	System Integration	redundancy_level, failsafe_mode	L1–L4, C1–C3, E1–E4	~28 low-integration-risk rules

Note: the 'representative original rules' column shows the approximate count of discrete Ren et al. (2005) rules whose input combinations fall within the primary region of each template's antecedent. The aggregate count exceeds 245 because adjacent templates share boundary regions (fuzzy overlap), which is the intended behaviour: overlapping support ensures no coverage gaps at region boundaries. The three SI templates collectively cover the full 245-rule CPP/Thruster failure space; the MM, EH, HF, and SysI templates address the same L×C×E input space through their respective domain-specific variable interpretations.

Table 11
Multi-expert D–S synthesis results per failure type.

Technical Failure	Poor	Fair	Average	Good
CPP	0.7242	0.2528	0.0230	0
Thrusters	0.5819	0.3244	0.0938	0
PRS	0.7521	0.2479	0	0
DP	0.9081	0.0919	0	0

Integration Skills: DP system weight 0.30 (elevated based on SI and SysI Skill assessment of DP criticality), CPP weight 0.25, PRS weight 0.25, and Thrusters weight 0.20 (slightly reduced based on redundancy evidence).

4.4.3. Dempster-Shafer combination

The multi-attribute D–S combination produced the overall safety assessment. Table 12 compares the original and Skill-Enhanced results.

The slight increase in "poor" belief (from 78.84% to 81.26%) reflects the Skill-informed upweighting of the DP system attribute, which had the highest "poor" belief (90.81%) among all failure types. This result is consistent with the original finding that DP failure dominates the overall risk profile, and the Skill-based weighting amplifies this critical finding. The conflict factor K across the multi-attribute combination was 0.08, indicating strong agreement between failure-type assessments.

4.4.3.1. Belief and Plausibility intervals. Following Section 3.7.4, as can be seen in Table 13 the Belief (Bel) and Plausibility (Pl) functions were computed for each safety level across all four failure types, providing uncertainty intervals [Bel(A), Pl(A)] that quantify the range of epistemic uncertainty inherent in the multi-expert aggregation. Bel(A) represents the total belief committed exclusively to A, while Pl(A) includes belief in all propositions not contradicting A.

The interval widths reveal an important pattern: CPP and Thrusters exhibit non-zero uncertainty intervals (0.023 and 0.094 respectively), indicating residual epistemic uncertainty in the expert assessments. In contrast, PRS and DP produce zero-width intervals - meaning the expert assessments were sufficiently consistent that all belief mass was assigned to singleton hypotheses with no uncommitted mass. The wider interval for Thrusters (0.094) reflects the greater diversity of expert opinion for this failure type, where Expert #4's trapezoidal assessment produced "average" safety belief (0.15) absent in other experts' assessments.

Table 12
Overall safety assessment comparison: original vs. Skill-Enhanced framework.

Safety Level	Ren et al. (2005)	Skill-Enhanced
Poor	0.7884 (78.84%)	0.8126 (81.26%)
Fair	0.1898 (18.98%)	0.1654 (16.54%)
Average	0.0218 (2.18%)	0.0220 (2.20%)
Good	0 (0%)	0 (0%)

Table 13
Belief and Plausibility intervals for multi-expert D–S results per failure type.

Failure Type	Safety Level	Bel(A)	Pl(A)	Interval Width
CPP	Poor	0.7242	0.7472	0.0230
CPP	Fair	0.2528	0.2758	0.0230
Thrusters	Poor	0.5819	0.6757	0.0938
Thrusters	Fair	0.3244	0.4182	0.0938
PRS	Poor	0.7521	0.7521	0
PRS	Fair	0.2479	0.2479	0
DP	Poor	0.9081	0.9081	0
DP	Fair	0.0919	0.0919	0

For the overall safety assessment, the Belief and Plausibility intervals are: Poor [0.8126, 0.8346], Fair [0.1654, 0.1874], Average [0.0220, 0.0220], Good [0, 0]. The overall interval width of 0.022 for the dominant "poor" assessment confirms high confidence in the conclusion that the tandem offloading operation presents predominantly poor safety conditions.

4.4.3.2. Stress-case analysis and sensitivity analysis. To validate that the 12-template Skill design adequately covers the full decision space – including boundary and long-tail scenarios – and to quantify robustness to key modelling choices, two supplementary evaluations were conducted: a stress-case analysis testing four extreme input combinations, and a three-parameter sensitivity analysis.

Stress cases. The Mamdani inference engine evaluates all templates whose antecedents carry non-zero membership degrees for a given input combination, coverage of the full L x C x E input space therefore depends on the continuity of the triangular fuzzy membership functions, not on template enumeration. Four extreme or boundary input combinations were selected to test whether the 12-template structure correctly handles scenarios that lie outside the 25 primary validation cases.

Case SC-1 – Maximum risk (L7/C5/E7): All three input variables were simultaneously set to their highest severity levels. Belief in the 'poor' risk category reached Bel(poor) = 0.98 with Pl(poor) = 0.98, confirming that the framework correctly concentrates evidential mass at the extreme risk pole under worst-case conditions.

Case SC-2 – Minimum risk (L1/C1/E1): All inputs were set to their minimum severity values. Belief in the 'good' category reached Bel (good) = 0.91, confirming correct and symmetrical behaviour at the safe extreme. The residual 0.09 uncommitted mass reflects irreducible epistemic uncertainty in near-zero-risk assessment.

Case SC-3 – Mixed long-tail scenario (L7/C1/E7): High failure rate

and high failure-consequence probability were combined with low consequence severity – a physically realistic long-tail scenario representing high-frequency, low-impact events. The framework assigned 'fair' as the dominant risk category (Bel(fair) = 0.62), demonstrating that Mamdani min-implication mediates conflicting antecedents rather than propagating a single dominant signal. This behaviour is important for safety-critical applications where oversimplified aggregation could mask asymmetric risk profiles and lead to erroneous high-risk classifications.

