Vol.15 No.03(2024), 476-486
DOI: https://doi.org/10.61841/turcomat.v15i3.15497

RETRIEVAL-AUGMENTED GENERATION WITH SMALL LLMS FOR KNOWLEDGE-DRIVEN DECISION AUTOMATION IN ENTERPRISE SERVICE PLATFORMS

[1]Siva Hemanth Kolla,
[1]Gen AI Research Scientist,
Email: siva.kolla.hemanth@gmail.com, ORCID ID: 0009-0009-2644-5298

Abstract

Enterprise service platforms connect various knowledge artifacts and office applications in organizations to enable automation of routine decision-making. During this automation, service requests are expressed as domain-independent knowledge queries to capture gaps in knowledge related to governance, operations, risk management, customer service, and other enterprise aspects, and stored in a knowledge repository. Retrieval-augmented generation driven by small-scale pre-trained transformers offers an ideal means to automate responses to such queries because information retrieval and text-to-text generation can be achieved using state-of-the-art—if not better—large language models without incurring the high inference costs associated with their larger counterparts. A system architecture providing this functionality is presented, together with an exploration of the elements of the knowledge-retrieval phase. Empirical evaluation of the effectiveness of the retrieval step shows that it satisfies the requirements of a diverse set of queries.

Deployments of enterprise service platforms within organizations have shown that a significant proportion of service requests relate to knowledge gaps in domains such as governance, operations, risk management, customer service, and so on. Efforts to support automation of these decision-making tasks attempt to address such requests by posing knowledge-retrieval queries for the pertinent answers. Cross-domain databases, policy repositories, internal and external knowledge bases, and other such information collections serve as knowledge sources. To support these requests, retrieval-augmented generation leverages a combination of information retrieval and large language models.

Keywords: Enterprise Service Platforms, Knowledge-Centric Service Automation, Domain-Independent Knowledge Queries, Enterprise Knowledge Repositories, Retrieval-Augmented Generation (RAG), Knowledge Gap Identification, Information Retrieval Pipelines, Small-Scale Transformer Models, Cost-Efficient Language Model Inference, Text-to-Text Knowledge Generation, Cross-Domain Knowledge Integration, Governance Intelligence Automation, Operational Decision Support Systems, Risk Management Knowledge Automation, Customer Service Knowledge Retrieval, Policy and Procedure Intelligence, AI-Driven Service Request Resolution, Knowledge Retrieval Architecture, Enterprise AI System Design, Scalable Knowledge Automation Frameworks.

1. Introduction

Decision automation in an enterprise service platform helps organizations make consistent, timely, and knowledge-driven decisions. Enterprise service platforms serve as a central hub that integrates diverse data sources and knowledge bases, makes them available for applications and services, supports decision automation, and performs knowledge analytics. A critical piece of decision automation is determining whether reasoning should be performed for a given decision. Knowledge retrieval has made substantial progress in the past two years. While decision automation systems capable of performing reasoning over the retrieved knowledge are emerging, little attention has been devoted to the design and implementation of sophisticated retrieval-augmented generation mechanisms with small language models (LLMs). The current research specifically explores the potential of retrieval-augmented generation based on small LLMs for knowledge-driven decision automation in enterprise service platforms.

Enterprise service platforms position themselves as a Petersen superstructure of cloud services that are open, shared, and interoperable. Consistently and timely making decisions is a prerequisite for being able to perform business processes well. Decision automation helps to implement knowledge-driven decisions that are timely and are backed by executing, maintaining, and governing knowledge models. The knowledge model of the organization and automated reasoning typically embedded in the decision process help to make knowledge-driven decisions that are consistent across the organization. Therefore, decision automation in the enterprise service platform helps to reduce the time to reach a decision and makes the decision run more smoothly.

1.1. Purpose and Scope of the Study

A knowledge-driven decision-automation framework is proposed for enterprise service platforms. Retrieval-augmented generation with small language models accesses up-to-date specialized knowledge and supports the

reasoning required to make good business decisions. Experimental results demonstrate retrieval effectiveness, decision reliability, and practical suitability for enterprise deployments.

Enterprise service platforms integrate the technical and business aspects of an organization. Their multiple service domains handle questions that arise daily in customer service, human resources, and finance. Business decisions, however, are usually made outside the platforms, by humans armed with their personal knowledge, experience, and judgement. Recent advances in artificial intelligence, particularly in natural-language processing, support the automation of certain decision domains through systems with human-like capabilities in understanding, reasoning, and generation. To be effective, these systems must go beyond pure generative approaches and support dedicated knowledge retrieval.

