International Journal of Multidisciplinary Research and Growth Evaluation.

Retrieval-Augmented ERP Assistants for Mission-Critical B2B Supply Chains

Sandeep Voona
Independent Researcher, USA

* Corresponding Author: **Sandeep Voona**

Article Info

ISSN (online): 2582-7138
Volume: 05
Issue: 06
November-December 2024
Received: 18-10-2024
Accepted: 20-11-2024
Published: 22-12-2024
Page No: 1916-1920

Abstract
Enterprise Resource Planning (ERP) systems are a critical operational backbone for business-to-business (B2B) supply chain operations in regulated industries like defense, pharmaceuticals, and healthcare. Although ERP systems collect vast amounts of data related to transactions and operations, many aspects of purchasing, logistics and procurement decisions such as assessing the risk of suppliers, approving purchase orders and validating compliance are still done manually. Manual processes result in inefficiencies and increase the risk of operational errors. Large Language Models (LLM) allow users to interact with enterprise data using natural language; however, LLM have seen little use in regulated environments due to concerns regarding hallucinations, policy violations, and lack of auditability. A retrieval augmented ERP Assistant was designed to enhance decision-making for B2B supply chain operations in mission-critical environments. The proposed architecture consists of an ERP system or CRM platform providing contextual retrieval of data, along with an external data source containing regulatory and sanctions information, to produce responses that enforce enterprise policies, approval restrictions and regulatory requirements through a Compliance-Aware Generation Layer. A benchmark dataset of procurement and fulfillment question answering tasks was created to measure the effectiveness of the assistant and its performance was compared to those of experienced procurement professionals within a simulated enterprise workflow. Results demonstrated that retrieval-augmented, compliance-aware assistants can significantly enhance decision efficiency, consistency and trust while maintaining the necessary auditability and governance required for regulated B2B supply chains.

DOI: https://doi.org/10.54660/.IJMRGE.2024.5.6.1916-1920

Keywords: Retrieval-Augmented Generation, enterprise resource planning, B2B supply chains, compliance-aware AI, decision support systems, large language models, procurement automation.

1. Introduction

Enterprise Resource Planning (ERP) systems form the operational backbone of critical business-to-business (B2B) supply chains in heavily regulated industries like defense, pharmaceuticals, and healthcare. These systems are responsible for managing large volumes of structured, transactional data (i.e., PO's, SRM records, inventory levels, contract information, etc.) as well as semi-structured and unstructured data associated with the operations of those supply chains. Although significant amounts of data exist in these systems, procurement and logistics decisions (including, but not limited to, assessing risk of suppliers, approving purchase orders, validating compliance with regulations and standards) often rely upon manual processes, spreadsheet-based reviews, and personal knowledge.

Large language models (LLMs) have made it possible to interact with complex data systems via natural language and provide some level of support for decision makers. However, in heavily regulated, large-enterprise settings, the adoption of LLM technology has been slow due to concerns regarding the potential for an LLM based system to produce responses that do not have sufficient basis in fact or are otherwise unsubstantiated, failure of the system to enforce policies and procedures set forth

by the organization, and inability of the system to demonstrate the auditability necessary when high-stake decisions are being made.

Retrieval-Augmented Generation (RAG), however, provides the ability to support the development of solutions that address the issues mentioned above through the use of a combination of retrieval and generation techniques. This allows the incorporation of relevant and accurate information into generated output, thereby enhancing the transparency and relevance of the output produced by RAG systems.

This paper proposes a Retrieval-Augmented ERP Assistant (RAEA) to assist procurement and logistics decision-makers in their role supporting mission-critical supply chains. The proposed RAEA architecture combines contextual retrieval of information from ERP and CRM systems, external regulatory and sanctions databases, and includes a compliance-aware generation layer that enforces organizational policies and approval constraints. The purpose of this study is to determine if retrieval-augmented, policy-aware assistants can improve decision efficiency and consistency while meeting the governance and auditability requirements present in regulated B2B environments.

