AI-Powered Decision Support for Modern Enterprises

Shashikala Valiki
Independent Researcher
shashikala.valiki.researcher@gmail.com

Abstract

Enterprises generate large volumes of unstructured information in reports, presentations, and conversations. This information often resides in diverse, distributed data sources, limiting its use for direct decision-making and leading to the adoption of adhoc workflows. The delay, inaccuracy, and poor governance of such work increase risk. A variety of solutions have sought to improve the integration of knowledge work into enterprise processes.

A Retrieval-Augmented Generation framework, augmented with a context-preserving knowledge-retrieval layer, provides decision-support capabilities across preset operation, logistics, planning, and compliance domains. Tools, interfaces, and monitoring dashboards allow for real-time context creation, approval, and rollback during LLM inference. Data provenance tracks root sources, promotes trust in responses, and accelerates compliance audits.

Keywords : Context-Preserving Generative AI,Enterprise Knowledge Workflows,LLM-RAG Framework,Retrieval-Augmented Generation (RAG),Real-Time Decision Support,Management Decision Intelligence,Enterprise Knowledge Management (EKM),Semantic Search,Hybrid Retrieval (Sparse + Dense),Contextual Grounding,Policy-Aware Generation,Role-Based Access Control (RBAC),Data Governance & Compliance,Human-in-the-Loop (HITL),Explainability & Provenance Tracking.

1. Introduction

Enterprise knowledge work relies heavily on accessible, accurate knowledge. Management decision-making, service delivery, business analytics, and operational execution depend on advanced, often confidential knowledge. The sources, age, accuracy, and appropriateness of knowledge become critical as tasks progress. Newer LLMs (i.e. Generative AI) can simplify and accelerate workflow via verbal dialogues. Yet, even with their integration, latency and reliability concerns hamper enterprise-class deployment. Data privacy, data protection, and knowledge confidentiality issues likewise remain significant.

Generative AI must consider data provenance and the stage-appropriate credibility of knowledge at all times to be fit for use in enterprise knowledge workflows. Thus, its integration requires additional architectural constraints to preserve context—provenance, structure, appropriate governance, user monitoring, and compliance checks in workflow dialogues across use cases.

The LLM-RAG (Language Model with Retrieval-Augmented Generation) instance proposed meets these context-preserving requirements for a specific real-time management decision-support application. An LLM-RAG-context-preserving framework identifies the design requirements for a Layer of Knowledge Retrieval, a Layer of Memory, Memory of Contextual Information, and overlay Governance that can monitor use in real time and roll back research with prior versions.

1.1. Background and Significance

Enterprise knowledge workflows encompass the systematic and consistent engagement with knowledge asset domains that are essential for achieving a company's vision, strategy, or objectives. Generic stakeholders engage with knowledge asset domains based on the level of effort required by the relevant knowledge asset, with minimal effort at one end and extensive effort at the other. An input-output framework maps decision-related knowledge assets to stakeholder and workflow contexts, with different configurations representing distinct aspects. The structure enables adaptation for specific conversations, supporting Latent-Class Models that classify such conversations into supporting operations and logistics functions, strategic planning initiatives, or risk management and compliance perspectives.

Generative AI refers to a subfield of artificial intelligence that focuses on systems that can generate new content, such as text, images, and audio, based on input data. The emergence of Generative AI has generated tremendous excitement, with recent versions of established tools attracting millions of users in a short time. Retrieval-Augmented Generation (RAG) supports the generation of contextually relevant content by integrating knowledge retrieval with context-aware generative models. While traditional Large Language Models (LLMs) generate text based on model parameters and user-provided prompts, RAG systems query additional knowledge sources throughout the inference process to provide more informed responses. An LLM-RAG framework empowers enterprise knowledge workflows by supplying contextually relevant knowledge to operating and logistics functions, strategic planning initiatives, and risk management. The resulting models assist stakeholders in making decisions that leverage enterprise knowledge assets while preserving data provenance and lineage.