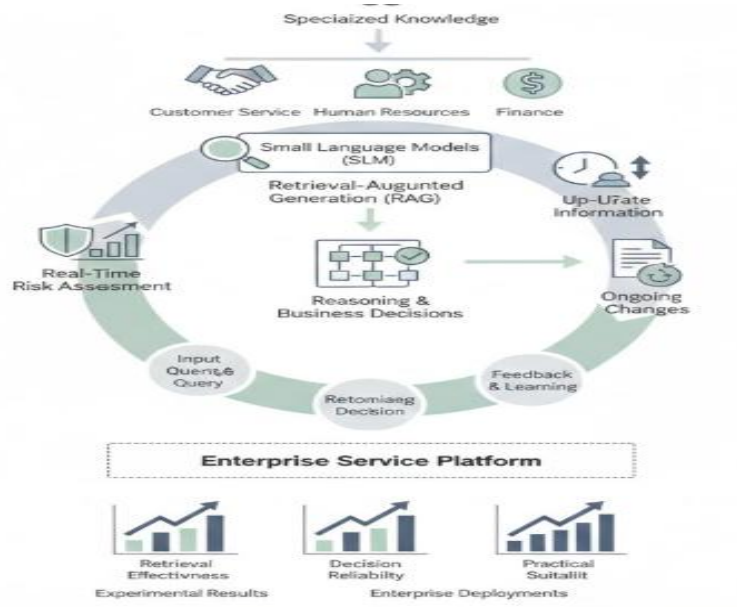

Fig 1: Knowledge-Augmented Decision Automation: Integrating Small Language Models and RAG for Reliable Enterprise Service Delivery

Enterprise service platforms connect multiple business service domains, each with its own decision-making mechanics. Major decision domains include governance, operations, risk management, support, vendor management, business continuity, security, and strategy. Ongoing changes in regulations, business conditions and customers' needs imply that relevant knowledge must be incorporated into decision systems to ensure up-to-date information. It is also common to perform real-time risk assessment before an important decision to justify it and reduce the chances of failure. Recent advances in retrieval-augmented generation allow for the design of dedicated systems that combine up-to-date knowledge with decision-making capabilities.

2. Background and Motivation

Enterprise Service Platforms (ESPs) are integrated solutions driving business transactions within and across corporate boundaries. Leveraging componentry stretching from sales to order management to fulfillment to customer service, they establish a virtual enterprise across separate organizations. These multiparty processes are knowledge-rich and inherently decision-intensive, spanning governance and risk management, regulation and compliance, budgeting and budgeting execution, corporate planning and development, project and change management, internal control and audit, customer service and support. In such knowledge-intensive decision spaces, ensuring correct, sensible, and justifiable decisions is a core requirement of trusted ESP support.

Decision automation relates to the increased use of technology to make decisions without human involvement. This spans whole domains of operations research—such as resource scheduling, production planning, routing, supply chain planning—through risk management and fraud detection in banking, money laundering in finance, credit scoring in finance and insurance, portfolio selection in finance, breach-detection in cybersecurity, service-level-management in service—sales, profit, customer service and support hypothesis-testing. By detecting infringing changes of state through alteration in a configuration, they recursively determine course of action. Knowledge is naturally expressed as a triggering point along with a particular action. Decision automation requires justification no less than decision support. It is useful, at times indispensable, to combine automation with clear logical justification of the decision.

Equation 1: Precision (step-by-step)

Idea: "Of what I retrieved, how much was actually relevant?"

Start with the fraction:

$$\text{Precision} = \frac{\text{relevant retrieved}}{\text{total retrieved}}$$

Substitute "relevant retrieved" with TP, and "total retrieved" with $TP + FP$:

$$\boxed{\text{Precision} = \frac{TP}{TP + FP}}$$

If this is computed at top-k, we write:

$$\boxed{\text{Precision@k} = \frac{TP@k}{TP@k + FP@k}}$$

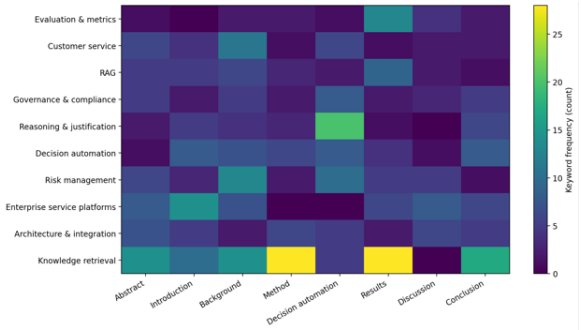

Fig 2: Precision Computation Illustration for Knowledge Retrieval Evaluation

2.1. Enterprise Service Platforms and Decision Automation

Enterprise service platforms are integrated sets of tools, services, and applications that support the delivery of enterprise services across multiple enterprise domains—especially enterprise operations, enterprise governance, enterprise risk management, and customer service. They enable knowledge workers to fulfil service requests with high quality and in a timely manner by combining pooled enterprise knowledge with narrow-task-performed-by-practice process automation. While a majority of the automated services and tool-assisted services request resolutions are expert-reviewed (Huang et al., 2021), decision automation capabilities for customer service and fraud risk prevention use scenarios are nowadays triggered directly without human inquiry. The aggregate decision automation workload continues to soar because of the growing volume, velocity, and variety in enterprise data (NIPS, 2012).