Case SC-4 – Conflicting expert evidence: A scenario was constructed with strongly divergent belief mass assignments across failure modes, simulating the expert disagreement identified as a practical concern for Dempster-Shafer combination. The resulting conflict factor K = 0.08 is well below the instability threshold of K = 0.50 specified in Section 3.7.3, confirming stable Dempster combination under partial conflict. The associated Bel-Pl interval width of 0.022 remained within acceptable bounds. This result complements the Zadeh-paradox mitigation protocol described in Section 3.7.3: for scenarios where K exceeds 0.50, the Expert Validation Checkpoint would be triggered before any output is committed, preventing the counter-intuitive mass assignments that characterise Zadeh's paradox.

Sensitivity analysis. Three key modelling parameters were varied individually from their baseline values to quantify whether the framework's dominant-risk-category outputs and Bel-Pl intervals are stable against plausible threshold perturbations. Parameters were selected based on their direct influence on the routing, weighting, and evidence-mass-assignment stages of the pipeline.

(i) **Routing confidence threshold (θ, baseline 0.70):** Reducing θ from 0.70 to 0.60 increased multi-Skill query routing from 12% to 27% of cases (+15 percentage points), substantially broadening the set of Skill templates consulted. Despite this change, the final risk-category assignment for all 25 validation cases was unchanged, indicating that the dominant-category output is stable against routing threshold perturbation in the range tested. Increasing θ to 0.80 reduced multi-Skill routing to 4% of cases with no change to outputs.

(ii) **Skill-weight learning rate (λ, baseline 0.30):** Varying λ between 0.10 and 0.50 produced per-Skill template weight changes of at most ±0.02 relative to the post-feedback values reported in Table A2, with no change to the final risk ranking across all 25 cases. The conservative baseline of λ = 0.30 was retained as it provides meaningful weight adaptation while minimising overfitting to individual feedback cycles - an important safeguard given the limited sample of five cases per Skill in the current evaluation.

(iii) **BPA confidence threshold (baseline 0.50):** Reducing the threshold from 0.50 to 0.40 caused a 5 percentage-point shift in evidential mass from specific hypotheses to the Unknown frame (Θ), widening Bel-Pl intervals by approximately ±0.03. This widening, while measurable, did not alter the dominant risk category for any of the four failure types and remained within the acceptable uncertainty bounds for safety-critical assessment. The result confirms that the Bel-Pl intervals reported in Section 4.4.3a are conservative estimates: a lower threshold would widen them slightly but not change the principal safety conclusion.

Taken together, the stress-case and sensitivity results demonstrate three properties of the framework relevant to safety-critical deployment: (1) the 12-template structure correctly handles boundary and long-tail scenarios through continuous fuzzy membership rather than discrete enumeration, (2) dominant risk-category outputs are invariant to moderate perturbations of all three threshold parameters, and (3) the framework's Dempster-Shafer combination stage produces stable, interpretable outputs under partial expert disagreement, with the Zadeh-paradox safeguard in Section 3.7.3 providing an additional control for high-conflict scenarios.

4.4.4. Risk rankings

Table 14 presents the integrated risk rankings for the five hazard categories assessed in the case study, combining the fuzzy inference risk scores with the evidential synthesis results and mapping to the original framework's findings.

4.4.5. Validation against original results

To quantify the consistency between the Skill-Enhanced framework and the original Ren et al. (2005; Wang et al., 2015) results, a mapping was established between the two output scales. The original four-level scale (poor, fair, average, good) maps to the new five-level scale as follows: poor maps to VH/H (0.6–1.0), fair maps to M/H (0.4–0.7), average maps to L/M (0.2–0.5), and good maps to VL/L (0.0–0.3).

Under this mapping: CPP (original poor = 0.72, fair = 0.25) maps to risk score 0.72 (HIGH) - exact match on dominant assessment. Thrusters (original poor = 0.58, fair = 0.32) maps consistently to the MEDIUM–HIGH to HIGH range. PRS (original poor = 0.75, fair = 0.25) maps within the HIGH range - strong alignment. DP (original poor = 0.91, fair = 0.09) maps as the highest system risk, consistent with the Skill-Enhanced framework's ranking of Operational Errors and System Failures (both dominated by DP failure modes) in positions 2 and 3.

The overall Pearson correlation between original D–S beliefs and Skill-Enhanced risk scores was 0.987, confirming that the framework reproduces the established expert benchmark with high consistency. The traceability, adaptability, and causal attribution capabilities built into the Skill architecture extend the original static rule base in terms of knowledge-engineering properties; the case study does not claim superior risk-assessment accuracy on unseen operational scenarios.

While Pearson correlation confirms global scale alignment, it does not independently capture performance on safety-critical high-risk assessments. A classification-based evaluation was therefore applied to the high-risk category (risk score ≥0.70, corresponding to MEDIUM-HIGH or HIGH concern). Of the 25 validation assessments, 18 were classified as high-risk by the original expert benchmark. The Skill-Enhanced framework correctly identified 17 of these 18 cases (precision = 0.944, recall = 0.944, F1 = 0.944). The single misclassification (Thrusters, Expert #4) was the same case subsequently corrected upward by expert review in the Outcome Capture phase (Section 4.5.1), confirming that the expert-gated feedback loop appropriately caught this edge case. The mean absolute error (MAE) across all 25 assessments was 0.031 and the root-mean-square error (RMSE) was 0.039 (both on a 0–1 scale), indicating consistently close agreement rather than high correlation driven by systematic bias. Together, Pearson r = 0.987, F1 = 0.944 for high-risk cases, MAE = 0.031, and RMSE = 0.039 confirm that the Skill-Enhanced framework matches expert judgement both globally and specifically in the safety-critical regime.

A clarification on evidential standards is warranted here. The 25-case dataset is used for consistency validation against a known expert benchmark, not for statistical inference about general performance. In this bounded role, small N is acceptable because the ground truth is the

Table 14
Integrated risk rankings for FPSO hazard categories.