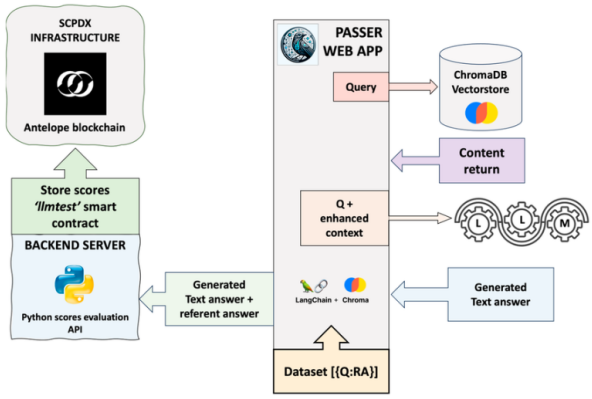


Fig 1: Workflow diagram for RAG LLM query processing and score storage

1.2. Research design

Enterprise management and decision support have traditionally relied on standard business intelligence approaches to aggregate, analyze, and make sense of operational data. Although well-established, business intelligence methods cannot operate in real time. With the increasing use of data lakes for enterprise data management, information provisioning cycles are shrinking—a trend that is likely to continue. Real-time decision-support solutions have emerged for certain operational processes. For management decision-support systems requiring foresight, dialogue with the AI-generated thumbnail plans may help ensure that the plans would achieve the desired objectives, all else being equal.

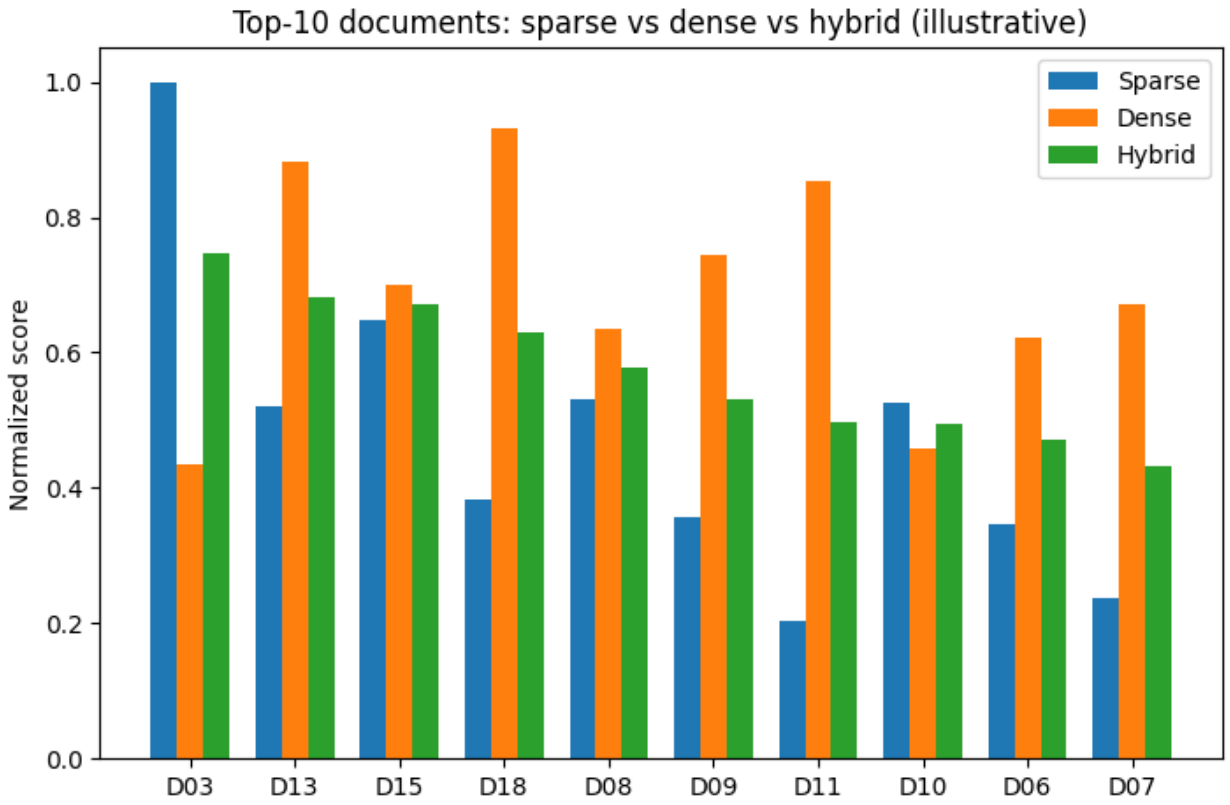
Many emerging technologies continue to receive attention from practitioners and researchers alike. Generative AI, Retrieval-Augmented Generation (RAG), and Large Language Models (LLMs) are but a few. Early adoption trends indicate a high level of curiosity among senior managers and a willingness to experiment with and invest in new solutions. Building system architectures, data models, and implementation procedures also requires caution and care. Given the huge financial, operational, and reputational consequences of a major decision, management remains aware of the hazards and challenges of generative AI when applied to support real-time enterprise knowledge workflows.