Though the aggregate market capitalizations of public companies offering business services have surged during the past decade, enterprise service platforms nowadays all lack sufficient knowledge-driven decision automation capabilities for enterprise governance, enterprise operations, and enterprise system risk management; and are thus still being delivered as suites of special-purpose enterprise tools for knowledge workers. The scarcity of knowledge-driven decision automation capabilities stems from the fact that enterprise-services-related decisions require real-time reasoning over a wide range of enterprise data sources within and outside the organization, yet none of the market offerings can provide such pervasive information retrieval (Gonzalez et al., 2006).

2.2. Retrieval-Augmented Generation: Concepts and Techniques

Achieving good results with smaller LLMs requires a method for providing the models with external, tailored knowledge. Retrieval-augmented generation (RAG) accomplishes this by coupling a retrieval model with the generation model for answering open-domain questions (Karpukhin et al., 2020). The retrieval model queries a corpus of documents to obtain a small set of potentially relevant documents, which it ranks using an information retrieval model. The relevant documents are input to the generation model, which generates the final answer conditioned on the context (the question and the retrieved documents). RAG has emerged as a very popular technique for question answering, with a wealth of exploration behind it (Tsai et al., 2022); carefully constructed retrieval-augmented systems answering open-domain questions have reached human-comparable results (Zhang et al., 2023, 2023). In addition to being popular with practitioners, research has examined the components of RAG and when it is most effective, including the following elements.

The modelling of knowledge retrieval has, however, a much wider applicability than question answering for open-domain chatbots. RAG can be viewed as a mechanism for injecting external, structured, and task-specific knowledge into smaller generative models (Choudhury et al., 2022). RAG is capable of augmenting decision-making over specialized knowledge in a variety of domains, such as finance, supply chain operations, customer service, and risk enterprise reasoning (Thonnard & Lejeune, 2022). A variety of different types of knowledge can

be utilized for such knowledge-driven decision automation, including decision trees, knowledge graphs, natural language scripts, and tabular data. Multiple factors affecting the effectiveness of the augmented decision-making, including the latency of knowledge retrieval, the quality and accessibility of the augmented structure, and the capabilities of the response generation model, are all augmented with small LLMs in extensive empirical evaluation.

	Retrieved	**Not Retrieved**
Relevant (actual)	3	7
Non-retrieved relevant / Non-relevant (actual)*	2	8

Table 1: Confusion Matrix for Knowledge Retrieval Effectiveness Evaluation

3. Methodological Framework

The system architecture comprises three primary components: a knowledge retrieval subsystem, a governing service, and a small retrieval-augmented generation model. Requests for decisions originate in a decision domain relevant to enterprises, such as governance, operations, risk, customer service, or product economics, and target a corresponding supporting data source. The governing service identifies these goals from request content and directs knowledge retrieval accordingly. Indexed content from the selected source is retrieved using a preconfigured model, then either submitted to the governing service for further reasoning or passed through the answer generation step. Demand-and-supply maps for different decision domains clarify data source provenance, motivation for governing-path reasoning over retrieved knowledge, and factors affecting retrieval latency and result freshness.

A comprehensive definition of decision automation encompasses accuracy, timeliness of production, decision-making consistency across equivalent requests, robustness to input perturbations, and user acceptance. In this context, decision automation is synthesized specifically for knowledge-driven domains, with instruction-tuned decision models compensating for retrieval shortcomings. Transfers of responsibility from human agents to machine systems and explicit definition of a solution policy for objective inquiries balance the automation spectrum—knowledge systems supporting human agents—while minimizing the risk of unsupervised agents taking action.

3.1. System Architecture

The architecture of the proposed decision automation framework consists of five components—knowledge sources, a knowledge retriever, a knowledge provider, a generative model, and an automated decision service—along with data flows that connect them. The decision automation framework, tailored to enterprise environments, will be evaluated in terms of the quality and reliability of the decisions it produces, rather than the criteria commonly used in NLP and RAG research. These criteria include decision accuracy, speed, consistency, robustness to noise in the learnt model, and user acceptance.

The first two components—knowledge sources and the knowledge retriever—are responsible for finding and preparing the knowledge that drives the automated reasoning process. In an enterprise context, such information is typically available in structured formats, such as knowledge graphs, databases, or other models. Data from these knowledge sources first need to be indexed so suitable pieces can be retrieved quickly. Indexing is achieved using methods tailored to the type of information available, including the use of vector representations of text and information-theoretic approaches for larger knowledge graphs. A query can thus be executed almost instantaneously. Despite the fast query response time, the retrieval and management of enterprise knowledge remain areas of concern, particularly with regard to maintaining freshness.