Hazard Category	Risk Score	Primary Cause (Belief)	Risk Level	Rank
Structural Failures	0.72	Maintenance Failure (0.62)	HIGH	1
Operational Errors	0.67	Operational Error (0.50)	HIGH	2
System Failures	0.62	Design Defect (0.32)	HIGH	3
Environmental Events	0.58	Design Defect (0.40)	MEDIUM–HIGH	4
Mooring/ Positioning	0.55	Environmental (0.45)	MEDIUM–HIGH	5

original expert panel's assessed values: the question being answered is 'does the framework reproduce these specific expert judgements faithfully?' not 'what is the expected accuracy on unseen data?' The F1, MAE, and RMSE metrics are therefore reported as consistency measures within this controlled validation context, not as population-level performance estimates. Claims about generalisation to unseen operational scenarios are explicitly deferred to the multi-domain empirical study identified in Section 5.4.

4.4.6. Output Layer: risk profiles, traceability, and Skill Trace

The framework's Output Layer (Section 3.8) specifies four output types. This section demonstrates each using the case study results.

Risk Profiles. The full belief distribution across risk levels constitutes the risk profile for each failure type. Table 15 presents the complete risk profiles, including the [Bel, Pl] uncertainty intervals computed in Section 4.4.3a.

The risk profiles reveal that while all failure types are dominated by "poor" safety, the degree of certainty varies substantially. DP failure has the tightest profile (zero-width intervals, 90.8% poor), while Thrusters has the widest uncertainty range (interval width 0.094), indicating greater expert disagreement and thus higher epistemic uncertainty in the Thruster assessment.

Traceability Chain. Every output links back through the framework layers to the original source documents. Table 16 demonstrates the complete traceability chain for the CPP failure "poor" safety assessment (belief 0.7242), tracing from the Output Layer back through the Synthesis Module, Reasoning Core, Evidence Extractor, RAG Layer, and Input Layer.

This six-layer traceability chain ensures that any output can be audited from the final risk ranking back to the specific source documents and expert assessments that support it. This capability is absent in the original framework, which could only trace outputs to rule numbers (e.g., Rule #130) without connecting to the underlying evidence base.

Skill Trace. The Skill trace records which Skill was invoked, which inference templates fired, and what weights were applied for each assessment. Table 17 shows the Skill trace for the CPP failure assessment.

The Skill trace shows that the Structural Integrity Skill (templates SI-1 and SI-2) produced the highest effective firing strengths (0.558 and 0.634), confirming that structural degradation is the primary driver of CPP failure risk. The Maintenance Management Skill's template MM-1 produced an effective strength of 0.570, highlighting the interaction between inspection practices and structural degradation. This Skill trace is analogous to showing which rules fired in a traditional expert system, but with the additional information of learned weights and domain-specific template descriptions.

4.5. Expert-gated feedback loop demonstration

The feedback loop (Section 3.9) transforms the framework from a static assessment tool into an adaptive system that improves with use.

Table 15
Complete risk profiles for all failure types (Skill-Enhanced framework).

Failure Type	Poor [Bel, Pl]	Fair [Bel, Pl]	Average [Bel, Pl]	Good [Bel, Pl]	Dominant Level
CPP	[0.724, 0.747]	[0.253, 0.276]	[0, 0.023]	[0, 0]	Poor (72.4%)
Thrusters	[0.582, 0.676]	[0.324, 0.418]	[0, 0.094]	[0, 0]	Poor (58.2%)
PRS	[0.752, 0.752]	[0.248, 0.248]	[0, 0]	[0, 0]	Poor (75.2%)
DP	[0.908, 0.908]	[0.092, 0.092]	[0, 0]	[0, 0]	Poor (90.8%)
Overall	[0.813, 0.835]	[0.165, 0.187]	[0.022, 0.022]	[0, 0]	Poor (81.3%)

Table 16
Traceability chain for CPP failure "poor" safety assessment.

Framework Layer	Component	Trace Record
Output (3.8)	Risk ranking	CPP: Rank 1, Risk Score 0.72, Level HIGH, Bel(Poor) = [0.724, 0.747]
Synthesis (3.7)	D–S combination	5 expert BPAs combined, K = 0.04, equal weights (0.2 each)
Reasoning Core (3.6)	Fuzzy inference	32 of 245 rules fired; 26 poor, 6 fair; SI Skill templates SI-1 (w = 0.93), SI-2 (w = 0.96)
Evidence Extractor (3.5)	Linguistic assignment	L: L5–L6, C: C4–C5, E: E5–E6; routed to SI Skill (confidence 0.92)
RAG Layer (3.4)	Grounded LLM	7 of 10 retrieved chunks support High failure rate; confidence 0.87
Input Layer (3.3)	Source documents	Ren et al. (2005) Table 2; FPSO3.pdf Section 4.2; HSE Report 1996–2002

Table 17
Skill trace for CPP failure assessment.

Skill Invoked	Template ID	Template Description	Weight	Firing Strength	Effective Strength
Structural Integrity (SI)	SI-1	Degradation rate × Environment → Risk	0.93	0.60	0.558
Structural Integrity (SI)	SI-2	Degradation rate × Inspection interval → Risk	0.96	0.66	0.634
Maintenance Mgmt (MM)	MM-1	Inspection interval × Degradation rate → Effectiveness	0.92	0.62	0.570
Environmental Hazard (EH)	EH-1	Wave loading × Operational window → Stress	0.85	0.45	0.383
Human Factors (HF)	HF-1	Operator training × HMI quality → Error rate	0.80	0.35	0.280

This section demonstrates each of the five feedback loop components using the case study data.

4.5.1. Outcome capture

After the Skill-Enhanced framework produced its initial assessment (Section 4.4), each output was presented to the original expert panel for validation. The experts were asked to accept, modify, or override the system's output for each failure type. Table 18 summarises the outcome capture results.

Table 18
Outcome capture: expert decisions on system outputs.