2. Background and Motivation

Enterprise knowledge workflows support complex management decision-making, requiring diverse knowledge inputs and prompt collation for contextual accessibility. Generative AI, particularly Retrieval-Augmented Generation, offers rapid yet often contextually ungoverned insights. Optimising GenAI systems for context preservation allows rapid, compliant deployment for real-time requests within decision-support scenarios.

Enterprise knowledge workflows rest on the premise that effective management decision-making relies not only on relevant historical experience but also on diverse knowledge inputs. Such knowledge underpins various business operations, including demand fulfillment and supply allocation. Although traditionally considered semi-structured and tacit, enterprise knowledge is increasingly captured as structured records, unstructured narratives, and numerical data. When a strategic decision must be made, the tacit contextual knowledge of planning specialists is relied upon to collate relevant information. Yet the collation of structured, unstructured, and numerical data relevant to the same decision point cannot be fully automated and must still rely on ad-hoc requests to separate systems involved in that process. In these scenarios, decision-support prompts are able to influence but not determine final decisions. Therefore, the speed of the response is more critical than its accuracy. Generative AI can efficiently scan contextually disparate information assets and suggest conclusions; however, Generation AI delivers these insights without any guarantee of accuracy or validity. Such weaknesses have elevated hallucination to a common topic of discussion within the field of Natural Language Processing.

ISSN: 3067-283X
AMERICAN JOURNAL OF ANALYTICS AND ARTIFICIAL INTELLIGENCE
VOLUME: 04 ISSUE: 01
RECEIVED: JANUARY 28 REVISED: FEBRUARY 04 ACCEPTED: FEBRUARY 23 PUBLISHED: MARCH 17



2.1. Enterprise Knowledge Workflows

Enterprise knowledge workflows involve the structured application of specialist knowledge within and between enterprises to address the explicit needs of stakeholders. Principal stakeholders are consuming enterprises, knowledge suppliers, and the specialists who address requirements from various sources. Knowledge workflows require specialist know-how—often tacit and not readily available from knowledge repositories—to create contextually informative blends of information or data. At scale and linked through an enterprise's content systems, specialist knowledge and know-how collectively form enterprise knowledge. These systems connect with people to contextualise knowledge-rich content. Knowledge workflows are tangible and repeatable activities linked to decision making. For management, high-level categorisation simplifies governance and understanding of control, risk, and compliance.

In the real-time decision-support context, commonly used management-data concepts—including KPIs, risks, and dependencies—align to create data for Constant Data Feeding decisions. Enterprise internal systems, including risk-management repositories and operational (assets and geography) and external (macro) data, connect to emerging needs. Three categories of Constant Data Feeding use cases have been identified: Operations & Logistics Support, Strategic Planning, and Risk Management & Compliance. Support uses apply decisions derived from Scenarios prepared through Strategic Planning or Risk Management.

2.2. Generative AI and Retrieval-Augmented

Generative AI refers to AI and machine learning systems that leverage existing data sources to generate new outputs, such as text, images, audio, and video. Unlike traditional AI systems that rely exclusively on rules, heuristics, and classification models, the output of Generative AI systems is supplied via retrieval from large external databases, enabling novel combinations and reframing of data across specific use cases. Generative Pre-trained Transformers (GPTs) — like DALLE-2, Midjourney, ChatGPT, and Google Bard — have sparked public interest and intensified academic inquiry into real-time-name-generative AI models further over time. However, while LLMs deliver remarkable capabilities, their current formulations miss the potential to differentiate and manage context by preserving data provenance. Retrieval-Augmented Generation (RAG) is emerging as a powerful way to refill gaps and augment GPT-like solutions in business domains. RAG combines LLM with modular machine-learning capabilities, delivering the potential to retrieve specific data bursts and combine these with input requests to generate elaborated, structured outputs even for repeat queries. Although a RAG system adds latency to individual queries, RAG systems continuously fill and extend (dramatically) the knowledge base, producing an always precise and evolving answer set that progressively reduces latency.

The contextual embedding structure of LLM-RAG enables the new RAG model to integrate and leverage context-specific decision data and expertise for contextualized, situation-relevant output generation. A generic version of the new model can be applied in any new context while achieving additional performance with reuse, iterative learning, and real-time error detection. However, although the LLM-RAG model can significantly reduce error rates, hallucination may still remain a factor. . Validation and confidence-checking mechanisms are paramount when deploying the LLM-RAG model within critical and sensitive domains such as defence, finance, law, and human resources.