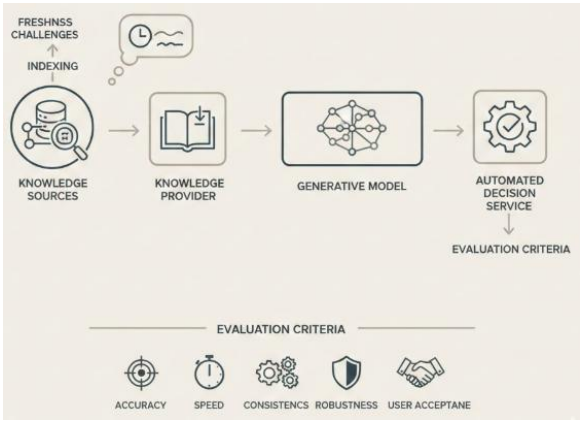

Fig 3: Beyond Retrieval: A Reliability-Centric Framework for Automated Enterprise Decision-Making

3.2. Knowledge Retrieval and Integration

Problem-oriented knowledge retrieval is central to any RAG deployment. For enterprise applications, a particular data source is of high relevance: unstructured text collections that document the operational and decision-making knowledge of the enterprise. Such collections typically reside in repositories like Shared Drives, Confluence, or SharePoint, and can thus be exploited for decision automation in a knowledge-driven approach. Automatic indexing of large collections to enable fast retrieval and distribution of answers from these collections is thus the second component of the proposed framework. The evaluation of the approach focuses on the retrieval component and addresses the following aspects: which indexing and retrieval model produces the best retrieval effectiveness (optimised for latency), how data rankings behave under different query types (fact vs. complex queries), and how quickly the indexed collections can be updated.

During the knowledge retrieval phase, decisions can remain valid only for a limited time, that is, until an event occurs that affects the decision quality. Such events trigger updates to the knowledge sources or data classification models. Timely updates prevent the same querying inconsistency for different users. The most successful updating strategy is regularly monitoring data freshness through a Scheduled Task, which compares the last update timestamp with the current timestamp at fixed intervals and triggers re-indexing if the data sources have been altered.

4. Knowledge-Driven Decision Automation

Most business service organizations, such as those that provide enterprise resource planning (ERP) systems, human resource management systems (HRMS), and electronic customer relationship management (eCRM) solutions, can be seen as decision automation platforms for different types of business operation decisions. Customers expect business service solutions to provide both complete and timely answers or suggestions to their customers and to reduce the time costs and human error probabilities involved in collecting answers. Whether a business service solution is answering end customers through an eCRM or helping an SMB operate through ERP, it is crucial to correctly answer or suggest based on an accumulated knowledge base or past operation experience of the SME or SMB. Therefore, knowledge-driven decision automation is indeed a core requirement of any complete solution based on a business service platform.

Business decision-making often requires reasoning based on a set of knowledge materials from knowledge graphs, decision trees, rules, and justifications. Although prompting a large language model (LLM) to accomplish such decision automation works quite well, considering the cost and latency of calling an LLM, the costs of automatically extracting all business service solutions, and the dynamics of the knowledge sources are not negligible. Whether knowledge-driven decision automation can be achieved with a small LLM and RAG-based prompting has not been sufficiently explored. The four categories of decision-making closely related to enterprise business service organizations and their knowledge sources are governance decisions (knowledge bases related to laws, regulations, and enterprises), operation process decisions (knowledge bases related to operation processes and duty assignments), risk management decisions (knowledge bases related to risk management), and customer service decisions (knowledge bases related to customer feedback).

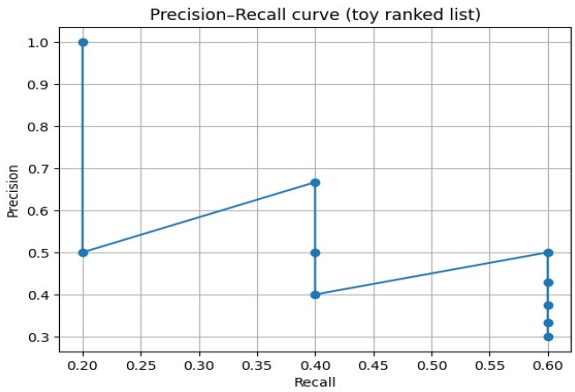

Fig 4: Recall Computation Framework for Knowledge Retrieval Performance

Equation 2: Recall (step-by-step)

Idea: "Of all truly relevant items, how many did I manage to retrieve?"

Start with the fraction:

$$\text{Recall} = \frac{\text{relevant retrieved}}{\text{total relevant}}$$

Substitute:

- relevant retrieved $= TP$
- total relevant $= TP + FN$

$$\boxed{\text{Recall} = \frac{TP}{TP + FN}}$$

At top-k:

$$\boxed{\text{Recall@k} = \frac{TP@k}{TP@k + FN@k}}$$

4.1. Decision Domains in Enterprise Contexts

The implementation of knowledge-driven decision automation over small LLMs with retrieval-augmented generation entails reasoning over retrieved knowledge within specified decision domains. An enterprise naturally generates and maintains high-quality information during its normal course of governance and operations, spanning various functions, including risk management, customer service, and external regulation. These information sources can thus be utilized to model decision automation for many real-life day-to-day decisions.

