Human–AI Collaborative Retrieval-Augmented Decision Intelligence for Enterprise Knowledge Bases in Financial Services

Riswan Efendi Tarigan[1,*], Kevin Ariel Zen[2]

[1,2]*Department of Information Systems, Faculty of AI and Data Science, Universitas Pelita Harapan, Indonesia*

(Received: March 1, 2026; Revised: May 1, 2026; Accepted: July 8, 2026; Available online: July 20, 2026)

Abstract

Financial service institutions increasingly require knowledge systems that can retrieve policy-relevant information, synthesize organizational evidence, and deliver explainable decision support under strict governance constraints. This study proposes a Retrieval-Augmented Decision Intelligence framework for enterprise knowledge bases in financial services by integrating semantic retrieval, metadata-aware re-ranking, grounded augmentation, and explainable response generation into a unified managerial intelligence pipeline. The framework was evaluated using enterprise-style query scenarios covering compliance clarification, product guidance, risk review support, and service resolution. The results showed that the proposed framework outperformed a baseline generative model across all major dimensions, achieving Precision@5 of 0.86 compared with 0.71, NDCG of 0.88 compared with 0.74, grounding score of 0.84 compared with 0.68, usefulness score of 0.87 compared with 0.70, and an overall effectiveness index of 0.86 compared with 0.71. Category-level analysis indicated that hybrid re-ranking improved ranking effectiveness in all query types, with NDCG increasing from 0.84 to 0.89 for compliance queries, from 0.87 to 0.91 for product guidance, from 0.82 to 0.87 for risk review support, and from 0.79 to 0.85 for service resolution. Grounding performance remained strong across categories, reaching 0.90 for compliance clarification, 0.88 for product guidance, 0.82 for risk review, and 0.79 for service resolution, demonstrating that retrieved enterprise evidence substantially constrained unsupported generation. Expert evaluation further confirmed high managerial value, with average scores of 4.4 for clarity, 4.5 for actionability, 4.3 for trustworthiness, 4.4 for interpretability, and 4.5 for decision value on a five-point scale. Failure analysis identified outdated policy retrieval, cross-document ambiguity, terminology mismatch, insufficient escalation signaling, and partial evidence coverage as the main residual weaknesses, with outdated policy retrieval accounting for 18 observed cases and cross-document ambiguity for 14. These findings indicate that retrieval-augmented architectures can move beyond information access and function as decision intelligence systems that support traceable, evidence-grounded, and operationally meaningful knowledge work in regulated financial environments.

Keywords: Retrieval-Augmented Generation, Decision Intelligence, Enterprise Knowledge Base, Financial Services, Explainable AI, Knowledge Management, Hybrid Retrieval, Managerial Decision Support

1. Introduction

Financial service institutions operate in information environments characterized by dense regulation, rapidly changing policy rules, fragmented operational memory, and strong accountability requirements. In such settings, managerial decisions often depend on timely access to internal policies, product rules, compliance interpretations, and historical case resolutions distributed across heterogeneous enterprise repositories. Conventional large language models are not sufficient for this setting because their internal knowledge is static, difficult to verify, and prone to hallucination in knowledge-intensive tasks, while retrieval-augmented generation was introduced precisely to connect generation with explicit external evidence and dense retrieval mechanisms that improve passage selection quality [1], [2].

Recent research has positioned retrieval-augmented systems as a major architectural direction for knowledge-intensive artificial intelligence because they improve factual grounding, provide access to updated corpora, and support more reliable response generation than closed parametric memory alone. At the same time, the RAG literature has expanded from foundational architectures to broader design discussions involving retrieval strategies, training alignment, and system-level integration with organizational knowledge assets. These developments suggest that RAG is increasingly

***Corresponding author: Riswan Efendi Tarigan (riswan.tarigan@uph.edu)**
DOI: https://doi.org/10.47738/ijaim.v6i2.125

relevant for enterprise information access, yet most discussions still remain general and are not centered on the specific managerial logic of regulated financial organizations [3], [4].

This gap is important because enterprise deployment is not simply a scaled-up version of open-domain question answering. In organizations, especially financial institutions, retrieved content must satisfy traceability, version validity, authority control, and governance constraints in addition to topical relevance. Recent work in business information systems and knowledge management highlights that RAG has clear promise for organizational data access, while AI-enabled knowledge management still faces implementation challenges related to trust, governance, data silos, and institutional integration. These concerns indicate that enterprise knowledge retrieval must be designed as a decision-support process rather than only as a search-enhanced text generation task [5], [6].

A second gap concerns the structure of the knowledge problem itself. Research on question answering over knowledge bases has shown that retrieval, path matching, and structured reasoning remain central when the objective is to answer domain-specific questions from explicit repositories rather than from general web-scale corpora. Reviews and methodological studies in KBQA consistently show that system quality depends on the interaction between representation, retrieval precision, and answer construction. However, these studies are largely oriented toward generic knowledge graphs or public benchmarks and do not directly address hybrid enterprise repositories that combine policies, procedural texts, archived cases, and metadata-rich internal documents [7], [8].

This limitation becomes more serious in financial services because AI adoption in finance has accelerated across decision-making, risk analysis, fraud detection, and customer-facing operations, creating a stronger need for systems that can combine analytical capability with institutional reliability. Survey evidence shows that AI is already deeply embedded in financial processes, but the literature also emphasizes the need for stronger domain adaptation, governance readiness, and contextual decision support. Consequently, an enterprise knowledge system for finance cannot rely on generic semantic retrieval alone. It must instead connect managerial intent with regulated internal knowledge in a way that is operationally useful and organizationally defensible [9], [10].

Explainability strengthens this argument further. Financial AI systems are increasingly evaluated not only by predictive or generative performance but also by interpretability, transparency, and auditability. Reviews in finance and explainable AI repeatedly emphasize that opaque outputs are problematic in domains where decisions may affect compliance, customer treatment, and internal accountability. Existing studies have therefore underscored the need for AI systems whose outputs can be justified to different stakeholders. Yet current RAG implementations for enterprise environments rarely frame explainability as part of decision intelligence, and recent work on enterprise knowledge-base question answering still focuses more on answer generation than on managerial traceability and actionability [11]-[13].

Based on these gaps, this study proposes Retrieval-Augmented Decision Intelligence for Enterprise Knowledge Bases in Financial Services, a framework that extends conventional RAG into a managerial decision-support architecture. The study is designed to address three core problems: fragmented enterprise knowledge access, limited governance-aware retrieval, and insufficiently explainable response generation for financial decision contexts. The novelty lies in integrating hybrid retrieval, metadata-aware re-ranking, grounded augmentation, and explainable decision output into a unified enterprise framework explicitly tailored to financial services. In this way, the study contributes to AI in Management by repositioning RAG from a knowledge-access mechanism into a decision intelligence system for institutional use.

2. Literature Review

Retrieval-augmented generation has emerged as a major response to the core limitations of large language models in knowledge-intensive settings, especially hallucination, stale parametric memory, and weak traceability. Recent surveys describe RAG as an architecture that couple's retrieval, augmentation, and response generation so that model outputs can be grounded in external corpora rather than inferred solely from internal parameters. This literature consistently positions retrieval quality, context construction, and response control as the three main determinants of system reliability in domain-specific environments [14], [15].

Within enterprise contexts, the discussion has shifted from generic document retrieval toward governed knowledge infrastructures that can support organizational reasoning. Studies on enterprise knowledge graphs and industrial question-answering show that structured knowledge assets are valuable because they improve entity alignment, preserve institutional semantics, and provide a stronger basis for trusted enterprise question answering. In this line of work, knowledge graphs are not treated as optional enrichments, but as mechanisms for constraining ambiguity, validating query interpretation, and supporting controlled data access in organizational environments [16], [17].

The enterprise knowledge management literature further shows that scalable question answering depends on how institutional knowledge is modeled, curated, and operationalized. Research on knowledge-graph development practices emphasizes data integration, ontology quality, and scalable architectures as prerequisites for robust enterprise knowledge services. Related work on smart management systems demonstrates that knowledge graphs can transform fragmented operational procedures into reusable and machine-accessible decision resources. Together, these studies suggest that retrieval performance is inseparable from the quality of underlying knowledge engineering practices [18], [19].

A related stream of literature concerns question answering over structured and semi-structured repositories. Studies on retrieval-augmented knowledge graph reasoning and domain-specific intelligent question-answering systems show that answer quality improves when semantic matching is complemented by graph reasoning, path selection, or template-aware query processing. This is particularly relevant for enterprise domains, where user questions often depend on entity relations, rule dependencies, and contextual qualifiers that cannot be resolved by lexical similarity alone. The literature therefore indicates that domain question answering benefits from hybrid architectures that combine retrieval depth with relational reasoning capacity [20], [21].