Equation 1: Semantic similarity score (prompt ↔ retrieved context)

Step-by-step derivation (cosine similarity)

1. Embed prompt text into a vector:

$$\mathbf{q} = E(\text{prompt}) \in \mathbb{R}^d$$

2. Embed a candidate passage/document chunk:

$$\mathbf{d} = E(\text{chunk}) \in \mathbb{R}^d$$

3. Dot product measures alignment:

$$\mathbf{q} \cdot \mathbf{d} = \sum_{i=1}^{d} q_i \, d_i$$

4. Vector length (magnitude):

$$\| \mathbf{q} \| = \sqrt{\sum_{i=1}^{d} q_i^2}, \| \mathbf{d} \| = \sqrt{\sum_{i=1}^{d} d_i^2}$$

5. Normalize dot product to remove dependence on magnitude ⇒ cosine similarity:

$$\cos(\mathbf{q}, \mathbf{d}) = \frac{\mathbf{q} \cdot \mathbf{d}}{\| \mathbf{q} \| \| \mathbf{d} \|}$$

2.3. Real-Time Management Decision Support Requirements

Service-level requirements for real-time management decision support reflect pressure on enterprise knowledge workflows to provide timely answers to frequently raised business questions. Initiatives from operations, logistics, and other business units depend on sizable amounts of data for planning, execution, and revision. When well-defined, such operations ask business question types for which latest data should be considered but for which static, historical reporting may not be timely enough. Ideally, a review of past operations, forecasts, and related parameters should all be accessible through a natural-language conversation, so that sources and checks of quality, accuracy, and bias are evident, and so that a requested conclusion can be evaluated on the same basis as the answer to a similar question posed 6 months earlier. Ongoing discussions must take into account possible raw-material supply constraints, potential future needs for downstream consolidation versus expansion, new developments in competitors' capacities or costs, and emerging dangers and vulnerabilities—all these by nature often requiring human judgment in combination with greatly varying amounts of available and known-quality information. In other words, late-stage confirmation with a small set of key individuals of next steps permitted and targeted, together with resultant progress and risk reassessment, should happen quickly. Business units feel this demand, too, for results from scenario analysis, determining when sales of key products to the retail sector are too high relative to customer traffic, or when arrival patterns for raw materials impose conflicting patterns in downstream processing for different destinations.

Any Generative AI system supporting such decision processes must operate in a timeliness, governance, training-sampling, and cost configuration aligned with that context. Prompt engineering becomes paramount—not only the framing of each prompt but the collection and filtering of training data chosen for whatever underlying guidance, ranging from response quality to frequency. For the key training set, user requests must be monitored for search volume as well as correctness and confirmability, suspected gaps addressed, any needed reductions in dataset quality managed, and updates injected where models are sufficiently leveraged. The supplies of raw information feeding real-time updates may have hidden additional noise but can never fully be eliminated—managers must monitor rates of confirmed error and attempt to merge them with question-answer inspection and prompt-frequency assessments from a model sitting "behind" the public interface.

3. Conceptual Framework

Context preservation in LLM-RAG systems is defined and key mechanisms for maintaining provenance are delineated, followed by an architecture overview that includes data flow, components, and interfaces.

ISSN: 3067-283X
AMERICAN JOURNAL OF ANALYTICS AND ARTIFICIAL INTELLIGENCE
VOLUME: 04 ISSUE: 01
RECEIVED: JANUARY 28 REVISED: FEBRUARY 04 ACCEPTED: FEBRUARY 23 PUBLISHED: MARCH 17

Context-preserving Support for Enterprise Knowledge Workflows

The proposed LLM-RAG framework preserves the context of a knowledge workflow by maintaining the origin of the data used for generation and validation, thus serving as a more interpretable and auditable information source when compared to conventional LLM configurations. The architecture consists of a knowledge retrieval layer, a contextual embedding structure, and a governance mechanism supporting real-time inference. These foundational elements are supported by action data specifying prompts and conditions for the actual use cases.

The first requirement is that contextual embeddings are created and persisted for data that will fuel a discussion within the workflow. Such embeddings must use a representation learning model whose role is to enlarge the set of future potential retrieval results. The second requirement is that such persistent storage of contextual embeddings is monitored, and that the distribution of content within the context can empirically signal possible degradation in the retrieval quality of the contextual support for attending to the action's focal event, namely, the Action if for the case of an operational decision, or the Action then for the case of a structured long legislation-text-formatted enterprise knowledge that must be relied on.