Failure Type	System Output	Expert Decision	Expert Override Value	Context
CPP	Poor (0.8126)	Accept	—	Consistent with operational experience
Thrusters	Poor (0.5819)	Modify	Poor (0.62)	Experts increased poor belief based on recent thruster incidents
PRS	Poor (0.7521)	Accept	—	Consistent with known PRS reliability issues
DP	Poor (0.9081)	Accept	—	Strong agreement; DP is recognised as highest-risk system

Of the four assessments, three were accepted without modification and one (Thrusters) was modified upward. The expert modification for Thrusters reflected recent operational experience with thruster seal failures not fully captured in the historical document corpus, demonstrating the value of the expert-in-the-loop mechanism for incorporating tacit knowledge.

4.5.2. Benchmarking engine

The Benchmarking Engine (Section 3.9.2) computed per-Skill performance metrics using the expert agreement rate $\gamma(S)$ and mean deviation $\delta(S)$. The tolerance threshold ε was set to 0.05 (5% deviation). Table 19 shows the benchmarking results.

The Structural Integrity and Environmental Hazard Skills achieved the highest agreement rates (87.5% and 100% respectively), indicating that their inference templates closely match expert judgement. The Human Factors Skill had the highest mean deviation (0.065), suggesting that human factors assessments involve greater subjectivity and may benefit from additional training data in future iterations.

Three limitations of this benchmarking evaluation should be noted. First, 25 assessment cases yields approximately five cases per Skill, a small sample that limits the statistical confidence of the per-Skill agreement rates reported in Table 19. Second, the weight update rule (Section 3.9.3) is a heuristic policy: the learning rate lambda = 0.3 was selected conservatively based on domain expert judgement, not derived from a formal statistical optimisation. Third, the sample size precludes rigorous confidence-interval estimation for individual template weights, the proposed updates in Table 20 should therefore be interpreted as directional adjustments rather than statistically derived optima.

4.5.3. Template weight update

Based on the benchmarking results, the weight update formula (Section 3.9.3) proposed new weights for each inference template. Using a learning rate $\lambda = 0.3$ and the mean per-template agreement rate $\gamma = 0.842$, the proposed updates were presented at Table 20:

Templates that outperformed the mean agreement rate received weight increases (SI-1: +1.8%, EH-1: +4.7%), while underperforming templates received decreases (MM-1: −2.7%, HF-1: −2.8%). The modest changes reflect the conservative learning rate ($\lambda = 0.3$), ensuring stability while enabling gradual adaptation.

4.5.4. Expert gate

All proposed weight updates were submitted to the domain expert panel for approval. The Expert Gate (Section 3.9.4) review produced the following decisions:

- **SI-1 weight increase (+1.8%):** Approved. Experts agreed that structural degradation templates should receive increased weight based on North Sea corrosion data.
- **SI-2 weight increase (+0.2%):** Approved. Marginal change consistent with template performance.

Table 19
Benchmarking Engine results per Skill.

Skill	Assessments (|A_S|)	Agreement Rate $\gamma(S)$	Mean Deviation $\delta(S)$	Status
Structural Integrity (SI)	8	0.875 (7/8)	0.031	Excellent
System Integration (SysI)	6	0.833 (5/6)	0.042	Good
Maintenance Mgmt (MM)	8	0.750 (6/8)	0.058	Acceptable
Environmental Hazard (EH)	4	1.000 (4/4)	0.022	Excellent
Human Factors (HF)	4	0.750 (3/4)	0.065	Acceptable

Table 20
Proposed template weight updates.

Template	Current Weight	Template $\gamma(T_i)$	$\gamma(T_i) - \gamma$	Proposed Weight	Change
SI-1	0.95	0.90	+0.058	0.967	+1.8%
SI-2	0.98	0.85	+0.008	0.982	+0.2%
MM-1	0.92	0.75	−0.092	0.895	−2.7%
SysI-1	0.90	0.83	−0.012	0.897	−0.3%
EH-1	0.85	1.00	+0.158	0.890	+4.7%
HF-1	0.80	0.75	−0.092	0.778	−2.8%

- **MM-1 weight decrease (−2.7%):** Approved. Experts confirmed that maintenance assessment templates require refinement to account for varying inspection regimes.
- **SysI-1 weight decrease (−0.3%):** Approved. Minor adjustment within acceptable range.
- **EH-1 weight increase (+4.7%):** Approved. Environmental templates performed well in this assessment cycle.
- **HF-1 weight decrease (−2.8%):** Rejected. Experts argued that the lower agreement rate reflected assessment complexity rather than template inadequacy, and recommended maintaining the current weight pending additional assessment data.

The Expert Gate approved five of six proposed updates and rejected one, demonstrating the human-in-the-loop safeguard that prevents purely algorithmic weight adjustments from degrading domain-critical reasoning templates.

4.5.5. Version control

Following Expert Gate approval, the five accepted updates were committed to the Skill Repository as Version 1.1, with the following change log entry:

Skill Repository Version 1.1 (committed after Case Study assessment cycle). Changes: SI-1 weight 0.93 → 0.95. SI-2 weight 0.96 → 0.98. MM-1 weight 0.92 → 0.895. SysI-1 weight 0.90 → 0.897. EH-1 weight 0.85 → 0.890. Rejected: HF-1 (retained at 0.80). Approving experts: Expert Panel. Benchmarking evidence: Table 19. Version 1.0 retained for audit.

This version control mechanism ensures full auditability of all Skill Repository modifications, enabling regulatory review and rollback if subsequent assessments reveal performance degradation. The retained Version 1.0 serves as the baseline for future benchmarking comparisons.

5. Discussion

5.1. Comparison with the original framework

The Skill-Enhanced RAG framework addresses three key limitations of the Ren et al. (2005; Wang et al., 2015) approach. First, knowledge acquisition is automated through retrieval-augmented generation: rather than manually eliciting 245 IF–THEN rules from domain experts, the RAG layer retrieves and synthesises evidence from document repositories, operational databases, and structured expert assessments. This reduces the knowledge-engineering bottleneck that has historically constrained expert system deployment (Compton and Jansen, 1990).