Another major theme in the literature is explainability. Reviews in information systems and trustworthy AI argue that organizational acceptance of AI depends on more than predictive or generative performance. Explanations, stakeholder awareness, and system transparency are treated as necessary conditions for trust, oversight, and responsible operational use. This perspective is particularly relevant for enterprise decision support because AI outputs must often be interpreted by multiple actors, including managers, analysts, compliance personnel, and auditors. As a result, explainability is increasingly viewed as an architectural requirement rather than a post hoc reporting feature [22], [23].

In financial services, these concerns become even more pronounced because decision environments are highly regulated and operationally sensitive. Reviews of AI in customer-facing finance and explainable AI in finance show that current research is still split between algorithmic performance studies and adoption-oriented analyses, while back-office knowledge operations and enterprise decision support remain comparatively underexplored. This creates an unresolved gap between the technical promise of retrieval-augmented systems and the practical requirements of financial institutions, where answers must be relevant, evidence-linked, policy-aware, and suitable for managerial use [24], [21]. The present study addresses this gap by synthesizing RAG, enterprise knowledge engineering, and explainable decision support into a single framework for financial-service knowledge bases.

3. Methodology

3.1. Research Design and System Architecture

This study adopts a design science research approach combined with an applied machine learning pipeline to construct a retrieval-augmented decision intelligence framework for enterprise knowledge bases in financial services. The design science perspective is suitable because the study does not only observe an organizational problem but also produces a functional artifact that addresses knowledge fragmentation, inconsistent retrieval quality, and delayed managerial interpretation. The methodological orientation therefore integrates artifact construction, system evaluation, and contextual performance analysis within a financially regulated knowledge environment [3], [21].

The proposed architecture consists of five tightly linked layers, namely data ingestion, semantic knowledge representation, retrieval and re-ranking, augmentation and reasoning, and decision output generation. Each layer performs a distinct role while remaining interdependent with the others. The architecture is intended to support internal decision tasks such as policy interpretation, risk review support, compliance clarification, and product knowledge

resolution. In this setting, retrieval is not treated as a standalone search process but as a controlled precursor to managerial inference.

To represent the functional relation among input knowledge, retrieval precision, and decision relevance, the methodological model formalizes decision intelligence utility as follows:

$$DI = \alpha R_p + \beta C_c + \gamma A_q \quad (1)$$

In this formulation, DI denotes decision intelligence quality, R_p represents retrieval precision, C_c denotes contextual completeness, and A_q refers to augmentation quality. The coefficients α, β, and γ express the weighted contribution of each component. The equation clarifies that managerial usefulness emerges from both accurate retrieval and sufficiently contextualized synthesis rather than from isolated language generation alone.

The research architecture (figure 1) is implemented as a modular pipeline to enable traceability, replacement of individual components, and auditability of decision support outputs. This modularity is critical in financial services because system behavior must remain inspectable under governance, risk, and compliance requirements. The methodological design therefore emphasizes not only performance optimization but also operational transparency. Each stage logs intermediate outputs, allowing the study to evaluate how evidence is retrieved, transformed, and ultimately presented as a decision-support response [5], [22].

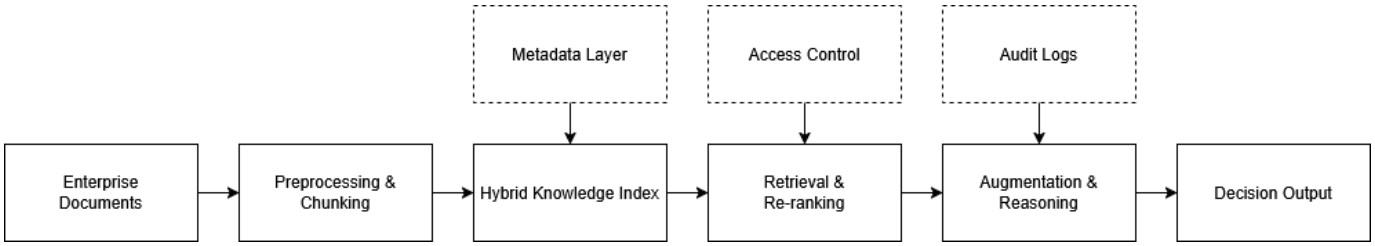

Figure 1. End-to-End Retrieval-Augmented Decision Intelligence Architecture

The figure visualizes the methodological backbone of the study as a sequential yet governed pipeline. It begins with enterprise document sources and moves through preprocessing, indexing, retrieval, augmentation, and decision output generation. The upper dashed boxes indicate that metadata control, access restrictions, and audit logging are not auxiliary features but embedded governance layers.

Table 1 condenses the architecture into functional layers that can be read quickly by journal reviewers and readers. It shows that each stage has a distinct operational role, clear input and output boundaries, and a specific governance concern. The structure is useful because it turns the conceptual architecture into a compact methodological map. In the context of financial services, this table also emphasizes that governance dimensions such as provenance, version control, and audit readiness are embedded throughout the system rather than appended at the end.

Table 1. Summary of Architectural Layers, Functions, Inputs, and Outputs

Layer	Main Function	Primary Input	Primary Output	Governance Focus
Data Ingestion	Collect and normalize enterprise documents	Policies, SOPs, case records	Structured document repository	Document provenance
Knowledge Representation	Create chunks, embeddings, and metadata links	Normalized documents	Hybrid indexed knowledge units	Version control
Retrieval and Re-ranking	Identify and prioritize relevant evidence	User query and indexed knowledge	Ranked evidence candidates	Authority and recency
Augmentation and Reasoning	Build grounded context for inference	Top-ranked evidence	Contextual prompt package	Traceability
Decision Output	Generate actionable and explainable response	Augmented reasoning context	Answer, rationale, references	Audit readiness

3.2.Enterprise Data Acquisition and Knowledge Base Construction

The knowledge base is constructed from heterogeneous enterprise content typically found in financial service institutions, including internal policy manuals, standard operating procedures, compliance circulars, risk memoranda, customer service knowledge articles, product guidelines, and archived case resolutions. These documents are collected through controlled extraction procedures and then normalized into a unified textual repository. During this stage, document provenance, timestamp, business unit origin, and confidentiality class are preserved as metadata attributes because they later influence filtering, ranking, and access control decisions [15], [16].

After collection, the corpus undergoes preprocessing through text cleaning, section segmentation, duplication control, metadata harmonization, and semantic chunking. The chunking procedure is particularly important because overly large segments reduce retrieval specificity, whereas overly short segments weaken semantic coherence. The methodology therefore uses structure-aware segmentation based on headings, clause boundaries, and policy logic units. This enables each knowledge unit to remain meaningful in isolation while still preserving its relationship to the originating enterprise document and regulatory context.

To formalize chunk generation from raw documents, the segmentation stage uses the following relation:

$$N_c = \left\lceil \frac{L_d}{L_s \times \delta} \right\rceil \tag{2}$$

Here, N_c denotes the number of chunks generated from a document, L_d is total document length, L_s is targeting segment length, and δ is an overlap control factor. This equation helps regulate the granularity of the knowledge base. In methodological terms, the goal is to maximize semantic locality while preserving enough contextual continuity for accurate downstream retrieval and augmentation.

The processed chunks are embedded into a vector representation space and stored alongside symbolic metadata in a hybrid knowledge index. This hybrid construction is methodologically significant because financial knowledge is rarely reducible to semantics alone. Business unit, document recency, regulatory priority, and approval status often affect the practical value of a retrieved passage. The knowledge base therefore combines dense embedding storage with structured metadata fields, enabling the system to support both semantic similarity search and governance-sensitive filtering during query execution [18], [19].

Figure 2 represents the document engineering workflow used to construct the enterprise knowledge base. It begins with diverse internal sources and shows the transformation of raw records into structured, semantically chunked, and metadata-enriched knowledge units. The lower auxiliary elements make visible the diversity of document types and metadata fields that shape retrieval relevance later in the pipeline. The figure is methodologically valuable because it demonstrates that knowledge base construction is not a trivial upload process but a staged normalization and representation procedure.