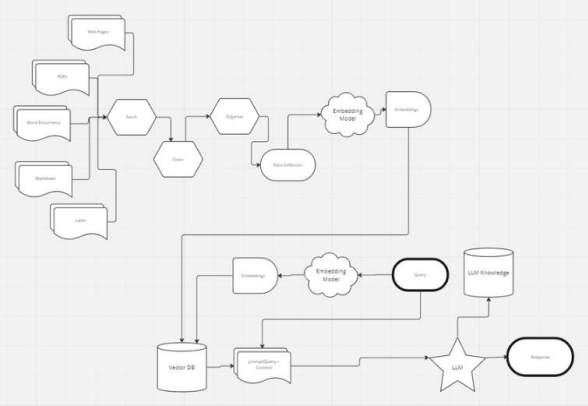


Fig 2: Conceptual framework for RAG-LLM

3.1. Context Preservation in LLM-RAG Systems

The examined generative AI architecture implements a context-preserving RAG capability and ensures a high degree of data governance and auditability. Context preservation in LLM-RAG systems refers to the capacity to maintain provenance over the information flow throughout the data life cycle. Specific mechanisms that accomplish provenance management for data-at-rest and data-in-motion are delineated, addressing aspects of data versioning, lineage, and auditability. It delineates constraints that govern the inclusion of information into the model memory.

The architecture overview depicts the general data flow, supporting components, and respective interfaces. It examines the complete contextual data life cycle within the information supply layer, from data generation to persistence and recovery, focusing on provenance facets. The description of data provenance covers all activities within Enterprise Knowledge Workflows

associated with the generation, ingestion, representation, and persistence of data, including Data Collection, Data Versioning, Data Lineage, and Data Auditability.

3.2. Architecture Overview

Data flow is characterized by three main phases: 1) knowledge retrieval; 2) context-aware retrieval-augmented generation; and 3) contextual embedding. The system consists of five functionally distinct components: a knowledge retrieval layer; a contextual embedding structure; a knowledge-retrieval-augmented-generation layer; and supervision modules that monitor feedback loops. Internal and external interfaces are defined at three levels—data supply, management, and querying. Context-maintenance mechanisms cover data provenance, historical-traceability requirements, and the handling of potentially malicious prompts.

Context preservation is enabled through data validation, temporality assurance, real-time supervision, and additional provenance tags. Knowledge sources are scrutinized for their sensitivity to hallucination and are updated—via revision rules—only when the confidence of predictions drops below a certain threshold. For cases requiring a rollback, the embedding structure serves as a memory stack, preserving input-response records from which past embeddings can be retrieved and subsequently combined with new data in their original response windows.

3.3. Data Lifecycle and Provenance

The data lifecycle encompasses data collection, versioning, lineage, and auditability. Context-preserving context relies on memory structures embedded with contextual knowledge and a knowledge retrieval layer that maintains information freshness and relevance. To capture dynamic information, support real-time inference at low latency, and enable context preservation, a context-preserving framework features persistent memory structures augmented by a knowledge retrieval layer. The data ingested by the knowledge retrieval layer serves as the underlying context for this layer, responding to knowledge requests seeking domain knowledge such as terminology definitions, rules, regulations, and framework descriptions. Functioning in production mode, the knowledge retrieval layer operates in a bottom-up manner, often guided by an LLM, dynamically injecting fresh context information into the underlying memory structures.

Maintaining the provenance of the data collected and utilized ensures traceability and auditability, enabling data scrutiny and validation for roles abiding by data governance and compliance frameworks. Data provenance delineates the lifecycle of data across every step of the data lifecycle, facilitating verification of data source authenticity and acknowledgement in its use. Data versioning preserves the history of data changes, serving an audit role at every stage of the data lifecycle, enabling rollback if required, and verifying compliance with privacy regulations and policies. Data lineage tracks data flow from origin to its final destination or consequence. Periodic, automatic versioning, lineage tracking, and logging of provenance acts as a governance control during the collection-persistent memory flow and validates compliance to regulations, directives, and policies.

4. Methods and Implementation

The realisation of the proposed framework entails a sequence of methodical operations that establish a context-preserving generative AI environment for real-time enterprise knowledge workflows. The context-preserving capability aligns the response generation with data provenance, audit trail, and compliance requirements. Control components ensure temporal alignment of the

knowledge-prompting data with the ongoing decision context. Largely automated processes govern memory creation and maintenance, taking into consideration the evolving socio-technological landscape of the enterprise and broader society.