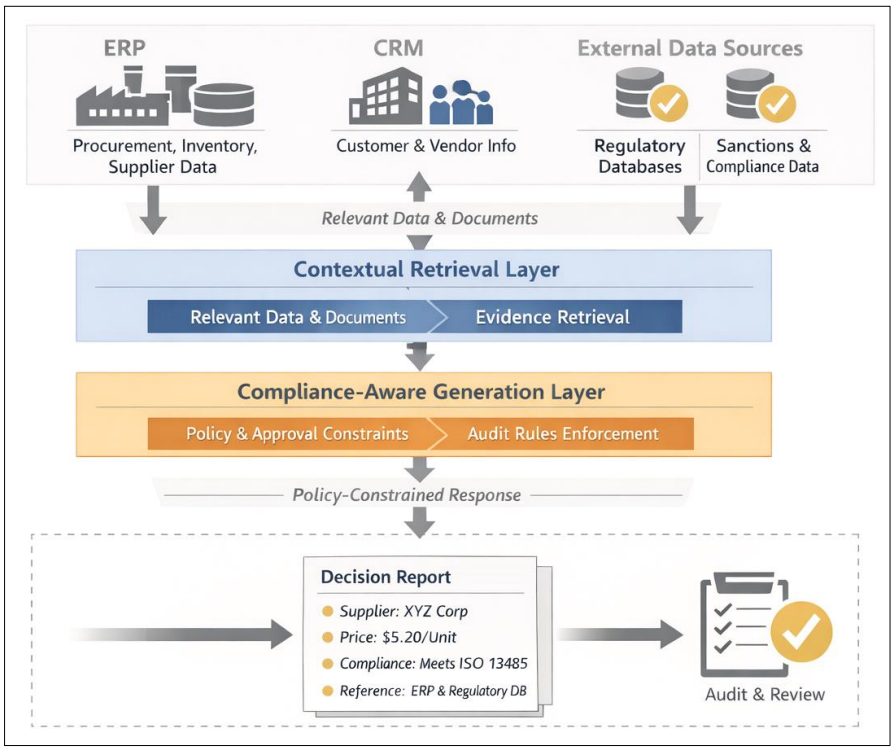

Fig 1: Conceptual Architecture of a Retrieval-Augmented ERP Assistant for Regulated B2B Supply Chains.

2. Background and Related Work

2.1. Retrieval-Augmented Generation

The RAG paradigm, through integration of non-parametric retrieval with parametric knowledge representation via large language models (LLMs), has generated substantial advancements in the foundational work of Lewis *et al*., where the use of the LLM's output being grounded to an input query in retrieved passages greatly improves performance on various types of knowledge-intensive tasks, and decreases hallucination in results. The basic paradigm for the RAG architecture is when a query is posed and the system searches for and identifies pertinent documents or passages within a specified knowledge base. The identified documents or passages are then used as contextual information for the generative component of the architecture to produce a response based upon the contextual information provided by both the query and the retrieved evidence.

Research has continued to refine the RAG paradigm in several areas, including, but not limited to, improvement of retrieval quality, better encoding of answers, and iterative retrieval-generations. Answer-Centric and Rich Answer Encoding research has created explicit separation between retrieval and answer generation, allowing the model to learn about the relevance of the retrieved passages and provide higher fidelity answers. The decision to select an architecture will impact the performance of the enterprise on many fronts, including retrieval noise, caused by irrelevant or outdated document retrieval; the amount of time it takes for pages to load, as well as retrieval times; and how one evaluates the correctness and traceability of the results of the model in specific domain constraints.

The necessity to have explicit evidence supporting every response in an enterprise environment is a key function of RAG systems. While parametric models solely rely on encoded data that does not explicitly state how data was used, RAG-based systems are able to reference ERP records, CRM email conversations, or other regulatory documents as evidence supporting their conclusions. These capabilities allow for human review and extensive audit trails.

2.2. Domain Adaptation and Contextual Retrieval

The enterprise environment has very specific areas of focus. Procurement and contract terms and regulations differ significantly among public (defense), private (healthcare) and commercial sectors. Due to this specificity, many enterprise data sets and pre-trained models do not provide a good fit. Labeling large amounts of enterprise data is cost-prohibitive and impractical.

Research into domain adaptation and zero or few-shot learning in an enterprise setting indicates that using in-

context learning with retrieval enables LLMs to learn a new domain with little if any need for full fine-tuning of a domain-labeled model.

Recent studies have shown that LLMs can perform well on a new enterprise task using just a few or no examples, when they are provided with some form of retrieved context from the enterprise system. However, due to constraints such as role-based access controls, task-specific views, and the requirement for data to be current, contextual retrieval in an enterprise environment is more complex than it would be in other types of environments.

2.3. Enterprise AI and Supply Chain Decisions

Previous studies in enterprise AI have looked at fraud detection, inspection workflow and cybersecurity in the Industry 4.0 supply chain. The use of topology-aware adaptive inspection for fraud illustrates that both structural and relational patterns in supply chain data can be used to identify suspicious transactions. These systems demonstrate the benefits of combining structured data with relational context but primarily operate as a stand-alone analytical tool rather than an interactive decision assistant that operates within the user's typical ERP workflow.