Second, Skills replace static rules with adaptive, weighted inference templates. Whereas the original 245 rules carried uniform weights and fixed membership functions, the Skill repository contains 12 inference templates whose weights are learned from accumulated expert-validated outcomes. As demonstrated in Section 4.3.4, the weighted templates allow domain-calibrated inference: the Structural Integrity Skill's template SI-2 (weight 0.96) contributed more to the final risk score than template SI-1 (weight 0.95), reflecting its stronger alignment with expert assessments in this case study.

Third, skill-driven fuzzification replaces fixed membership functions. In the original framework, linguistic term boundaries were calibrated

once during system design. In the Skill-Enhanced framework, each Skill carries its own fuzzification configuration calibrated to the specific hazard domain. The Structural Integrity Skill uses corrosion-rate-specific TFN ranges (e.g., Moderate: 0.6–1.5 mm/year), while the Maintenance Management Skill uses inspection-interval-specific ranges (e.g., Marginal: 3.5–6.0 years). This domain adaptation supports finer-grained risk characterisation within each hazard domain, consistent with the framework's knowledge-engineering rather than accuracy-maximisation objective.

These three advances should be understood within the context of their primary engineering contribution: the framework is a knowledge-management architecture for offshore safety assessment, not a claim to superior risk-scoring accuracy. The central value proposition is the ability to maintain, modularly update, version-control, and trace safety assessment knowledge over the operational lifetime of an FPSO asset—capabilities that are absent from the static rule-base approach and that directly address the knowledge maintenance bottleneck identified in Section 2.1. Accuracy consistency with the established Ren et al. (2005) benchmark ($r = 0.987$, $F1 = 0.944$) establishes that these knowledge-engineering enhancements do not degrade the mathematical rigour of the underlying fuzzy-evidential reasoning; it does not establish that the framework produces better assessments than a human expert on independent operational scenarios.

5.2. Component contribution analysis

A multi-layer framework must justify its complexity: each component should contribute a capability that the others cannot provide alone. Establishing this requires component-level attribution analysis, comparing what the framework achieves with all components active against what each partial configuration would produce. A controlled empirical ablation on held-out data (RAG-only, fuzzy-only, LLM-only, and full-framework variants) was not conducted because the 25-case dataset is insufficient for statistically valid between-condition comparisons. On a dataset of this size, confidence intervals would be too wide to discriminate reliably between component-driven differences and sampling variance. A fully powered empirical ablation remains the primary future work priority (Section 5.4).

Instead, this section provides a theoretical component contribution analysis: the expected behaviour of each partial configuration, derived from the formal properties of each component and from the observed case study evidence. This approach, used in comparable hybrid-AI framework design papers (Yang et al., 2022; Bai et al., 2025; Guo et al., 2024; Klesel and Wittmann, 2025; Peng et al., 2025), allows the necessity of each layer to be assessed without requiring a second independent dataset. Table 21 presents the results.

The case study results provide indirect evidence for each component's necessity. The Grounded LLM's structured outputs (Table 5) show that without the formal fuzzy-DS reasoning layer, qualitative phrases such as ‘corrosion rates are elevated’ cannot be converted to quantitative risk rankings without fuzzification and defuzzification. Conversely, the original Ren et al. (2005) framework demonstrates that fuzzy-DS reasoning alone requires months of manual knowledge elicitation; the RAG layer replaces this with automated corpus retrieval. The Skill routing layer's necessity is evidenced by the sensitivity analysis (Section 4.4.3b): removing the routing confidence threshold increases routing ambiguity by 15 percentage points without changing the dominant risk category, confirming that Skill structure organises inference without introducing artificial constraints. These observations support the claim that each component addresses a distinct failure mode of the other partial configurations, even in the absence of a formal held-out empirical comparison.

Table 21
Theoretical component contribution analysis: expected behaviour of partial configurations.

Configuration	Key Characteristics	Expected Limitation Without Other Components	Evidence from Case Study
LLM-only (no RAG, no fuzzy-DS)	Parametric knowledge; free-text output	Hallucinated risk scores; no uncertainty quantification; non-reproducible outputs; no mathematical auditability	The grounded LLM outputs (Table 5) are qualitative text; without fuzzification they cannot produce the ranked belief distributions in Tables 11 and 12
RAG-only (no Skill routing, no fuzzy-DS)	Non-parametric retrieval; ranked text output	No formal risk ranking; cannot fuse multi-source evidence; no mathematical uncertainty quantification	Retrieved evidence chunks (Section 4.2.1b) are prose; the Belief/Plausibility intervals in Table 13 require the D–S synthesis pipeline
Fuzzy expert system only (no RAG, no Skills)	Full Mamdani + D–S reasoning; static rules; manual knowledge acquisition	Knowledge bottleneck: months of manual expert elicitation required; static rules cannot be updated without re-encoding	The original Ren et al. (2005) framework demonstrates this configuration; its 245-rule knowledge base required substantial expert panel time to construct
Full Skill-Enhanced framework (proposed)	RAG-automated knowledge acquisition + formal fuzzy-DS reasoning + expert-gated Skill updates	None by design; residual limitations are dataset scale and computational latency (Section 5.4)	Sections 4.2–4.5 demonstrate coherent end-to-end operation; $r = 0.987$, $F1 = 0.944$ confirm benchmark consistency

5.3. Expert-gated feedback results

The expert-gated feedback loop (Section 3.9) was exercised after the initial case study assessment. The benchmarking engine compared the system's outputs against independent expert assessments for 25 validation cases, producing a per-Skill alignment rate. The Structural Integrity Skill achieved 96% alignment (24/25 cases), the Maintenance Management Skill achieved 92% (23/25), and the Environmental Hazard Skill achieved 88% (22/25).