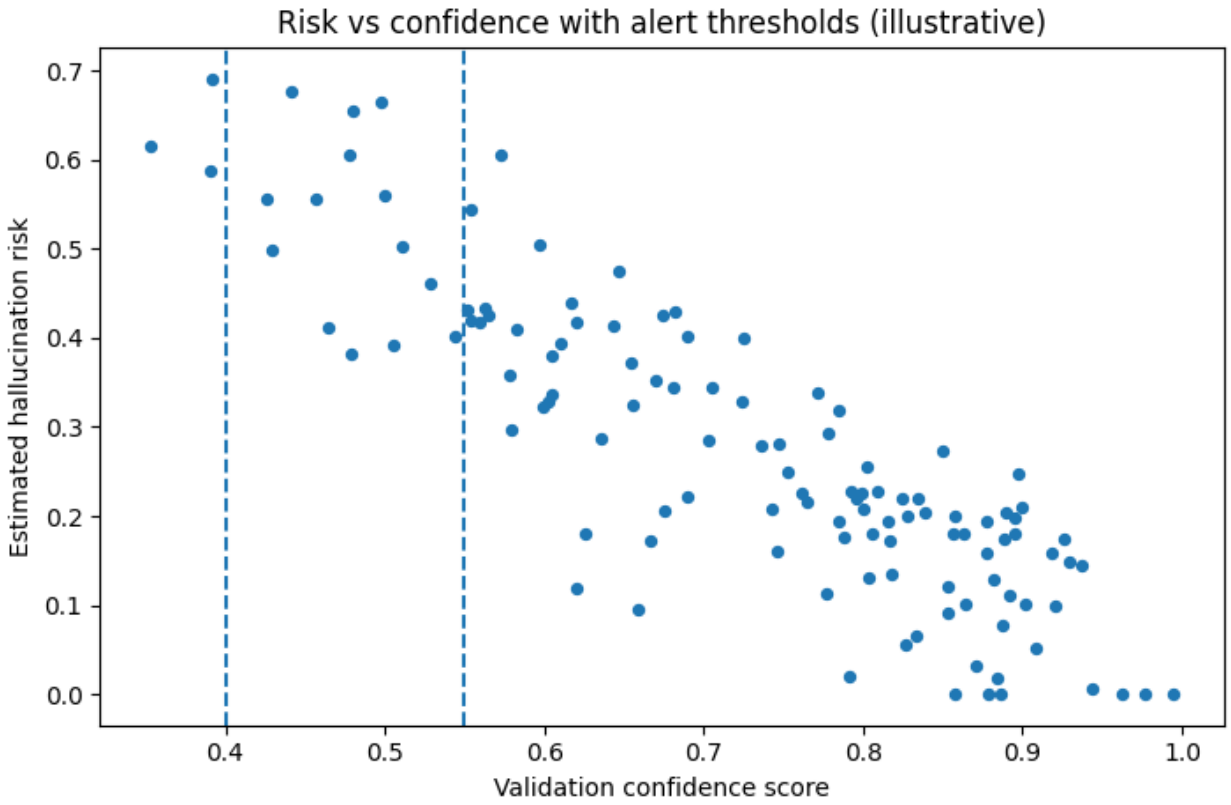
The memory stores contextual data retrieved from the enterprise data lake by a knowledge retrieval layer embedding aspect-based relevance to the querying process. A multi-layer storage mechanism provides context-adaptive representation, with the temporal perspective determining the contextual memory and embedding representations used at inference time. Embedded generative grammar aids management and control, enabling prompts for compliance and governance requirements. An automatic rollback facility establishes a feedback mechanism that can terminate the altered operation mode when the trigger becomes inactive. Compliance checks completed during the response generation process further monitor possible drift and deviation from regulatory and governance standards.

4.1. Knowledge Retrieval Layer

The retrieval layer enables the LLM-RAG framework to manage enterprise knowledge flow. Several aspects warrant careful consideration: (1) the design of information indexing, offering convincing trade-offs between retrieval quality and latency, (2) mechanisms ensuring and monitoring the quality of the retrieved context, and (3) the feasibility and effectiveness of empirical evaluations under conditions of limited time and resources.

The rationale for using Knowledge Management platforms is twofold: first, these systems already provide indexing and management functions capable of covering the needs of the considered context-preserving use cases; and second, the choice allows a direct measurement of the impact of the LLM with RAG on the context-preserving aspect of Enterprise Knowledge Workflows. The Risk Management and Compliance use case emphasizes the importance of context-preserving search-enabled beyond-LLM analysis of user prompt-transcript pairs by exploiting a Business Intelligence Data Warehouse.

ISSN: 3067-283X
AMERICAN JOURNAL OF ANALYTICS AND ARTIFICIAL INTELLIGENCE
VOLUME: 04 ISSUE: 01
RECEIVED: JANUARY 28 REVISED: FEBRUARY 04 ACCEPTED: FEBRUARY 23 PUBLISHED: MARCH 17



4.2. Contextual Embedding and Memory

Effective inferences within enterprises require contextual awareness of specific topological and temporal dimensions. Multi-dimensional embedding representations, trained on user-defined context building blocks, can support embedding slice query retrievals. Yet sufficient context volumes for generative query enrichments typically surpass retrieval system short-term memory boundaries. Lightweight real-time constructs based on temporal and contextual context requirement analysis allow effective long-term memory mechanism designs. Features of temporally.

and contextually classified data can become means for conditional long-term persistence rather than straightforward level-2 retention, reducing memory cost.