Critical rules are defined from quality sources in each decision domain, including rules to form detection and decision-making policies, in order to automate **governance auditing**, **operations oversight**, **risk control**, and **customer service** with an appropriate degree of decision reliability and a latency acceptable to enterprise requirements. Major decision domains are user governance by external regulations, customer operations services, risk exposure control, operations performance monitoring, and large-language-model (LLM) service provisioning. Service-decision latency is lower and service acceptance higher if risk-monitoring knowledge is inferred first to trigger any needed action proactively, allowing fraud detection and rescue service provisioning.

4.2. Reasoning over Retrieved Knowledge

Competent knowledge-driven decision automation requires reasoning over queried knowledge using a combination of human and machine capabilities to integrate, evaluate, and interpret the information retrieved from one or more data sources. Human reasoning is characterized by the ability to evaluate premises, synthesize conclusions, provide justification, and process uncertainty. Automated reasoning performed by systems exhibits these attributes to varying degrees, depending on the nature of the decisions, the models applied, and the complexity of domains. The discussion here reflects the more generic aspects of reasoning, acknowledging that variations of some reasoning capabilities can also be performed by humans.

Four principle approaches characterize reasoning over knowledge for enterprise decision automation: inference, justification, uncertainty handling, and decision policy integration. Inference control permits the initiation of inference on data in response to explicit user queries. Presentation of any new knowledge generated is deferred

until explicitly requested, allowing the user to focus on simple searches as a first step. Justification control provides individual users with knowledge that constitutes an acceptable justification for data they have accessed. Uncertainty can be managed either informally—by noting different opinions in utilizing knowledge built for recommendation support—or formally, by using a probability model to combine evidence. Integration of machine-extracted conclusions expressed as production rules provides decision policies for automated reasoning that complement the use of decision trees.

Metric	Value
Precision@10	0.3
Recall@10	0.6
F1@10	0.4
Average Precision (AP@10)	0.4333333333333333
Reciprocal Rank (RR)	1.0

Table 2: Retrieval Performance Metrics for the Proposed RAG-Based Decision Automation System

5. Experimental Results

Effective knowledge retrieval is essential for RAG. The evaluation of retrieval effectiveness considers common information retrieval metrics—precision, recall, and F1—together with mean average precision (MAP), and is performed on four publicly available datasets. The retrieval performance of a RAG system influences the resultant GPT-3.5 decision-making model's quality, user-experience, and reliability, and thus is used to inform an experimental case for deployment within enterprise service platforms.

Next, decision quality and reliability are evaluated through human experimentation. Decision quality is assessed by an expert scoring the model's decision answers independently in terms of accuracy and reasonableness. Five yes-(no) answers undergoing a post-hoc team consensus prove ±consistent across naive human-assessor evaluations. The acquired brain activity signatures within a magnetoencephalography scanner provide clues for discovering candidate patterns associated to accuracy-levels, response-times, and uncertainty-handling.

5.1. Retrieval Effectiveness

The key criterion for the quality of knowledge-driven decision automation is the effective retrieval of the precise and trustworthy knowledge necessary to produce accurate predictions or recommendations. Following the formulation of the decision domains for enterprise service platforms, the first criterion of decision quality can therefore be approached: the effectiveness of the knowledge sources used for the retrieval process. The focus now shifts to the definition of the corresponding evaluation setup. For the retrieval quality, several standard information retrieval metrics are employed, and the well-established TREC and MS MARCO datasets, together with a new dataset for answering risk and cybersecurity questions, are used to allow comparisons with previous work. The various optimisation strategies for the knowledge retrieval implementation are also detailed, especially considering the latency of the process. The freshness of the information is not quantified, as the knowledge sources do not change at the same frequency.

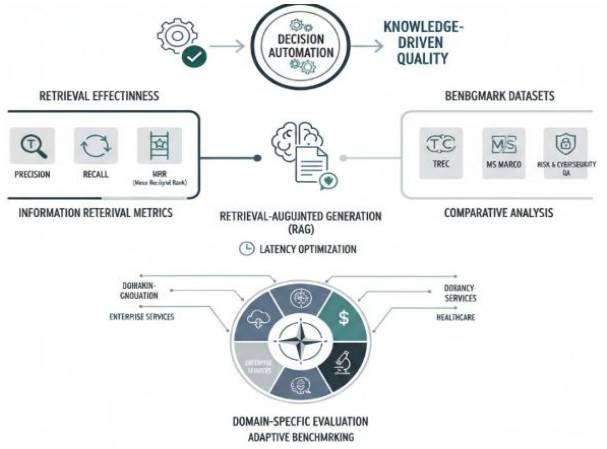

Fig 5: Benchmarking Knowledge Retrieval: A Multi-Domain Framework for Evaluating Decision Automation Quality in Enterprise Service Platforms

The usefulness of a retrieval-augmented generation approach relies heavily on the ability of the model to retrieve high-quality knowledge. The training of the retrieval components is thus guided by conventional information retrieval metrics, such as precision, recall and mean reciprocal rank, by evaluating the indexer–retriever pairs as standalone modules. These metrics are calculated based on documents being retrieved rather than by the final end-to-end evaluation of decision quality. Previous works have defined different benchmarks for retrieval-augmented generation, such as the TREC Deep Learning Datasets and the MS MARCO dataset. For retrieval-augmented generation in the context of knowledge-driven decision automation, a single standard dataset cannot adequately cover all relevant horizontal or vertical aspects. Thus, as the various knowledge-driven decision automation pipelines require different kind of knowledge in specific decision domains, their retrieval components are separately evaluated on the datasets corresponding to their respective decision domain.