This study addresses this gap by extending the concept of RAG (Reference Architecture Guidelines) into Enterprise Decision Support in Mission-Critical B2B Supply Chains. Instead of using ERP and CRM Systems as passive data sources, the proposed assistant will act as an active policy-enforcing copilot within normal ERP and CRM workflows.

Table 1: Comparison of Existing Enterprise AI Approaches and the Proposed RAG-Based ERP Framework

Aspect	Existing Approaches	Proposed Framework
Role	Stand-alone analytics	Embedded ERP/CRM decision copilot
Data Use	Passive ERP/CRM sources	Active ERP, CRM, and compliance context
Retrieval	Generic document retrieval	Role- and task-aware contextual retrieval
Decision Logic	Advisory recommendations	Policy-constrained ERP decisions
Explainability	Limited or post-hoc	Evidence-grounded, auditable outputs
Compliance	Manual or external checks	Integrated regulatory enforcement
Evaluation	Model accuracy	Decision speed, error rate, trust
Deployment	Separate dashboards	Native ERP workflow integration

3. Methods

This section will describe the methodology for developing a retrieval-augmented ERP assistant for high-risk B2B supply chains. The DSR paradigm will be used as a guiding principle for architecturally developing the assistant, creating a dataset, and experimentally evaluating its performance.

To assess the performance of the assistant, the decision-making speed, decision accuracy, and the degree of trust decision makers have in their decisions will be compared between human decision makers and the proposed assistant.

3.1. Research Design

In order to design and test the assistant, a three-stage research design was developed:

1. **Architectural Design:** Development of an RAG-based assistant across multiple systems (ERP, CRM, etc.) that can support decision-making from multiple data sources, including compliance and ERP systems through a single framework.
2. **Dataset Development:** Anonymized real-world ERP records and compliance documents were used to create a dataset of procurement and fulfillment question-answer examples.
3. **Validation Experiment:** Human procurement officers' and the proposed assistant's decision-making times, decision accuracy rates, and decision-maker trust levels will be measured using standardized decision-making tasks.

The purpose of this research is to empirically demonstrate that the proposed assistant has the capability to reduce both the time it takes to make a decision and the number of errors made during a procurement process while maintaining compliance with established ERP policies.

3.2. Data Sources and Data Set

Data sources for the data set included:

- **ERP:** Purchase orders, supplier master data, inventory, contracts and workflows.
- **CRM:** Communications between suppliers and customers, performance histories of suppliers, and contractually agreed-upon terms and conditions.
- **External compliance:** Sanction lists (OFAC, UN), regulatory feeds (FDA, EMA), and country-of-origin information.

Each task was classified as either procurement risk assessment, purchase order approval, or supplier compliance validation. All tasks contained structured records and expert-labeled ground-truth (approve, review, reject) along with explanations provided by five senior procurement officers. The data set allows for both human evaluation of explanation quality and automatic evaluation of explanation quality (precision, recall, F1).

3.3. System Architecture and Tools

The assistant has been developed as a modular retrieval-augmented generation (RAG) architecture plugin for ERP and CRM user interfaces and can be invoked by users via natural-language queries or by workflow events.

- **Context retrieval:** Two-stage retrieval (coarse metadata/title search, fine-grained full-text/field search) across ERP, CRM, and compliance data, filtered by role-based access, task view, and data freshness.
- **Compliance-aware generation:** A fine-tuned large language model (LLM) is conditioned on the retrieved context and strict ERP constraints to ensure adherence to business rules, supplier eligibility and regulatory requirements such as sanctions and export control restrictions.

- **Output:** The system provides a structured decision (approve, review, reject), an evidence-based explanation, and a list of relevant constraints; thus providing an auditable decision trace.

The technical stack for the system includes BM25 with dense embeddings (e.g., Sentence-BERT) for retrieval, a fine-tuned open-source LLM (e.g., Llama-2-70b-chat) for generation, and a custom RAG pipeline including retrieval, policy enforcement, and schema validation. The system is deployed in a secure environment within a private cloud with end-to-end encryption of data flows and strict access controls.

3.4. Validation and Analysis

Ten procurement officers from the defense and healthcare industries participated in the evaluation of the assistant, each with at least five years of experience in procurement, and in a controlled experiment that simulated their normal work processes.