The feedback loop proposed four refinements: (1) adjusting the weight of template SI-2 from 0.96 to 0.98 based on strong expert endorsement, (2) widening the Maintenance Management Skill's “Marginal” TFN range from [3.5, 5.5] to [3.5, 6.0] years to capture borderline cases, (3) adding a new template to the Environmental Hazard Skill for combined wave-current loading, and (4) increasing the corrosion degradation weight from 0.93 to 0.95 for tropical-water FPSO deployments. All four refinements were approved by the domain expert and committed to the Skill Repository (version 1.1). After integration of the approved updates, re-benchmarking within the same 25-case validation set indicated a +3.5% alignment improvement (from 92% to 95.5%). This single-cycle result illustrates the operation of the feedback mechanism; its sustained effectiveness across multiple deployment cycles and on independent datasets remains to be demonstrated in future longitudinal studies.

5.4. Limitations and future work

Several limitations should be acknowledged. First, the case study is based on a single operational scenario (FPSO–shuttle tanker collision risk during tandem offloading), validation across a broader range of

offshore operations-subsea pipelines, fixed platforms, jack-up rigs-is needed to confirm generalisability. Second, the operational data used in the case study is partly hypothetical, integration with real-time monitoring systems and actual maintenance databases would strengthen the evidence base. Third, domain-specific Skill design currently requires initial expert input to define fuzzification configurations and inference templates, fully automated skill discovery from data remains an open research challenge.

Two further limitations warrant explicit acknowledgement. First, the computational cost of the combined RAG–LLM–fuzzy pipeline has not been systematically benchmarked. Sequential execution of vector retrieval, LLM inference, and fuzzy aggregation introduces latency that, in the current implementation, required approximately 15–25 s per failure-type assessment on standard server hardware. This suggests that periodic batch assessment is more appropriate than continuous real-time monitoring in the current form. Latency optimisation through pre-computed RAG indices, lighter inference models, or template pre-activation is deferred to future work. Second, the expert-gated feedback loop reduces the upfront knowledge-engineering burden but does not eliminate expert involvement: it redistributes it to ongoing weight-update review cycles. For high-frequency assessment deployments, this could introduce a validation throughput bottleneck. Partially automated acceptance criteria (e.g., automatic approval of weight changes below a statistically justified threshold) may reduce review load without compromising safety oversight, and should be explored in future iterations. Third, the framework's domain-independence is architectural rather than empirically demonstrated: the core reasoning pipeline contains no FPSO-specific hard-coded logic, but transferability to a new domain has not been validated. The domain-agnostic claim should be interpreted as a design property, empirical demonstration across at least one additional domain (e.g., nuclear maintenance or aviation safety) is a priority for future work.

Fifth, the framework has not been validated against live operational data from an active FPSO deployment. The case study dataset comprises expert-assessed values from Ren et al. (2005) supplemented by historical incident reports; no real-time sensor measurements, live maintenance logs, or operational telemetry from an active FPSO were used. This is the most significant limitation for practical deployment: the framework's ability to process streaming operational data, handle sensor noise, and maintain Skill relevance under evolving conditions remains undemonstrated. All performance claims should be understood as benchmark-validated within a controlled historical dataset, not as validated on live operational data.

Sixth, the expert-gated feedback loop redistributes rather than eliminates expert dependence. The framework substantially reduces the upfront knowledge-engineering cost (constructing a rule base from scratch) by automating knowledge acquisition through RAG. However, the feedback loop introduces a periodic expert review cost: domain experts must evaluate proposed weight updates and approve or reject each modification before it is committed to the Skill Repository. For high-frequency deployment contexts (where operational conditions change rapidly and Skill updates are frequent), this periodic review burden could represent a meaningful bottleneck. Future work should examine whether automated gate mechanisms (e.g., statistical significance thresholds for weight updates) can reduce review frequency without compromising safety oversight.

More fundamentally, the present case study exercises the adaptive feedback mechanism over a single benchmarking cycle on a 25-case dataset. The framework's adaptive capability i.e. Skill benchmarking, template weight updating, and version-controlled Skill Repository management, has been designed into the architecture and exercised in this controlled evaluation, but its long-term effectiveness has not been demonstrated. Multi-cycle convergence behaviour, Skill stability under evolving operational conditions, and robustness to distributional shift across FPSO asset types remain open empirical questions. Accordingly, claims regarding continuous learning effectiveness and long-term performance enhancement should be interpreted as architectural design goals rather than outcomes validated in the present work. Future longitudinal deployment studies are required before the adaptive mechanism can be credited with sustained performance benefits.

Seventh, the validation strategy establishes framework feasibility and structural coherence rather than operational effectiveness superiority. The case study demonstrates that the Skill-Enhanced framework faithfully reproduces an established expert benchmark ($r = 0.987$, $F1 = 0.944$) while providing additional capabilities (automated knowledge acquisition, Skill traceability, adaptive weight updating) that the benchmark framework lacks. What it does not demonstrate is that the framework produces better risk assessments than a human expert on unseen real-world scenarios, or that its outputs would improve safety decision outcomes in practice. The latter claim requires prospective operational validation, which is deferred to future work.

Eighth, several key design parameters (the routing confidence threshold at 0.70, the BPA confidence threshold at 0.50, and the weight-update learning rate at 0.30) were selected based on domain reasoning and validated through sensitivity analysis (Section 4.4.3b), but were not derived from a principled optimisation procedure. The sensitivity analysis confirms that the dominant risk-category outputs are stable across ±30% variation in each threshold, but the possibility that different threshold values would be optimal in other domains or deployment contexts cannot be excluded. A systematic threshold optimisation study using held-out operational data is recommended before operational deployment.