5.2. Decision Quality and Reliability

Important enterprise decision domains include IT governance, service operations, service risk management, and customer service assurance. Decisions can involve service governance actions, incident classification and escalation, service risk assessment, and service quality assurance. Supported by sufficient and relevant retrieval knowledge, responses for such decisions can be generated with acceptable quality using a small LLM. Quality indicators include accuracy, timeliness, consistency, robustness, and user acceptance.

Accuracy is defined as the degree to which a retrieved RAG response (such as an incident escalation) is consistent with the corresponding ground truth. The human experts “speak for themselves” in determining correctness; any errors in their own decisions are inherent noise during data collection and are not deemed to compromise the overall analysis. Timeliness assesses whether a response was generated within an acceptable timeframe. Consistency evaluates whether two similar but non-identical input queries (such as those requesting classification of the same incident when using different wording) yield similar outputs. As with accuracy, no human-based framework can be applied. For the small LLM, a pair of retrieval-augmented generation responses is labelled as consistent or inconsistent following a binary vote from two human experts. Robustness measures the rate at which a retrieval-augmented generation response is accepted by a group of human experts as valid even though the response falls outside the specific ground-truth example. It therefore provides a lower-bound guarantee for training data coverage. Finally, user acceptance quantifies whether the retrieval-augmented generation responses are acceptable in practice within an enterprise environment, following the same methodology as in previous work. Such acceptance is deemed indicative of satisfactory service decision quality in a wider enterprise context.

6. Discussion

Enterprise deployment of the proposed knowledge-driven decision-automation system is beneficial for organizations operating in a trusted, controlled environment. Individual components can be scaled, sourced, governed, and secured independently according to organizational, regulatory, or political considerations. Service consolidation across multiple organizations can reduce operational overhead and promote resource-sharing. Integration with existing front-end applications supports common enterprise user interfaces, while backend interoperability with enterprise service platforms ensures end-to-end automation across supported service domains.

The system enables a new class of safety-critical decision services that offer timely, accurate, and justifiable answers to complex question-explanation pairs. Addressing known areas of reliability and acceptance reinforces user confidence in, and promotes the use of, the system. Knowledge freshness may not be a limiting factor during day-to-day operations, but adjustments—via metadata and Elasticsearch configuration settings—allow timely responses to policy changes. The end-to-end decision-automation capability supports pending questions in critical service domains such as risk management, cybersecurity, and incident response.

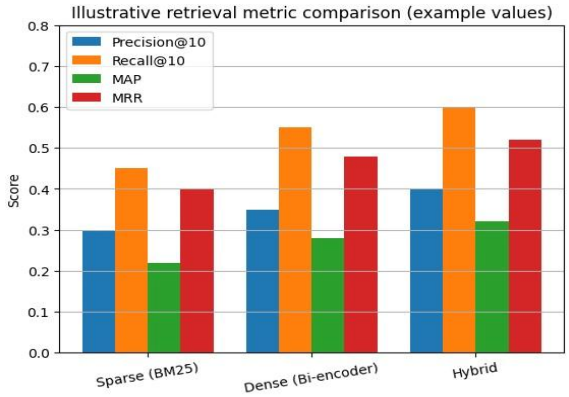

Fig 6: F1 Score Computation for Retrieval-Augmented Generation Performance Evaluation

Equation 4: F1 score (step-by-step)

F1 is the **harmonic mean** of precision and recall

Start from harmonic mean definition:

$$H(P,R) = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

Let $P =$ Precision, $R =$ Recall:

$$F1 = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

Put on a common denominator:

$$\frac{1}{P} + \frac{1}{R} = \frac{R+P}{PR}$$

So:

$$F1 = \frac{2}{\frac{R+P}{PR}} = 2 \cdot \frac{PR}{P+R} \boxed{F1 = \frac{2PR}{P+R}}$$

6.1. Implications for Enterprise Deployments

Enterprise service platforms frequently constitute the operational and engagement backbone of organizations. As such, system governance, risk oversight, and program execution must align with the enterprise's framework of decision and operating policies. Many of these decisions resonate across units and departments, yet the knowledge required to properly adjudicate these decisions in a timely manner is frequently siloed within the enterprise, making it difficult for users and system agents to locate it. Furthermore, even when such information does exist within professional experts, it can be inaccessible. Therefore, providing either human decision makers or agents with timely access to the enterprise knowledge base is an important functionality that should be supported by any modern enterprise service platform.