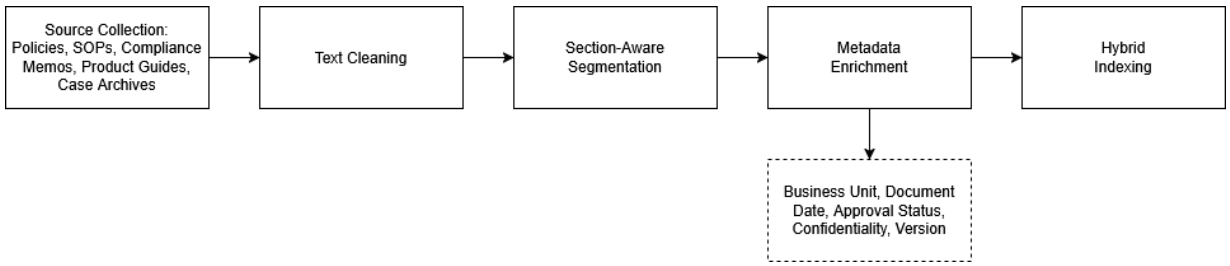

Figure 2. Enterprise Knowledge Base Construction Workflow

Table 2 translates the ingestion workflow into a concrete inventory of enterprise content types and their treatment within the methodology. Each row clarifies that different documents require different preprocessing and segmentation logic to preserve business meaning. The table also shows how metadata capture is aligned with document function, which is essential for later filtering and ranking.

Table 2. Enterprise Document Types, Preprocessing Steps, and Indexing Outputs

Document Type	Preprocessing Action	Segmentation Basis	Metadata Captured	Index Output
Policy Manuals	Cleaning and clause normalization	Heading and rule block	Version, approval date	Semantic-policy chunks
SOP Documents	Procedure extraction	Step sequence	Department, status	Operational chunks
Compliance Circulars	Terminology alignment	Regulatory clause	Effective date, authority	Compliance chunks
Product Guides	Noise removal	Feature and condition block	Product line, revision	Product knowledge units
Case Archives	Resolution summarization	Issue-response pair	Case type, closure date	Historical resolution chunks

3.3.Retrieval, Re-Ranking, and Augmentation Mechanism

The retrieval stage begins when a user submits a managerial or operational query related to financial products, internal controls, policy interpretation, or compliance obligations. The query is encoded into a dense vector and matched against the knowledge base to identify candidate passages. Candidate retrieval is intentionally broad in the first stage so that potentially relevant evidence is not prematurely excluded. This high-recall stage is then followed by re-ranking to improve relevance ordering and reduce noise before contextual augmentation is performed [1], [2].

The re-ranking mechanism evaluates the semantic compatibility between the query and the retrieved passages while also considering metadata alignment, business criticality, and recency. This is especially necessary in financial services, where a passage may appear semantically similar but still be unsuitable because it belongs to an outdated policy version or an unrelated product line. The retrieval methodology therefore uses a hybrid score that combines vector similarity with rule-based weighting derived from enterprise metadata and document authority levels.

The scoring function for candidate ranking is defined as follows:

$$S(d,q) = \lambda \text{sim}(e_q, e_d) + (1-\lambda)(\eta M_d + \theta T_d + \kappa A_d) \quad (3)$$

In this equation, $S(d,q)$ is the total relevance score between document chunk d and query q, $\text{sim}(e_q,e_d)$ denotes embedding similarity, M_d represents metadata relevance, T_d indicates temporal validity, and A_d denotes document authority. The scalar λ controls the balance between semantic and structured evidence. This formulation reflects the practical reality that enterprise retrieval quality depends on both linguistic proximity and institutional validity.

After re-ranking, the selected passages are assembled into an augmented prompt context for the reasoning module. The augmentation stage includes redundancy suppression, contradiction screening, and context window control to ensure that the generated response is grounded in the most relevant evidence. Rather than allowing unrestricted language generation, the system constrains output synthesis using retrieved financial knowledge units. This step is methodologically essential because it reduces hallucination risk, improves traceability, and anchors managerial recommendations to verifiable enterprise documentation [14], [20].

Figure 3 explains the operational core of the proposed model by separating retrieval into recall-oriented search and precision-oriented re-ranking. The lower scoring boxes indicate that document relevance is computed from multiple dimensions rather than semantic similarity alone. The post-processing blocks show that evidence augmentation is controlled through redundancy and conflict handling before the context is passed to the reasoning model.

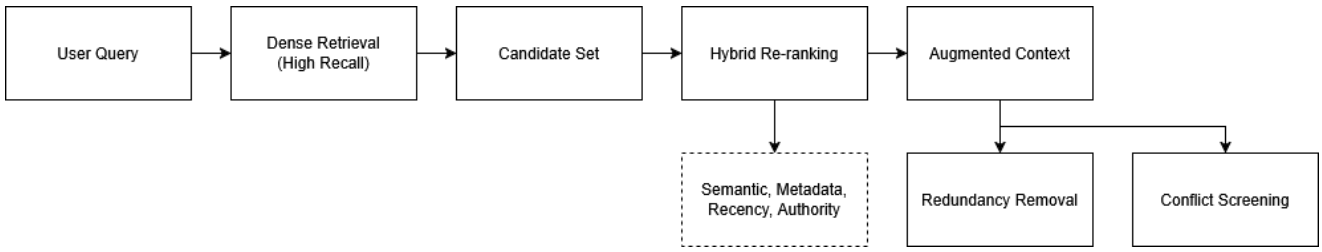

Figure 3. Two-Stage Retrieval and Augmentation Mechanism

Table 3 lists the principal computational signals that shape retrieval quality in the methodology. It shows that the model does not rely on a single similarity function, but on a composite mechanism that reflects semantic, temporal, and organizational validity. The table is especially useful because it links each component to a specific operational role and practical effect. In financial services, such explicit parameterization is important for explaining why a retrieved passage is considered fit for managerial decision support.

Table 3. Retrieval Parameters, Ranking Signals, and Augmentation Controls

Component	Purpose	Main Signal	Operational Role	Expected Effect
Dense Retrieval	Capture broad candidate evidence	Embedding similarity	High recall screening	Fewer missed relevant chunks
Metadata Matching	Align query with business context	Department and document tags	Context filtering	Improved enterprise relevance
Temporal Weighting	Prioritize valid and recent content	Effective date	Recency adjustment	Lower outdated retrieval risk
Authority Weighting	Prefer approved source material	Approval and policy level	Trust calibration	Higher institutional validity
Augmentation Control	Prepare grounded reasoning context	Redundancy and conflict checks	Context assembly	Cleaner evidence package

Algorithm 1 Retrieval-Augmented Decision Intelligence Pipeline

```
Input:
   Query q
   Hybrid knowledge base KB
   Top-k retrieval size k
   Re-ranking size r
Output:
   Decision-support response y
   Evidence set E*

Begin
   q_clean <- preprocess(q)
   q_embed <- encode(q_clean)
   C <- retrieve_top_k(KB, q_embed, k)
   For each chunk c in C do
      sem_score[c] <- semantic_similarity(q_embed, embedding(c))
      meta_score[c] <- metadata_match(q, metadata(c))
      time_score[c] <- temporal_validity(metadata(c))
      auth_score[c] <- authority_weight(metadata(c))
      total_score[c] <- combine(sem_score[c], meta_score[c], time_score[c], auth_score[c])
   End For
   R <- rank_descending(C, total_score)
   E <- select_top_r(R, r)
```

```
    E* <- remove_redundancy_and_conflicts(E)
    prompt <- build_augmented_context(q_clean, E*)
    y <- generate_grounded_response(prompt)
    return y, E*
End
```

3.4. Decision Intelligence Layer and Analytical Inference

The decision intelligence layer transforms retrieved enterprise evidence into interpretable analytical outputs that support managerial action. This layer does not merely summarize retrieved text. Instead, it restructures the evidence into decision-oriented forms such as recommended actions, risk flags, policy interpretations, exception pathways, or comparative option assessments. In the context of financial services, such outputs are valuable for frontline operations, supervisory review, internal governance, and knowledge-driven escalation workflows where consistency and justification are critical [11], [12].

To improve analytical usability, the inference procedure applies task framing based on the decision category associated with each query. Queries related to compliance clarification are treated differently from those concerning customer resolution, product suitability, or internal risk guidance. This task-aware prompting strategy helps the model produce outputs that align with the logic of the business process rather than generic summaries. As a result, the system can present retrieved evidence in a form that is operationally meaningful for a specific financial management function.

The confidence of the decision-support output is modeled through an evidence aggregation function:

$$C_f = \frac{1}{n}\sum_{i=1}^{n} w_i \cdot r_i \tag{4}$$

Here, C_f denotes confidence in the final response, r_i is the relevance contribution of evidence item i, and w_i is the importance weight assigned to that item. The formula expresses that output confidence should increase when multiple high-quality evidence units converge on a consistent conclusion. This is especially useful in enterprise environments where the quality of a recommendation depends on the density and coherence of documentary support.