Multi-dimensional contextual embedding representation learning combines representations of training domain-specific knowledge bases across contexts (user-specific knowledge). Custom embedding for individual slices of use-case complex task building blocks can supplement training and improve performance without relying on too many components. For example, the need to use too many behavioural cues operating in multi-user discussion/dialog/chat settings to enable user-independence and

contextual de-bias can make a multi-dimensional approach implausible due to excessive behavioural shape and hence computational cost. In general, attempting to represent all behaviours/utterances across all tasks is burdensome, and slice-level representation learning and learning only for user/agent-specific segments can yield reliable slice-independent inferences.

With a real-time need to fulfil for action decision-agenda prompts, when input to the LLM for user task support, the provision of timely but low-cost context help is naturally attractive. Helper buildings, although using historical data, use them for direct provision of help only on current happenings rather than storing fine-grained temporal information explicitly since they are not maintained long-term for continuous viewing but directly retrieved through a query to the user-generated, user-function-agent-domain data summations. Support need, Cost, and Completeness are main aspects to satisfy.

Equation 2: Sparse retrieval scoring (BM25) and why it looks like it does

Step-by-step derivation (from TF → TF-IDF → BM25 intuition)

1. **Term frequency (TF)**: count how often term $t$appears in document d:

$$tf(t,d) = \text{count of } t \text{ in } d$$

2. **Inverse document frequency (IDF)**: terms appearing in many docs are less informative. If $N$is total docs and $df(t)$docs contain term t, then (one common form):

$$idf(t) = \log\left(\frac{N - df(t) + 0.5}{df(t) + 0.5} + 1\right)$$

3. **Why BM25 modifies TF**: raw TF grows without bound, but usefulness saturates. So we use a saturating function:

$$\frac{tf(t,d)\,(k_1 + 1)}{tf(t,d) + k_1}$$

where $k_1 > 0$controls saturation.

4. **Length normalization**: longer docs naturally have larger TF, so normalize by doc length $|\,d\,|$relative to average length avgdl:

$$tf(t,d) + k_1\left(1 - b + b\frac{|\,d\,|}{\text{avgdl}}\right)$$

where $b \in [0,1]$controls how strongly length is penalized.

5. **BM25 score for query** q: sum over query terms:

$$s_{\mathrm{BM25}}(q,d) = \sum_{t \in q} idf(t) \cdot \frac{tf(t,d)\,(k_1+1)}{tf(t,d) + k_1\left(1 - b + b\dfrac{|\,d\,|}{\mathrm{avgdl}}\right)}$$

4.3. Real-Time Inference and Governance

Real-time inference must meet operational latency goals while ensuring responsible usage of LLM-RAG capabilities. Monitoring mechanisms that track system performance against the defined temporal constraints can mitigate situations when response times exceed acceptable bounds, alerting users of potential compliance concerns. A list of failure-prone cases, disallowed by business rules and/or triggering control checks, can aid such an approach. Users can be allowed to prudently disregard composite risk indications or else the system can use the warning for output rollback — reverting to the knowledge base with an affirmative-response choice upon receiving an indicative LLM no. Users can validate the memorialization of genuinely useful inference contributions (embedding or chat memory) with respect to their personal knowledge or experience. Such validation can also support the designation of a much richer retrieval context for adjacent probable-use cases meeting the domain condition of clearly definable surrounding informational elements. Sensitive content thus detected can aid governance models requiring strict non-disclosure and specially track domain matches through teaching-effective agent setups.

Finally, multiple alternative responses may be produced within a balanced recall range and appropriately monitored for possible hallucination warning by introducing a validation score predicting the prospective confidence-standard of each candidate choice. Decision usage with compatible temporal frequency below a certain approval ratio can also indicate a collective sign-warning of lack of user trust and therefore appositive attention.

5. Use Cases and Scenarios

Enterprise decision support objectives span the spectrum of management responsibility, from operational execution to strategic planning and change management. A prioritization exercise identified the following short-term use cases for context-preserving LLM-RAG technology:

1. **Operational and Logistics Decision Support**: Order management for high-velocity, low-margin products requires fine-tuned execution processes. Data and metadata from enterprise resource planning (ERP) software provide structured input, and the LLM generates consecutive prompts for process execution—filling order details, projecting availability, and recommending transport modes and service providers. Ongoing implementation is validating and extending the concept for replenishment buying and inventory transfers. Successful production deployment will confirm the utility of contextual retrieval augmentation.