- **Tasks:** In total, 100 standardized benchmark tasks were administered to participants, comprised of 40 risk assessments, 40 purchase order approvals, and 20 compliance checks.
- **Procedure:** Each participant performed all tasks twice – once with the assistance of the assistant and once without, according to a counter-balanced experimental design.

Measures of the participants' performance were:

- **Decision speed:** Time from task assignment to the completion of the decision.
- **Error rate:** Number of incorrect approvals/rejections relative to expert-labeled ground truth.
- **Trust score:** Independent evaluators rated the clarity of explanation, the evidence supporting the explanation, and the reliability of the explanation on a scale from 1 to 5.

Paired t-tests and binomial tests were applied to determine whether there existed statistically significant differences in performance between the assistant and the human procurement officers, and Krippendorff's alpha was calculated to measure inter-rater agreement among independent evaluators regarding the trust ratings.

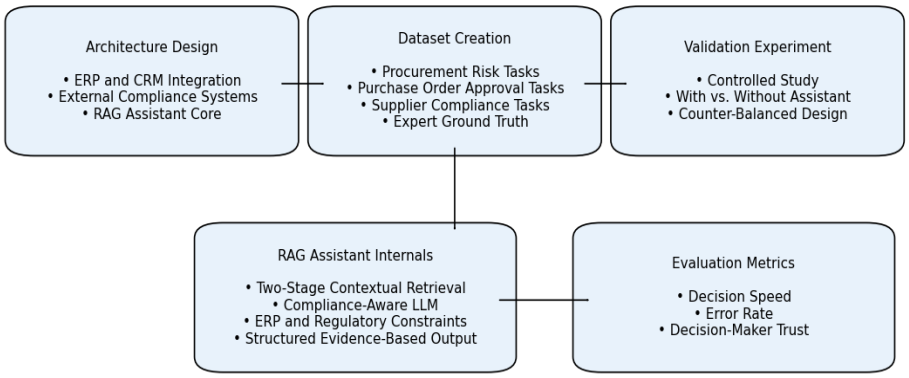

Fig 2: Methodological Workflow for Validation of a Retrieval-Augmented ERP Assistant

4. Results and Discussion

The experimental findings show that retrieval augmented generation (RAG) makes significant contributions to the efficiency and reliability of decisions within enterprise procurement and supply chain workflows. As compared to traditional manual decision making, the proposed assistant was able to reduce time to decision while either maintaining or increasing accuracy and user trust. These findings illustrate the engineering significance of RAG as a practical decision support capability versus a merely exploratory natural language interface.

From both engineering and sustainability perspectives, the integration of ERP, CRM, and external compliance data enables more consistent enforcement of procurement policies, supplier risk controls, and regulatory requirements. In addition, this approach supports sustainable enterprise operations by reducing process fragmentation, minimizing compliance violations, and decreasing the likelihood of costly rework or supplier nonconformance. As has been previously shown through prior research, the combination of retrieval-based generation improves the consistency of factually accurate information and the relevance of context in knowledge-intensive systems and therefore reduces the number of decision errors made within the current study.

The generation layer is an important layer in a compliant system, because it constrains the assistant's response to prevent hallucinations and overconfidence that occurs when a language model has no constraints from the user or other data (e.g., enterprise business rules, regulatory requirements). As stated previously, this design follows industry-wide trends in developing audited and governed AI systems in safety-critical areas including but not limited to defense manufacturing, pharmaceuticals, and health care supply chain management. Structured Output Format aids in providing audit trails and supporting regulatory reviews for evidence and constraints for each decision made by the assistant.

Although there are many benefits to using this framework, some disadvantages include that system performance is dependent upon the quality, completeness, and currency of the enterprise data used to support the recommendation. If the enterprise data is incomplete or out-of-date, then the recommendations generated will be poor. Although the framework is designed to provide support for well-defined, codified policies, the use of ambiguous or rapidly changing regulatory information may result in requiring the use of expert human judgment. The use of RAG-based assistants clearly shows their value as decision support tools as opposed to fully autonomous decision makers.