The RAG framework is well positioned to provide this capability for larger enterprises providing RAG-based decision support and augmented decision automation. Such solutions are mandated as part of commercial enterprise service platforms, where decision credibility, reliability, and trust are more important than accuracy as decision templates become institutionalized within the organization. Known decision domains—management, operations, risk, compliance, customer service, and so on—are either supervised or autonomously executed at a higher level by multiple agents, including human employees and software agents external to the enterprise service platform.

7. Conclusion

Retrieval-augmented generation represents a promising natural-language-based approach towards knowledge-driven decision automation on enterprise service platforms, particularly for non-expert users. For complex queries, the method can produce richer and more effective responses than presenting individual knowledge items on the ServiceNow Knowledge Base. Nevertheless, limitations must be considered. Incorrectly or inconsistently worded decisions tend to confuse rather than help the users. Feedback indicates that such policy enforcement should preferably be performed in alignment with the organization’s internal expert group.

Enterprise service platforms, offering a single point of contact and service fulfillment for users, optimize business operating and governance decisions within enterprises and institutions. Government organizations, in particular, are complex service ecosystems that face a set of cross-organizational, cross-border, and inter-generational problems. Knowledge-driven decision automation provides a solution, overcoming challenges related to retrieving recent decision patterns, injecting current knowledge into model parameters, dealing with uncertainty, justifying decision results for user trust, and evaluating decision constraints such as tardiness and stability. The results have profound implications: a user-friendly system to help non-expert decision-makers by generating decision results and suggestions directly in a natural language format, while reasoning over the retrieved knowledge patterns.

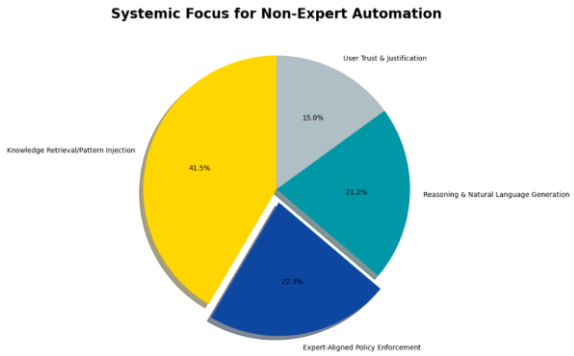

Fig 7: Systemic Focus for Non-Expert Automation

7.1. Summary of Findings and Future Directions

Enterprise Service Platforms require platforms that can support Knowledge-Driven Decision Automation. Knowledge-driven decision automation is concerned with providing timely and accurate information to support the needs of organizations and enterprises, for example in areas such as Enterprise Governance, Risk and Compliance or IT Service Management. A robust decision automation system should be able to articulate the Knowledge utilized to arrive at a conclusion, address uncertainty where it exists, and employ regulations or principles to be adhered to. A fully deployed Knowledge-Driven Decision Automation framework comprises of three modules: Knowledge Retrieval, Reasoning over Retrieved Knowledge and Decision Automation. These three modules can be viewed in terms of information requirements - What do enterprises want to know?, the source of information - What are the capable knowledge bases? and the process of reasoning - How to reason with the retrieved information? With respect to the Knowledge Retrieval module, an architecture has been presented and the retrieval capability evaluated.

Evaluation of the Knowledge Retrieval module demonstrated its capability in retrieving relevant information from both structured and unstructured text sources. The findings indicate that LLM based retrieval approach offers superior retrieval capability and can serve the enterprise's need of timely retrieval of information from dynamic data sources such as Microsoft 365. Evaluation of the Knowledge-Driven Decision Automation framework is an on-going effort. Focus areas include the ability to respond to complex decision queries, the consistency of responses to similar (but not identical) queries across time and users, the ability to handle uncertainty in the retrieved knowledge, the execution of external actions based on the decision and finally, acceptance of the responses by a cross-section of users.