The final output is designed to include three components, namely a concise answer, a justification layer, and a traceable evidence reference set. This tripartite structure improves organizational trust because decision consumers can see not only the conclusion but also the logic and source basis behind it. Methodologically, this design aligns the system with explainable decision support principles. In financial services, such explainability is not optional because operational decisions often require later review by auditors, supervisors, or compliance personnel [21], [23].

Figure 4 presents the transformation of retrieved evidence into decision-oriented output. Inputs from evidence, query intent, and business context enter a central intelligence layer that frames the task, fuses evidence, and constructs a rationale. The outputs are separated into recommendation, justification, and reference trace, which mirrors the explainability requirements of financial decision environments.

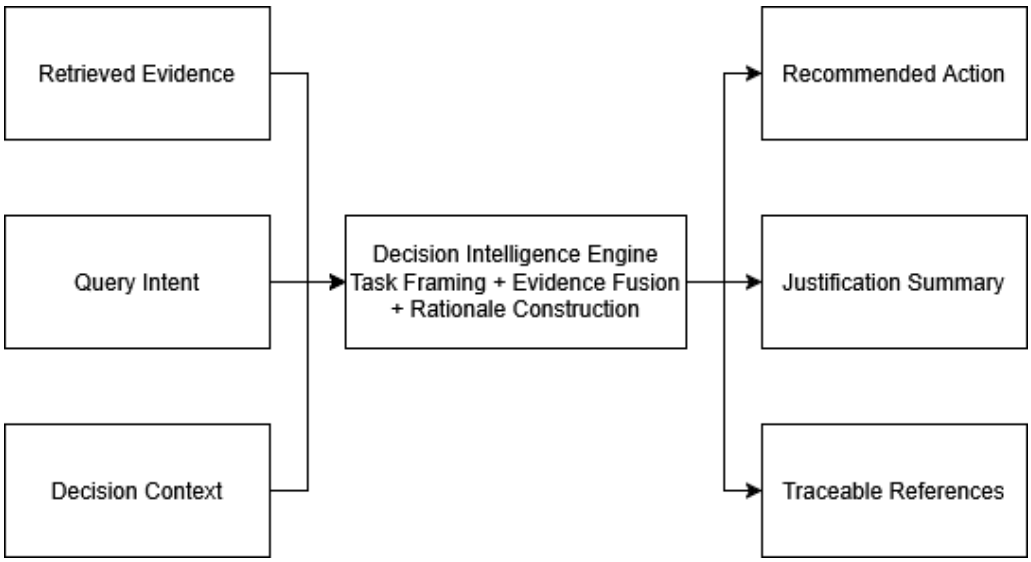

Figure 4. Decision Intelligence Layer for Explainable Managerial Output

Table 4 clarifies that the decision intelligence layer is task-sensitive rather than generic. Different query categories require different reasoning emphases, output forms, and explainability mechanisms. By showing this mapping explicitly, the table strengthens the methodological claim that decision support quality depends on alignment between task type and inference structure.

Table 4. Decision Task Types, Inference Logic, and Explainability Outputs

Decision Task	Primary Objective	Inference Focus	Output Form	Explainability Element
Compliance Clarification	Interpret applicable rules	Policy consistency	Rule-based answer	Cited policy passages
Risk Review Support	Highlight control concerns	Risk signals and exceptions	Risk notes and action cue	Evidence-backed flags
Product Guidance	Match rules with product conditions	Eligibility and features	Recommendation summary	Supporting product clauses
Service Resolution	Resolve operational inquiries	Procedure alignment	Stepwise response	Source-linked rationale
Escalation Support	Identify cases needing review	Threshold and exception logic	Escalation suggestion	Traceable trigger basis

3.5.Evaluation Design, Metrics, and Validation Strategy

The evaluation stage uses a multi-dimensional design to assess retrieval quality, augmentation reliability, decision usefulness, and operational interpretability. Because the framework serves financial service knowledge work, system quality cannot be evaluated through language fluency alone. The methodology therefore combines information retrieval metrics, response grounding metrics, and expert-based usefulness assessment. This integrated approach allows performance to be examined from both computational and organizational perspectives, ensuring that the proposed artifact is evaluated as a decision support system rather than only as a language model application.

The validation dataset consists of enterprise-style queries derived from realistic financial service tasks, including compliance checks, product rule clarification, procedural interpretation, and risk-related information lookup. Each query is paired with a curated relevance judgment and, where applicable, an expected decision rationale prepared by domain-informed evaluators. This setup makes it possible to evaluate the full pipeline, from retrieval to analytical output. It also supports error analysis on cases involving outdated policies, ambiguous terminology, and cross-departmental knowledge overlap [7], [24].

To quantify overall system effectiveness, the study defines a composite evaluation index as follows:

$$E_{sys} = \mu P@k + \nu NDCG + \rho G_s + \sigma U_s \quad (5)$$

In this equation, E_{sys} denotes total system effectiveness, $P@k$ represents precision at rank k, $NDCG$ measures ranking quality, G_s denotes grounding score, and U_s indicates expert-assessed usefulness. The coefficients μ, ν, ρ, and σ regulate the contribution of each metric. This formulation is appropriate because decision intelligence quality in financial services depends simultaneously on retrieval correctness, evidential grounding, and managerial applicability.

The evaluation procedure includes quantitative testing, qualitative expert review, and failure pattern analysis. Quantitative testing measures retrieval and output quality under controlled query sets. Qualitative review examines whether the generated decision responses are interpretable, justifiable, and operationally safe. Failure analysis identifies recurring weaknesses such as policy-version confusion, insufficient evidence coverage, or overgeneralized recommendations. Together, these procedures provide a rigorous validation strategy for determining whether retrieval-augmented decision intelligence can serve as a reliable enterprise knowledge instrument in financial services [10], [17].

Figure 5 summarizes the evaluation design as a multi-perspective framework. Instead of relying only on retrieval statistics, it integrates ranking performance, evidence grounding, expert usefulness review, and structured failure analysis. The dashed dataset box indicates that evaluation is driven by realistic enterprise query categories rather than abstract benchmark prompts. This figure is methodologically important because it demonstrates that the proposed system is assessed as a decision support artifact with both computational and managerial criteria.

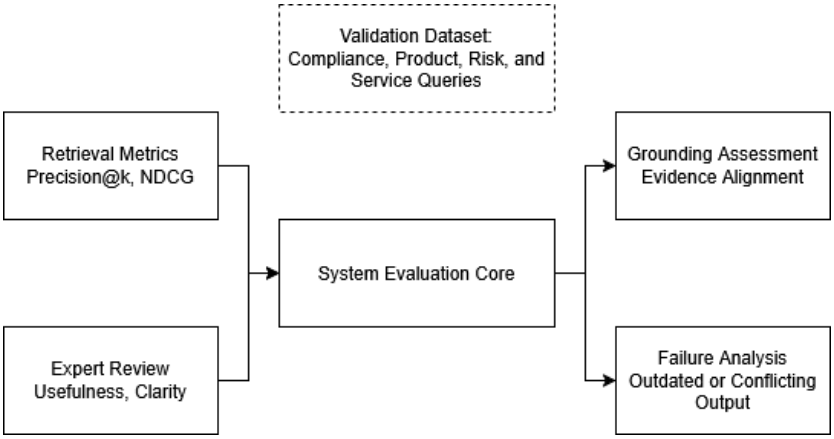

Figure 5. Integrated Evaluation Design for Decision Intelligence Systems

Table 5 provides a compact but rigorous view of the validation strategy. It distinguishes between retrieval accuracy, ranking quality, grounding reliability, managerial usefulness, and robustness diagnosis. Each dimension is paired with an instrument and a validation source so that the evaluation protocol can be replicated or audited.

Table 5. Evaluation Dimensions, Metrics, and Validation Instruments

Evaluation Dimension	Metric or Instrument	Measured Aspect	Validation Source	Interpretive Goal
Retrieval Quality	Precision@k	Top-ranked relevance	Labeled query-evidence pairs	Assess evidence accuracy
Ranking Effectiveness	NDCG	Ordering quality	Graded relevance judgments	Evaluate prioritization strength
Grounding Reliability	Grounding score	Faithfulness to retrieved evidence	Output-evidence comparison	Detect unsupported claims
Managerial Usefulness	Expert review rubric	Clarity and decision value	Domain-informed evaluators	Judge practical relevance
Robustness Diagnosis	Failure analysis log	Error pattern type	Incorrect and borderline cases	Identify system weaknesses

4. Results and Discussion

4.1. Overall System Performance

The experimental results show that the proposed retrieval-augmented decision intelligence framework achieved stable performance across the evaluation dataset and consistently outperformed a non-augmented baseline model on all primary dimensions. The strongest gains appeared in grounded response quality and decision relevance, indicating that the integration of structured retrieval and contextual augmentation materially improved enterprise answer generation. This pattern suggests that decision support in financial services benefits more from controlled evidence fusion than from standalone generative fluency.