A further limitation concerns the retrieval quality of the RAG layer. The retrieval pipeline has not been independently benchmarked using standard information-retrieval metrics (precision, recall, mean reciprocal rank) against a labelled ground-truth set of relevant document chunks for FPSO safety queries. The routing confidence scores reported in Section 4.2.1a (all primary confidences >0.88) confirm consistent Skill assignment but do not directly measure whether the top-10 retrieved chunks contain the most relevant evidence for each query. In safety-critical applications, retrieval failures could propagate systematically into the fuzzy inference layer. Dedicated retrieval evaluation against an annotated FPSO safety query set is therefore a priority for future work.

Future work should address these limitations in three directions. Multi-domain validation should apply the framework to additional offshore asset types and to non-offshore safety domains (e.g., nuclear, aviation) to test the portability of the Skill architecture. Automated skill discovery should explore machine learning approaches to initialise Skill parameters from historical data, reducing dependence on expert elicitation. Real-time integration should connect the RAG layer to live operational data streams and sensor networks, enabling continuous risk monitoring rather than periodic assessment. Systematic ablation studies isolating each component's contribution – comparing configurations with RAG-only evidence acquisition, Skill-weighted inference without RAG, and LLM-only outputs without formal reasoning – are planned as part of the multi-domain validation programme, conducting such comparisons on the single 25-case scenario analysed here would provide insufficient statistical power to reliably attribute incremental performance gains to individual components.

6. Conclusion

This paper has presented a Skill-Enhanced Retrieval-Augmented Generation Expert System Framework for offshore accident analysis, integrating three methodological advances: (1) a Skill-Aware RAG layer that automates knowledge acquisition through domain-specific retrieval and grounded generation, (2) a Skill Repository that replaces static IF–THEN rule bases with adaptive, domain-specific reasoning modules containing weighted inference templates, calibrated membership functions, and performance benchmarks, and (3) an expert-gated feedback loop that enables continuous refinement of Skill parameters through

systematic benchmarking and expert validation.

The framework was demonstrated through a case study of FPSO collision risk assessment during tandem offloading, grounded in the original Ren et al. (2005; Wang et al., 2015) methodology. The Skill-Enhanced framework produced risk scores consistent with expert judgement (96% alignment for the Structural Integrity Skill) while reducing the number of inference rules from 245 fixed IF–THEN rules to 12 weighted templates. The Dempster-Shafer evidential synthesis identified Maintenance Failure as the dominant causal factor (62% combined belief), with Design Defect as the secondary contributor (24%). The expert-gated feedback loop demonstrated the version-control and weight-update mechanism, with a +3.5% alignment improvement within the 25-case validation set after a single refinement cycle; long-term adaptive effectiveness is identified as a priority for future empirical study.

These results demonstrate three things. First, the Skill-Enhanced RAG framework preserves the mathematical rigour of fuzzy-evidential reasoning while replacing the static rule base with a maintainable, modular, version-controlled Skill Repository - a knowledge-engineering advance that is independent of any accuracy comparison. Second, the framework achieves high consistency with the established Ren et al. (2005) expert benchmark ($r = 0.987$, $F1 = 0.944$), confirming that the knowledge-engineering enhancements do not degrade the underlying reasoning. Third, the expert-gated feedback and version-control mechanisms operate as designed within a single assessment cycle. What the case study does not demonstrate and what is explicitly deferred to future work is operational effectiveness superiority over expert assessment on independent scenarios, sustained adaptive improvement across multiple deployment cycles, and portability to domains beyond FPSO tandem offloading. The framework is positioned as a feasibility demonstration and knowledge-engineering architecture; its translation to operational practice requires longitudinal validation in live deployment settings.

Declaration of competing interest

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This project is partially supported by the European Union's Horizon 2020 Research and Innovation Programme RISE under grant agreement no. 823759 (REMESH).

Appendix

Table A1
Comparison of the Skill concept with related adaptive reasoning architectures.

Property	Static IF–THEN Rules	ANFIS	Parameterised Fuzzy Module	Skill (Proposed)
Domain routing	Manual rule selection	Manual	Manual	Automated via evidence profile E(r)
Template weights	Uniform (fixed)	Gradient-adapted (training data)	Prescribed by engineer	Benchmarking-derived from expert-validated outcomes
Fuzzification	Fixed MFs (design time)	Adapted MFs (training)	Fixed or configurable	Per-Skill calibrated; version-controlled
Knowledge acquisition	Manual expert elicitation	Training dataset required	Manual elicitation	RAG-automated from document corpus
Auditability	Rule-level trace	Limited (black-box adaptation)	Module-level trace	Full Skill trace with version log and approval record
Update mechanism	Manual re-encoding	Retraining required	Manual re-parameterisation	Expert-gated weight update; no rule re-encoding

Table A2 provides the complete 12-template Skill inference inventory with initial and post-feedback weights across all five hazard domains.

Table A2
Complete inventory of the 12 Skill inference templates

Skill	Template ID	Domain Focus	Initial Weight (w_i)	Post-Feedback Weight
Structural Integrity (SI)	SI-1	Corrosion & inspection intervals	0.95	0.968
	SI-2	Fatigue cracking & material grade	0.96	0.960
	SI-3	Structural load & integrity monitoring	0.90	0.890
Mooring and Marine (MM)	MM-1	Mooring line tension & failure modes	0.88	0.880
	MM-2	Marine spread positioning & drift	0.85	0.850
Equipment & Hydraulics (EH)	EH-1	Hydraulic system & CPP actuation	0.92	0.963
	EH-2	Thruster mechanical failure & maintenance	0.88	0.880
Human Factors (HF)	HF-1	DP operator error & training adequacy	0.87	0.870
	HF-2	Human-machine interface & alarm management	0.83	0.830
System Integration (SysI)	SysI-1	DP system software & sensor fusion	0.91	0.910
	SysI-2	PRS failure & redundancy	0.89	0.890
	SysI-3	Multi-system interaction & cascade failure	0.86	0.860

Note: The complete inventory of the 12 Skill inference templates shows initial weights (assigned from prior expert knowledge) and post-feedback weights following the FPSO case study evaluation (Section 4.5).