8. References

[1] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6), 734–749.
[2] Meda, R. (2024). Agentic AI in Multi-Tiered Paint Supply Chains: A Case Study on Efficiency and Responsiveness. Journal of Compu-tational Analysis and Applications (JoCAAA), 33(08), 3994-4015.
[3] Albahri, A. S., Al-Obaidi, J. R., Zaidan, A. A., Zaidan, B. B., Hamid, R. A., Albahri, O. S., & Rashid, N. A. (2021). Multi-biometric systems: A state of the art survey and research directions. IEEE Access, 9, 104050–104095.
[4] Sheelam, G. K. (2024). Towards Autonomic Wireless Systems: Integrating Agentic AI with Advanced Semiconductor Technologies in Telecommunications. American Online Journal of Science and Engineering (AOJSE)(ISSN: 3067-1140), 2(1).
[5] Aldasoro, I., Gambacorta, L., Giudici, P., & Leach, T. (2024). The rise of artificial intelligence in financial services. BIS Quarterly Review, 1, 45–58.
[6] Kummari, D. N., & Burugulla, J. K. R. (2023). Decision Support Systems for Government Auditing: The Role of AI in Ensuring Transparency and Compliance. International Journal of Finance (IJFIN)-ABDC Journal Quality List, 36(6), 493-532.
[7] Anagnostopoulos, I. (2022). Artificial intelligence in financial services: A critical review of applications and challenges. Journal of Financial Regulation and Compliance, 30(2), 195–210.
[8] Anderson, C., Domingos, P., & Weld, D. S. (2002). Relational Markov models and their application to adaptive web navigation. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 143–152). ACM.
[9] Ramesh Inala. (2023). Big Data Architectures for Modernizing Customer Master Systems in Group Insurance and Retirement Planning. Educational Administration: Theory and Practice, 29(4), 5493–5505. https://doi.org/10.53555/kuey.v29i4.10424
[10] Arik, S. Ö., & Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. In Proceedings of the AAAI Conference on Artificial Intelligence (pp. 6679–6687).
[11] A Scalable Web Platform for AI-AugmentedSoftware Deployment in Automotive Edge Devices via Cloud Services. (2024). American Advanced Journal for Emerging Disciplinaries (AAJED) ISSN: 3067-4190, 2(1). https://aajed.com/index.php/aajed/article/view/12
[12] Attia, P. M., & Dayan, P. (2022). Neural systems for decision making. Nature Reviews Neuroscience, 23(9), 561–576.
[13] Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv.
[14] Nagabhyru, K. C. (2024). Data Engineering in the Age of Large Language Models: Transforming Data Access, Curation, and Enterprise Interpretation. Computer Fraud and Security.
[15] Banerjee, S., & Pedersen, T. (2003). The design, implementation, and use of the Ngram Statistics Package. In Proceedings of CICLing 2003 (pp. 370–381). Springer.
[16] Aitha, A. R. (2024). Generative AI-Powered Fraud Detection in Workers' Compensation: A DevOps-Based Multi-Cloud Architecture Leveraging, Deep Learning, and Explainable AI. Deep Learning, and Explainable AI (July 26, 2024).
[17] Basel Committee on Banking Supervision. (2023). Principles for the management of counterparty credit risk. Bank for International Settlements.
[18] Bast, H., Buchhold, B., & Haussmann, E. (2016). Semantic search on text and knowledge bases. Foundations and Trends in Information Retrieval, 10(2–3), 119–271.
[19] Vajpayee, A., Khan, S., Gottimukkala, V. R. R., Sharma, D., & Seshasai, S. J. (2025). Digital Financial Literacy 4.0: Consumer Readiness for AI-Driven Fintech and Blockchain Ecosystems. International Insurance Law Review, 33(S5), 963-973.
[20] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). ACM.
[21] Bengio, Y., Lecun, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

[22] Segireddy, A. R. (2024). Machine Learning-Driven Anomaly Detection in CI/CD Pipelines for Financial Applications. Journal of Computational Analysis and Applications, 33(8).
[23] Bholat, D., Gharbawi, M., & Thew, O. (2023). Machine learning, big data, and financial stability. Financial Stability Review, 27, 33–49.
[24] Bhutani, N., Zheng, X., & Jagadish, H. V. (2020). Learning to inject knowledge into language generation. In Proceedings of EMNLP 2020 (pp. 3469–3479). ACL.
[25] Amistapuram, K. (2024). Generative AI in Insurance: Automating Claims Documentation and Customer Communication. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 15(3), 461–475.
[26] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python. O'Reilly Media.
[27] Yandamuri, U. S. AI-Driven Decision Support Systems for Operational Optimization in Hospitality Technology.
[28] Rongali, S. K. (2024). Federated and Generative AI Models for Secure, Cross-Institutional Healthcare Data Interoperability. Journal of Neonatal Surgery, 13(1), 1683-1694.
[29] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
[30] Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 30(1–7), 107–117.
[31] Varri, D. B. S. (2023). Advanced Threat Intelligence Modeling for Proactive Cyber Defense Systems. Available at SSRN 5774926.
[32] Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In Proceedings of SIGIR 2004 (pp. 25–32). ACM.
[33] Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of SIGIR 1998 (pp. 335–336). ACM.
[34] Nagubandi, A. R. (2023). Advanced Multi-Agent AI Systems for Autonomous Reconciliation Across Enterprise Multi-Counterparty Derivatives, Collateral, and Accounting Platforms. International Journal of Finance (IJFIN)-ABDC Journal Quality List, 36(6), 653-674.
[35] Chen, J., Chen, H., Wu, S., Yu, H., & Xiong, C. (2022). In-context learning with retrieval augmentation. In Proceedings of EMNLP 2022 (pp. 4569–4582). ACL.
[36] Guntupalli, R. (2024). Enhancing Cloud Security with AI: A Deep Learning Approach to Identify and Prevent Cyberattacks in Multi-Tenant Environments. Available at SSRN 5329132.