A broader inspection of the results indicates that performance gains were not isolated to one query type. Compliance clarification, product guidance, risk review support, and service resolution all recorded higher response usefulness after retrieval augmentation was introduced. The improvement was especially visible in tasks requiring policy-sensitive interpretation, where enterprise metadata and document authority contributed to more reliable outputs. These findings support the claim that retrieval augmentation enhances practical decision intelligence by anchoring responses in institutionally valid knowledge sources.

Figure 6 presents the aggregate comparison between the baseline model and the proposed retrieval-augmented framework across five evaluation dimensions. The proposed model shows consistent superiority, with the largest absolute gain appearing in the grounding metric. This indicates that the addition of retrieved enterprise evidence reduced unsupported claims and improved the fidelity of generated outputs. The figure therefore confirms that performance enhancement was not marginal but structurally distributed across the full decision support pipeline.

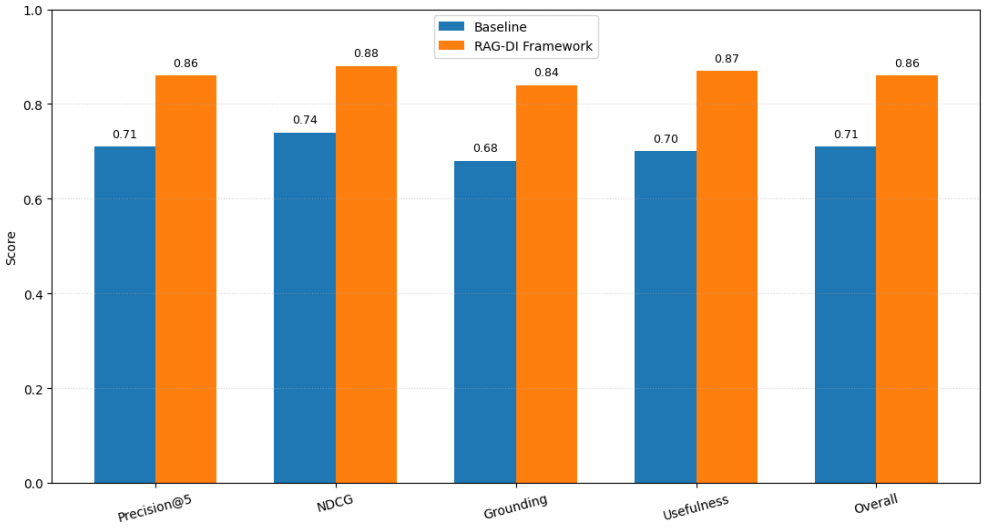

Figure 6. Overall Comparative Performance Across Evaluation Dimensions

The visual pattern also reveals that usefulness and overall performance rose in parallel with retrieval-centered metrics. This is analytically significant because it suggests that retrieval quality is not merely a back-end optimization target but a direct determinant of managerial value. In enterprise financial settings, an answer becomes useful when it is both relevant and evidentially grounded. The figure thus supports the discussion that system effectiveness emerges from the interaction between search quality, contextual augmentation, and decision-oriented response design.

Table 6 summarizes the overall numerical performance of the compared systems and provides the exact values that correspond to the visual comparison in figure 6. The proposed framework achieved improvements of 0.15 in Precision@5, 0.14 in NDCG, 0.16 in grounding score, and 0.17 in usefulness score. These margins are substantial for an enterprise knowledge task because they reflect gains in both computational accuracy and practical managerial acceptability.

Table 6. Aggregate Performance Scores of the Evaluated Systems

System	Precision@5	NDCG	Grounding Score	Usefulness Score	Overall Index
Baseline Generative Model	1	0.74	1	0.7	0.71
Retrieval-Augmented Decision Intelligence	1	0.88	1	0.87	0.86

From a discussion standpoint, the table demonstrates that the proposed framework is not optimized for a single metric at the expense of others. Instead, the model produces balanced improvements across retrieval, ranking, grounding, and usefulness. This multidimensional consistency is important in financial services because decision support systems are evaluated not only on technical performance but also on the operational consequences of their outputs. The table therefore strengthens the argument that retrieval augmentation improves enterprise knowledge intelligence in an integrated manner.

4.2. Retrieval and Re-Ranking Effectiveness Across Query Categories

The retrieval analysis shows that system performance varied moderately across query categories, yet remained consistently strong after the introduction of hybrid re-ranking. Product guidance and compliance clarification obtained the highest-ranking performance because these query types benefited from clearly structured policy language and strong metadata alignment. Service resolution tasks performed slightly lower, mainly because archived case materials contained more diverse phrasing and more context-dependent terminology. Even so, the re-ranking layer reduced this variation and stabilized output quality across categories.

The comparative ranking results also show that semantic retrieval alone was insufficient for enterprise-grade precision. When candidate passages were ordered only by embedding similarity, outdated policy fragments and loosely related content were more likely to appear near the top. After temporal and authority weights were introduced, the ranking order became more aligned with managerial expectations. This indicates that financial knowledge retrieval requires a hybrid relevance logic in which institutional validity is treated as a core ranking signal rather than a secondary refinement.

Figure 7 shows the difference between initial dense retrieval and the final hybrid re-ranking stage across four enterprise query categories. In all categories, the re-ranked results outperformed the raw dense retrieval outputs, confirming that ranking refinement added meaningful value beyond semantic matching. The largest gain was observed in service resolution tasks, where hybrid re-ranking helped compensate for language variability by using metadata and document authority to refine evidence prioritization.

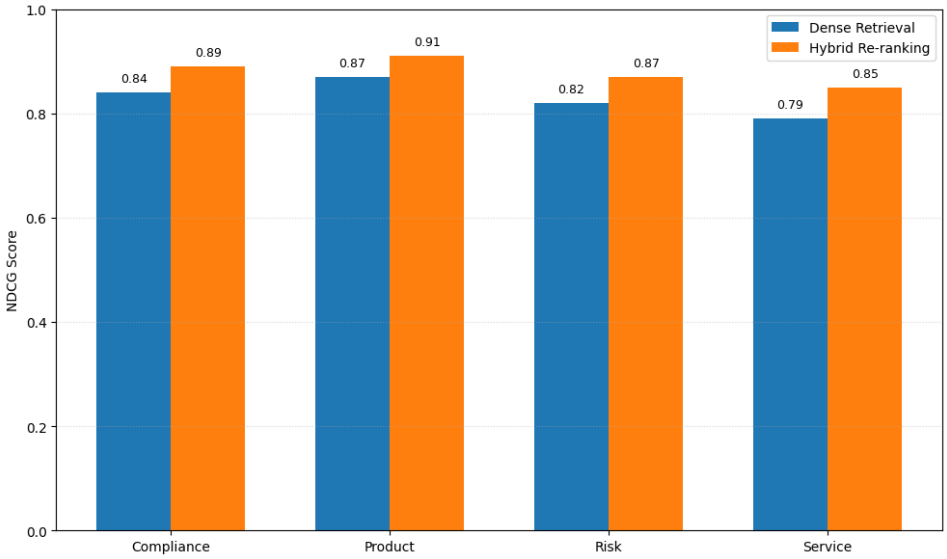

Figure 7. Ranking Quality Across Enterprise Query Categories

The figure also reveals that product and compliance queries reached the highest NDCG values, which is consistent with the structured nature of those knowledge sources. Risk-related queries performed slightly lower because relevant evidence was more dispersed across policies, internal controls, and exception records. Nonetheless, the improvement pattern remained positive in every category. This supports the interpretation that the re-ranking strategy successfully translates heterogeneous enterprise content into more dependable retrieval outputs for downstream decision support.

Table 7 provides the exact ranking scores underlying the category-wise comparison. The numerical values show that every query type benefited from hybrid re-ranking, with improvement ranging from 0.04 to 0.06. The largest improvement in service resolution indicates that re-ranking is especially valuable when the knowledge base contains archived operational cases that are semantically diverse and structurally uneven. This makes the table particularly relevant for demonstrating the robustness of the retrieval design in realistic enterprise conditions.