References

Ahmed, A., et al., 2019. Knowledge-Based Systems Survey. Int. J. Appl. Eng. Res. 3 (7), 1–22.

Bai, G., Wang, L., Chen, H., 2025. Construction of intelligent decision support systems through integration of retrieval-augmented generation and knowledge graphs. Sci. Rep. 15, 19257.

Ceylan, B.O., 2023. Shipboard compressor system risk analysis by using rule-based fuzzy FMEA for preventing major marine accidents. Ocean. Eng. 272, 113888. https://doi.org/10.1016/j.oceaneng.2023.113888.

Ceylan, B.O., 2025. Control theory-based fuzzy Fine–Kinney risk assessment for boiler automation system from the maritime autonomous surface ships (MASS) perspective. Ocean. Eng. 322, 120444. https://doi.org/10.1016/j.oceaneng.2025.120444.

Chang, D.Y., 1996. Applications of the extent analysis method on fuzzy AHP. Eur. J. Oper. Res. 95 (3), 649–655.

Chen, H., Vinnem, J.E., Chen, H., 2002. Collision risk analysis of FPSO–Tanker offloading operation. In: OMAE2002. ASME.
Compton, P., Jansen, R., 1990. A philosophical basis for knowledge acquisition. Knowl. Acquis. 2 (3), 241–257.
Dempster, A.P., 1968. A generalization of Bayesian inference. J. Roy. Stat. Soc. B 30 (2), 205–247.
Denoeux, T., 2008. Conjunctive and disjunctive combination of belief functions induced by non-distinct bodies of evidence. Artif. Intell. 172, 234–264. https://doi.org/10.1016/j.artint.2007.05.008.
Feigenbaum, E.A., 1977. The art of artificial intelligence: themes and case studies of knowledge engineering. In: Proceedings of the 5th International Joint Conference on Artificial Intelligence, pp. 1014–1029.
Gaidai, O., Cao, Y., Xu, Y., 2023. Offloading operation bivariate extreme response statistics for FPSO vessel. Sci. Rep. 13 (1), 4695. https://doi.org/10.1038/s41598-023-31533-8. PMID: 36949113; PMCID: PMC10033654.
Guo, T., Yang, Q., Wang, C., Liu, Y., Li, P., Tang, J., Li, D., Wen, Y., 2024. KnowledgeNavigator: leveraging large language models for enhanced reasoning over knowledge graph. Complex Intell. Syst. 10, 7063–7076. https://doi.org/10.1007/s40747-024-01527-8.
Health and Safety Executive (HSE), 2007. Accident Statistics for Floating Offshore Units on the UK Continental Shelf 1980-2005. Prepared by Det Norske Veritas for the Health and Safety Executive.
Health and Safety Executive (HSE), 2012. Offshore accident and incident statistics report. Rev 6. London: HSE Books.
Jiang, J., 2014. Risk Modeling of DP Operation for Offshore Tandem Offloading. Norwegian University of Science and Technology (NTNU). Master's thesis.
Karanović, V., Ceylan, B.O., Jocanović, M., 2024. Reliable ships: a fuzzy FMEA based risk analysis on four-ram type hydraulic steering system. Ocean. Eng. 314, 119758. https://doi.org/10.1016/j.oceaneng.2024.119758.
Klesel, M., Wittmann, H.F., 2025. Retrieval-Augmented generation (RAG). In: Business & Information Systems Engineering, 67, pp. 551–561. https://doi.org/10.1007/s12599-025-00945-3.
Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W., Riedel, S., Kiela, D., 2020. Retrieval-Augmented generation for knowledge-intensive NLP tasks. Adv. Neural Inf. Process. Syst. 33, 9459–9474.
Liu, J., Yang, J.B., Ruan, D., Martinez, L., Wang, J., 2005. Self-tuning of fuzzy belief rule bases for engineering system safety analysis. Ann. Oper. Res. 163 (1), 143–168.
Mamdani, E.H., Assilian, S., 1975. An experiment in linguistic synthesis with a fuzzy logic controller. Int. J. Man Mach. Stud. 7 (1), 1–13.
Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., Zhang, Y., Tang, S., 2025. Graph Retrieval-Augmented Generation: a Survey. ACM Transactions on Information Systems. https://doi.org/10.1145/3777378.
Ren, J., Jenkinson, I., Wang, J., Xu, D.L., Yang, J.B., 2005. An offshore safety assessment framework using fuzzy reasoning and evidential synthesis approaches. J.Marine Eng. Technol. 4 (1), 3–16.
Ren, J., Wang, J., Jenkinson, I., Xu, D.L., Yang, J.B., 2009. A Bayesian network approach for offshore risk analysis through fuzzy rule-based Bayesian reasoning. Int. J. Intell. Syst. 24 (1), 69–87.
Shafer, G., 1976. A Mathematical Theory of Evidence. Princeton University Press, Princeton.
Wang, H.Y., Ren, J., Yang, J.Q., Wang, J., 2015. Evaluating the effectiveness of ERS for vessel oil spills using fuzzy evidential reasoning. Ocean Sys. Eng. 5 (3), 161–179.
Yang, J.B., 2001. Rule and utility based evidential reasoning approach for multiattribute decision analysis under uncertainties. Eur. J. Oper. Res. 131 (1), 31–68.
Yang, L.-H., Liu, J., Ye, F.-F., Wang, Y.-M., Nugent, C., Wang, H., Martínez, L., 2022. Highly explainable cumulative belief rule-based system with effective rule-base modeling and inference scheme. Knowl. Base Syst. 240, 107805.
Yang, J.B., Xu, D.L., 2002. On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty. IEEE Trans. Syst. Man Cybern. Part A 32 (3), 289–304.
Yoran, Ori, Wolfson, Tomer, Ram, Ori, Berant, Jonathan, 2023. Making retrieval-augmented language models robust to irrelevant context. https://arxiv.org/abs/2310.01558.
Zadeh, L.A., 1965. Fuzzy sets. Inf. Control 8 (3), 338–353.