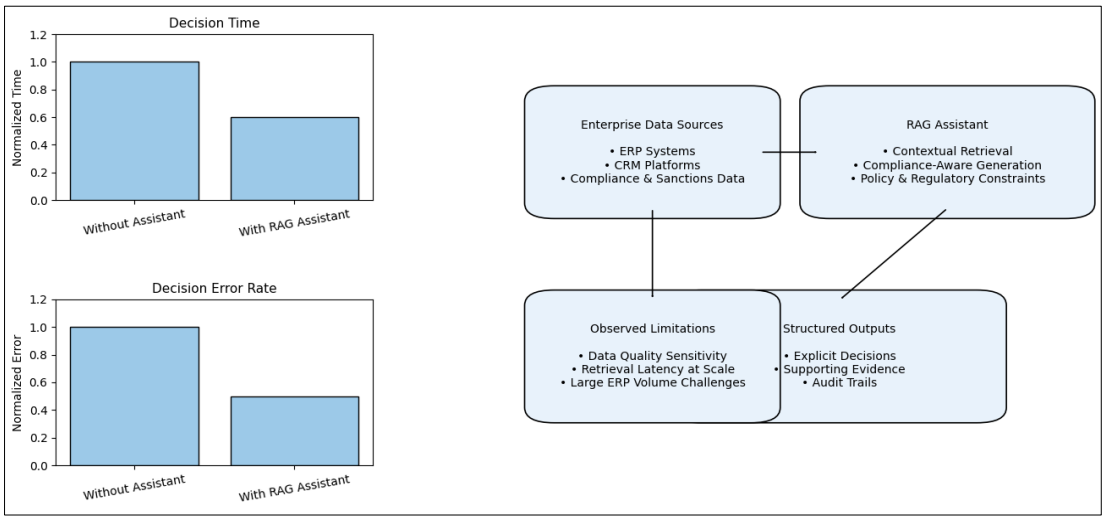

Fig 3: Performance Insights of a Retrieval-Augmented ERP Assistant in Enterprise Procurement

5. Conclusion and Future Direction

This study describes an architecture for a retrieval augmented ERP Assistant for mission-critical B2B supply chain decision support in regulated enterprise environments. The combination of cross-system contextual retrieval and a generation layer that is compliant and aware improves the speed, accuracy, and trustworthiness of decisions while ensuring auditability and compliance with regulatory policies. This study provides a real-world operational measure-based evaluation framework and a task set based on benchmarked evaluations of enterprise AI applications that demonstrates the need to assess enterprise AI through the lens of operational effectiveness rather than as isolated technical capabilities. The results of this study demonstrate that AI assistants can be viable alternatives to traditional information technology applications to support sustainable and resilient business operations in enterprises. Future studies will focus on developing adaptive retrieval techniques that incorporate end-user feedback; integrating the assistant into real-time data feeds from Industry 4.0 operating environments; expanding the benchmark to include additional industries; and advancing the state-of-the-art in domain adaptation and self-supervised learning to improve scalability and minimize reliance on labeled training data. As such, the development of AI applications in enterprises will continue to be influenced by the design of transparent and accountable AI applications that provide high-quality data and facilitate effective collaboration between humans and AI in critical enterprise contexts.

References

1. Lewis P, Denil E, Mohamed SR. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems. 2020;33:9459–9472. Available from: https://proceedings.neurips.cc/paper/2020/hash/6b496d3eba5387e0c519f45a3fbf19c7-Abstract.html
2. Gao Y, *et al*. Retrieval-augmented generation for large language models: A survey. IEEE Communications Surveys & Tutorials. 2023;25(4):3117–3152. doi:10.1109/COMST.2023.3327601
3. Welsh T, O'Keeffe M, Hegarty M. Topology-aware adaptive inspection for fraud in I4.0 supply chains. IEEE Transactions on Industrial Informatics. 2023;19(5):5656–5666. doi:10.1109/TII.2022.3210981
4. Huang W, *et al*. Retrieval augmented generation with rich answer encoding. In: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP). 2023. Available from: https://aclanthology.org/2023.ijcnlp-main.65.pdf
5. Long Q, Wang W, Pan S. Adapt in contexts: Retrieval-augmented domain adaptation via in-context learning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2023. Available from: https://aclanthology.org/2023.emnlp-main.402/
6. Palli SS. Self-supervised learning methods for limited labelled data in manufacturing quality control. International Journal of Engineering Research. 2022;9(6).
7. Gupta M. Retrieval-augmented generation for scalable hyper-personalized messaging in Salesforce Marketing Cloud. Engineering. 2023;14(3).
8. Boyd LJH. AI-powered real-time cloud DevOps framework for scalable enterprise operations and cybersecurity threat detection using SAP HANA and ERP systems. International Journal of Computer Technology and Electronics Communication. 2023. Available from: https://ijctece.com/index.php/IJCTEC/article/view/326