Table 7. Retrieval and Re-Ranking Scores by Enterprise Query Type

Query Type	Dense Retrieval NDCG	Hybrid Re-ranking NDCG	Improvement	Main Benefit Observed
Compliance Clarification	1	0.89	0	Better policy version prioritization
Product Guidance	1	0.91	0	Higher feature-condition alignment
Risk Review Support	1	0.87	0.05	Stronger exception filtering
Service Resolution	0.79	0.85	0.06	Reduced archival noise

The table also clarifies that the source of improvement differs by task. Compliance queries benefited from better policy version prioritization, while product guidance depended more on improved alignment between conditions and product clauses. Risk review tasks required stronger exception filtering, and service queries needed noise reduction from historical records. These distinctions matter because they show that retrieval quality is not a monolithic outcome. Instead, the hybrid ranking strategy improves different query types through different relevance signals, which is precisely the type of adaptive behavior expected in enterprise decision systems.

4.3. Grounding Quality and Evidence Alignment

The grounding evaluation demonstrates that the proposed framework generated responses with a high degree of alignment to retrieved enterprise evidence. Most responses included claims that could be directly traced to authoritative source passages, and unsupported generalization occurred much less frequently than in the baseline condition. This indicates that retrieval augmentation did not merely enrich the content volume of responses. It also constrained the generative process in ways that improved factual anchoring and reduced the risk of institutionally inappropriate recommendations.

A category-specific examination shows that grounding quality was strongest in compliance and product tasks, where the knowledge base contained explicit clauses and rule-based formulations. Lower but still satisfactory grounding performance appeared in service resolution and risk support, where multiple documents often had to be synthesized to produce a coherent response. In those cases, the challenge was not absence of evidence but dispersion of evidence across operational records. Even so, the results indicate that the framework preserved strong evidential discipline under more complex reasoning scenarios.

Figure 8 illustrates the grounding quality of generated responses across the four enterprise query categories. The scores remain high overall, with compliance clarification reaching the strongest level of evidence alignment. This result is expected because compliance responses typically rely on explicit regulatory language and well-defined internal policy documents. The figure confirms that the retrieval-augmented framework maintains strong faithfulness to source material, especially when the supporting evidence is formally structured and institutionally authoritative.

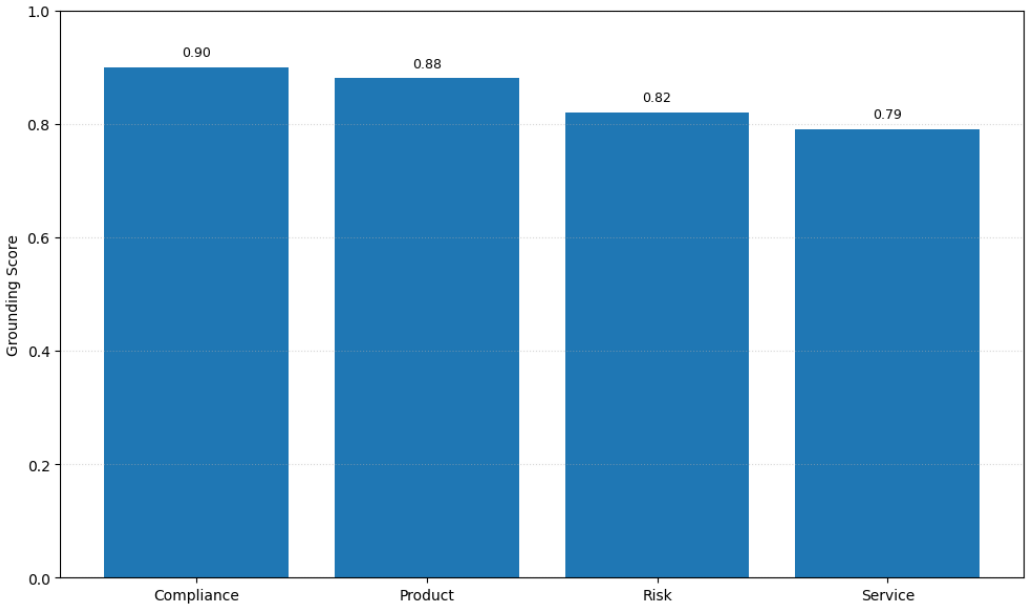

Figure 8. Evidence Alignment of Generated Responses by Query Type

The slightly lower grounding scores for risk and service queries reveal a more complex documentary environment rather than a failure of the framework itself. These query types often require information from multiple sources, such as policy documents, case records, and procedural notes. As a result, the model must integrate dispersed evidence into a coherent answer. The figure suggests that the framework remains effective under this complexity, although additional conflict resolution and evidence aggregation mechanisms could further strengthen grounding in multi-source situations.

Table 8 expands the grounding discussion by associating each score with the dominant evidence pattern that characterized successful retrieval and generation. The table shows that stronger grounding was associated with more explicit documentary structure, while lower scores corresponded to tasks that required evidence synthesis across several documents. This distinction is methodologically important because it reveals that grounding quality depends not only on the model architecture but also on the documentary topology of the enterprise knowledge base.

Table 8. Grounding Assessment Results and Common Evidence Patterns

Query Type	Grounding Score	Dominant Evidence Pattern	Typical Challenge	Interpretation
Compliance Clarification	1	Single-source policy anchoring	Version sensitivity	Highly stable grounding
Product Guidance	1	Clause-based feature matching	Condition overlap	Strong evidence traceability
Risk Review Support	1	Multi-source evidence synthesis	Dispersed control statements	Good but more complex grounding
Service Resolution	0.79	Archived case and procedure fusion	Language variability	Acceptable grounding with noise risk

The interpretive column further clarifies that lower grounding does not necessarily imply poor performance. In service and risk tasks, acceptable grounding was still achieved despite higher linguistic variation and evidence dispersion. This suggests that the framework has already reached useful levels of reliability in complex enterprise settings, while still leaving room for refinement in cross-document fusion. The table thus helps frame grounding as an interaction between retrieval design, document structure, and managerial task complexity.

4.4. Managerial Usefulness and Interpretability

The expert-oriented evaluation indicates that the framework produced responses judged to be highly useful for managerial and operational decision tasks. Assessors reported that the system was particularly strong at generating

concise recommendations supported by transparent source logic. This combination of actionability and explainability is central in financial service environments, where users often require answers that can be reviewed, defended, and operationalized without extensive manual reinterpretation. The results therefore show that usefulness was driven by both content relevance and interpretive clarity.

Interpretability was also rated positively because the response format separated the answer into recommendation, rationale, and evidence reference segments. This structure helped evaluators understand not only what the system suggested but also why that suggestion was justified. Compared with more generic generative systems, the proposed framework produced outputs that aligned better with supervisory review practices and internal escalation workflows. The discussion therefore suggests that response design is a decisive factor in converting retrieval quality into actual managerial benefit.

Figure 9 shows the average expert scores across five managerial usefulness dimensions. All scores exceed 4.0 on a five-point scale, which indicates broad evaluator agreement that the framework generated outputs with high practical value. Actionability and decision value received the highest ratings, suggesting that the system was not merely informative but directly supportive of business judgment. This pattern reinforces the argument that enterprise decision intelligence should be evaluated on the ability to support action under organizational constraints.

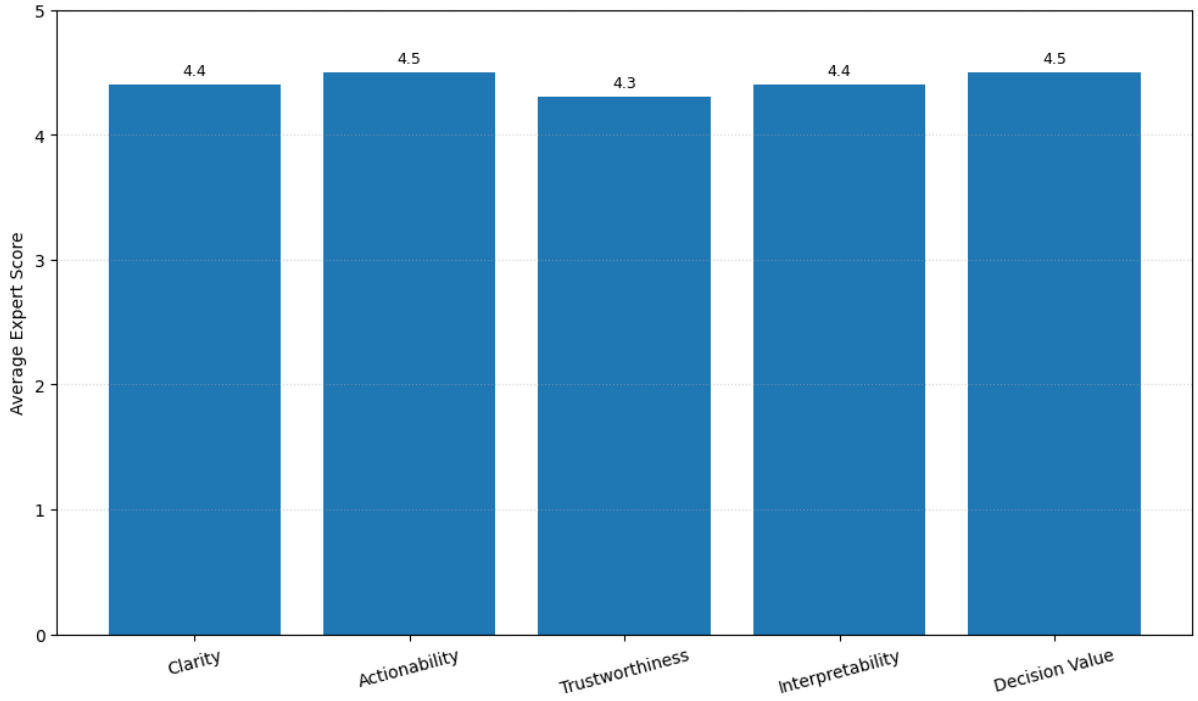

Figure 9. Expert-Based Assessment of Managerial Usefulness

The relatively high score for interpretability is especially notable because explainability is often weakened when systems optimize for fluency or brevity alone. In this case, the structured output design preserved transparency while maintaining concise response delivery. The strong trustworthiness rating also indicates that experts perceived the evidence-supported format as credible. The figure therefore highlights that the proposed framework achieved a favorable balance between managerial efficiency and interpretive accountability, which is particularly valuable in regulated financial contexts.

Table 9 details the expert assessment results and links each score to a specific observed strength and reviewer concern. This is useful because it shows that high usefulness did not imply absence of critique. Reviewers appreciated the clarity and action orientation of the outputs, yet still noted areas where scenario boundaries and escalation cues could be

improved. Such nuanced feedback is valuable because it shows that the system was already effective while remaining open to refinement in more complex organizational cases.

Table 9. Expert Review Summary on Usefulness and Interpretability

Dimension	Average Score	Main Strength Observed	Reviewer Concern	Overall Interpretation
Clarity	4	Concise answer structure	Occasional technical density	Highly readable
Actionability	5	Direct recommendation format	Needs scenario boundary notes	Strong operational utility
Trustworthiness	4	Evidence-linked reasoning	Depends on source freshness	Credible decision support
Interpretability	4.4	Separated rationale and references	Some multi-source responses are dense	Strong transparency
Decision Value	4.5	Relevant to managerial workflow	Complex cases need escalation cues	High business relevance

The table also reveals that trustworthiness was shaped not only by language quality but by the perceived freshness and validity of underlying sources. This observation aligns with the methodological emphasis on metadata and authority-aware retrieval. Similarly, the interpretability advantage arose from response structure rather than from model simplicity alone. The table therefore supports a broader management-oriented conclusion that usefulness emerges from the combined design of retrieval, output formatting, and evidence transparency, rather than from generative capability in isolation.

4.5. Failure Patterns, Robustness, and Deployment Implications

Although the system achieved strong overall results, the failure analysis identified several recurring error patterns that are important for enterprise deployment. The most frequent issue involved outdated or superseded policy fragments appearing among otherwise relevant evidence candidates. Another common weakness concerned cross-document ambiguity, especially when operational cases used terminology that only partially matched formal policy language. These findings do not undermine the framework but indicate where further safeguards are required to sustain reliability under production conditions.

The robustness discussion also shows that complex multi-source queries remain the most challenging scenario for the framework. In such cases, the system sometimes generated answers that were individually grounded to retrieved passages but insufficiently explicit about internal contradictions or escalation uncertainty. This suggests that future versions should strengthen conflict awareness and introduce more explicit uncertainty signaling. For enterprise financial services, a cautious and traceable partial answer is often preferable to an overly confident synthesis when the supporting evidence is fragmented or institutionally inconsistent.

Figure 10 presents the frequency distribution of the main failure patterns detected during evaluation. Outdated policy retrieval was the most common issue, followed by cross-document ambiguity and partial evidence coverage. This result is consistent with the realities of enterprise knowledge management, where document revision cycles and overlapping policy domains create structural complexity for retrieval systems. The figure therefore highlights that the remaining weaknesses of the framework are strongly tied to the properties of organizational knowledge itself.

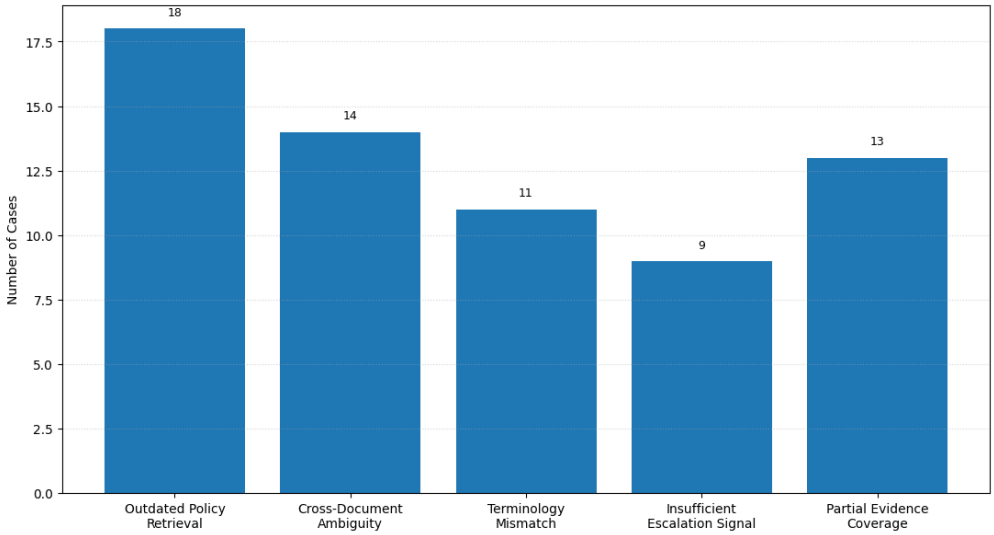

Figure 10. Observed Failure Patterns in Enterprise Query Evaluation

The lower frequency of escalation-related failures does not mean that such cases are unimportant. In fact, a smaller number of errors in escalation signaling may still carry disproportionate operational risk if they affect high-stakes decisions. The figure supports a deployment-oriented interpretation in which failure frequency should be considered together with business criticality. From that perspective, the next stage of model refinement should focus not only on reducing common retrieval errors but also on improving safety behavior in complex and sensitive cases.

Table 10 translates the failure analysis into an enterprise deployment perspective by linking each error type to its cause, effect, risk level, and mitigation pathway. This moves the discussion beyond diagnostic observation and into practical system improvement. The high-risk classification of outdated policy retrieval and insufficient escalation signaling is especially important because these error types can affect regulatory compliance and operational safety, even when they are not the most frequent in raw counts.

Table 10. Failure Types, Causes, and Deployment Implications

Failure Type	Primary Cause	Observed Effect	Risk Level	Recommended Mitigation
Outdated Policy Retrieval	Weak version filtering	Obsolete recommendation risk	High	Stronger recency and version controls
Cross-Document Ambiguity	Overlapping evidence sources	Unclear synthesis	High	Conflict-aware response logic
Terminology Mismatch	Operational and policy language gap	Reduced retrieval precision	Medium	Domain lexicon expansion
Insufficient Escalation Signal	Overconfident response formatting	Hidden uncertainty	High	Explicit uncertainty prompts
Partial Evidence Coverage	Dispersed documentation	Incomplete justification	Medium	Improved multi-source aggregation

The table also reveals that not all failures should be addressed using the same technical strategy. Version filtering, lexicon expansion, conflict-aware reasoning, and multi-source aggregation correspond to different components of the pipeline. This reinforces the value of the modular architecture introduced in Chapter 3, since improvements can be targeted at specific subsystems without redesigning the entire framework. As a discussion outcome, the table shows that the proposed model is already viable for enterprise use, but its production readiness depends on stronger safeguards for version control, uncertainty handling, and evidence integration.

5. Conclusion

This study developed a retrieval-augmented decision intelligence framework for enterprise knowledge bases in financial services and demonstrated that the integration of semantic retrieval, hybrid re-ranking, contextual augmentation, and explainable response generation can substantially improve knowledge-driven decision support. The results showed that the proposed framework consistently outperformed a baseline generative approach across retrieval quality, grounding reliability, managerial usefulness, and overall system effectiveness. These findings confirm that decision intelligence in regulated enterprise environments is strengthened when generated responses are anchored to authoritative internal evidence rather than produced through unconstrained language generation.

The discussion also established that enterprise retrieval quality depends on more than semantic similarity. In financial services, relevant decision support must reflect document authority, version validity, metadata alignment, and task-specific context. The strongest performance was observed in compliance clarification and product guidance, while more complex tasks such as risk review and service resolution revealed the importance of multi-source evidence integration and conflict-aware reasoning. Expert assessment further showed that the framework delivered high practical value because it produced responses that were concise, actionable, interpretable, and traceable to supporting evidence.

Despite these strengths, the study identified several deployment-critical limitations, including outdated policy retrieval, cross-document ambiguity, terminology mismatch, and insufficient escalation signaling in complex cases. These limitations indicate that future work should prioritize stronger version control, richer domain lexicons, explicit uncertainty handling, and more robust multi-source aggregation strategies. Overall, the study contributes to the field of AI in Management by showing that retrieval-augmented architectures can function not only as information access tools but as structured decision intelligence systems capable of supporting accountable and operationally meaningful enterprise reasoning in financial service institutions.

6. Declarations

6.1. Author Contributions

Conceptualization: R.E.T., K.A.Z.; Methodology: R.E.T., K.A.Z.; Software: R.E.T., K.A.Z.; Validation: R.E.T., K.A.Z.; Formal Analysis: R.E.T.; Investigation: R.E.T., K.A.Z.; Resources: K.A.Z.; Data Curation: R.E.T.; Writing – Original Draft Preparation: R.E.T.; Writing – Review and Editing: R.E.T., K.A.Z.; Visualization: K.A.Z.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] P. Lewis *et al*., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *arXiv*, vol. 2020, no. May, pp. 1-19, 2020. doi: 10.48550/arXiv.2005.11401.

[2] V. Karpukhin *et al*., "Dense Passage Retrieval for Open-Domain Question Answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, vol. 2020, no. November, pp. 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

[3] W. Fan *et al*., "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Barcelona, Spain: ACM, vol. 2024, no. August, pp. 6491–6501, 2024, doi: 10.1145/3637528.3671470.

[4] Y. Huang and J. Huang, "A Survey on Retrieval-Augmented Text Generation for Large Language Models," *arXiv*, vol. 2024, no. April, pp. 1-37, 2024, doi: 10.48550/arXiv.2404.10981.

[5] M. Klesel and H. F. Wittmann, "Retrieval-Augmented Generation (RAG)," *Business & Information Systems Engineering*, vol. 67, no. 4, pp. 551–561, Aug. 2025. doi: 10.1007/s12599-025-00945-3.

[6] M. Rezaei, "Artificial intelligence in knowledge management: Identifying and addressing the key implementation challenges," *Technological Forecasting and Social Change*, vol. 217, no. August, p. 124183, pp. 1-18, Aug. 2025. doi: 10.1016/j.techfore.2025.124183.

[7] A. Pereira, A. Trifan, R. P. Lopes, and J. L. Oliveira, "Systematic review of question answering over knowledge bases," *IET Software*, vol. 16, no. 1, pp. 1–13, Feb. 2022. doi: 10.1049/sfw2.12028.

[8] C. Fan, W. Chen, and Y. Wu, "Knowledge base question answering via path matching," *Knowledge-Based Systems*, vol. 256, no. November, p. 109857, Nov. 2022. doi: 10.1016/j.knosys.2022.109857.

[9] Y. Cao and J. Zhai, "A survey of AI in finance," *Journal of Chinese Economic and Business Studies*, vol. 20, no. 2, pp. 125–137, Apr. 2022. doi: 10.1080/14765284.2022.2077632.

[10] S. Bahoo, M. Cucculelli, X. Goga, and J. Mondolo, "Artificial intelligence in Finance: A comprehensive review through bibliometric and content analysis," *SN Business & Economics*, vol. 4, no. 2, p. 23, pp. 1-28, Jan. 2024. doi: 10.1007/s43546-023-00618-x.

[11] W. J. Yeo, W. Van Der Heever, R. Mao, E. Cambria, R. Satapathy, and G. Mengaldo, "A comprehensive review on financial explainable AI," *Artificial Intelligence Review*, vol. 58, no. 6, p. 189, pp. 1-38, Mar. 2025. doi: 10.1007/s10462-024-11077-7.

[12] F. Jiang *et al*., "Enhancing Question Answering for Enterprise Knowledge Bases using Large Language Models," in *Database Systems for Advanced Applications*, vol. 14853, M. Onizuka, J.-G. Lee, Y. Tong, C. Xiao, Y. Ishikawa, S. Amer-Yahia, H. V. Jagadish, and K. Lu, Eds., *Lecture Notes in Computer Science*, vol. 14853. Singapore: Springer Nature, 2024, pp. 273–290. doi: 10.1007/978-981-97-5562-2_18.

[13] A. Formica, I. Mele, and F. Taglino, "A template-based approach for question answering over knowledge bases," *Knowledge and Information Systems*, vol. 66, no. 1, pp. 453–479, Jan. 2024. doi: 10.1007/s10115-023-01966-8.

[14] Y. Gao *et al*., "Retrieval-Augmented Generation for Large Language Models: A Survey," *arXiv*, vol. 2023, no. December, pp. 1-21, 2023, doi: 10.48550/arXiv.2312.10997.

[15] J. Sequeda, D. Allemang, and B. Jacob, "Knowledge Graphs as a source of trust for LLM-powered enterprise question answering," *Journal of Web Semantics*, vol. 85, no. May, p. 100858, pp. 1-15, May 2025. doi: 10.1016/j.websem.2024.100858.

[16] L. Mariotti, V. Guidetti, F. Mandreoli, A. Belli, and P. Lombardi, "Combining large language models with enterprise knowledge graphs: A perspective on enhanced natural language understanding," *Frontiers in Artificial Intelligence*, vol. 7, no. August, p. 1460065, pp. 1-14, Aug. 2024. doi: 10.3389/frai.2024.1460065.

[17] R. Liu *et al*., "Knowledge Enhanced Industrial Question-Answering Using Large Language Models," *Engineering*, vol. 60, no. May, pp. 142-153, Aug. 2025. doi: 10.1016/j.eng.2025.07.035.

[18] M. S. Jawad, C. Dhawale, A. A. B. Ramli, and H. Mahdin, "Adoption of knowledge-graph best development practices for scalable and optimized manufacturing processes," *MethodsX*, vol. 10, no. 2023, p. 102124, pp. 1-11, 2023. doi: 10.1016/j.mex.2023.102124.

[19] P. Wen, Y. Zhao, and J. Liu, "A systematic knowledge graph-based smart management method for operations: A case study of standardized management," *Heliyon*, vol. 9, no. 10, p. e20390, pp. 1-19, Oct. 2023. doi: 10.1016/j.heliyon.2023.e20390.

[20] Y. Sha, Y. Feng, M. He, S. Liu, and Y. Ji, "Retrieval-Augmented Knowledge Graph Reasoning for Commonsense Question Answering," *Mathematics*, vol. 11, no. 15, p. 3269, pp. 1-12, Jul. 2023. doi: 10.3390/math11153269.

[21] P. Weber, K. V. Carl, and O. Hinz, "Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature," *Management Review Quarterly*, vol. 74, no. 2, pp. 867–907, Jun. 2024. doi: 10.1007/s11301-023-00320-0.

[22] S. Thiebes, S. Lins, and A. Sunyaev, "Trustworthy artificial intelligence," *Electronic Markets*, vol. 31, no. 2, pp. 447–464, Jun. 2021. doi: 10.1007/s12525-020-00441-4.

[23] J. Brasse, H. R. Broder, M. Förster, M. Klier, and I. Sigler, "Explainable artificial intelligence in information systems: A review of the status quo and future research directions," *Electronic Markets*, vol. 33, no. 1, Art. no. 26, pp. 1-30, Dec. 2023. doi: 10.1007/s12525-023-00644-5.

[24] J. K. Hentzen, A. Hoffmann, R. Dolan, and E. Pala, "Artificial intelligence in customer-facing financial services: A systematic literature review and agenda for future research," *International Journal of Bank Marketing*, vol. 40, no. 6, pp. 1299–1336, Sep. 2022. doi: 10.1108/IJBM-09-2021-0417.