2. **Strategic Planning Scenarios**: Senior leadership teams engage in periodic long-range planning exercises. To improve the quality of debate, both scenario-building and forecasting techniques are under consideration. Scheduled updates to longterm scenarios generate an appropriate timing consideration for projecting high-potential zero-based budgets. The LLM-RAG system supports both functions. For scenario building, it creates draft narratives from risk and opportunity factors. For forecasting, it captures leading indicators and embeds description Probability-Value charts into presentation structures.

3. **Risk Management and Compliance Control**: The goal is to formalize and fortify the response pipeline for high-velocity, low-visibility conditions, setting detection ranges and control measures. Dedicated scanning identifies stealth data and information shifts within the ecosystem, while retrieval augmentation detects emerging threat narratives, assess probabilities, and identify ownership and response responsibilities. A contextualized audit loop documents the scanning input and output, monitors control effectiveness, and encodes learnings for training and simulation purposes.

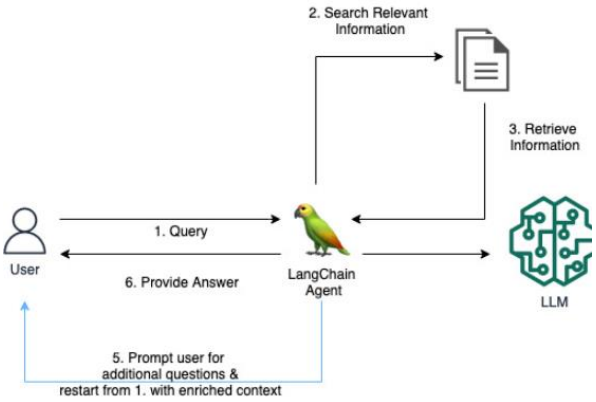


Fig 3: LLM responses in RAG use cases by interacting with the user

5.1. Operations and Logistics

Decision support for operations and logistics administrators hinges on effective knowledge retrieval from structured and unstructured enterprise data sources by a context-preserving LLM-RAG framework. Direct data pipelines provide real-time or near-real-time information into ongoing supply-chain operations, such as inventory levels and supplier performance. Prompt questions reflect strategic focus and control objectives for user-defined activities, such as analysis of transaction red-flags and upside opportunities.

Search, retrieval, and LLM inferencing for an operations business process require an end-to-end architecture (Figure 6). Structured operational data reported against fixed time schedules and periods are maintained in an enterprise data warehouse (EDW). Other enterprise data, like social media, financial market reports, and customer feedback, are collected into a data lake. Context-preserving embedding facilitates search and retrieval against both structured and unstructured data sources. As LLM inferencing is delayed, generated output is stored in a context-persistent LLM memory and provides knowledge to follow-on queries. Grouped queries further reduce LLM call volumes, improve LLM-response quality, and maintain response accuracy.

Supporting users' strategic and rebound-planning decisions requires LLM-RAG integration of longitudinal demand-and-supply data and business-model-market dynamics. Embedded memory caching retains context and makes shared alternative-generating prompts possible. A long-period focus supports scenarios like upcoming supply-market analysis, next-three-month forecasting, and overall market-change assessments for planning and rebound strategies. Support for end-to-end operational feedback

monitoring and prediction-control decisions is defined through the structural integration of regulatory feedback probes and LLM-prompted analytical work.

5.2. Strategic Planning

Strategic planning use cases focus on defining, pursuing, and monitoring long-term organizational objectives. Depending on the planning horizon, data content and shape can change markedly. This is especially true for forecasts resulting from statistical or machine-learning models. The illustration takes medium- and long-term planning as examples, where periodic assessments are accompanied by medium- and long-term feedback loops. Data source requirements currently covered include management- and business-intelligence dashboards for periodic evaluations, Wikipedia, and internal publications, which are relatively stable and highly structured.

Most application areas have common data sources and access workflows but differ in the prompt to the LLM. A governance process ensures that results are promptly checked by experts and bidder selection aligns closely with forecasting volatility. For highly volatile areas, the LLM explains the models it considered and why it discounted those with greater projected error within a monitoring case. For more stable situations, it supports forecasting recalibration with supporting analysis or literature to confirm potential changes within the planning context.

5.3. Risk Management and Compliance

Risk management and compliance help to protect the enterprise from disruptive events that may adversely impact performance. Regulatory compliance systems have to monitor compliance with complex rules that differ across jurisdictions. Risk management seeks to understand and manage threats before they occur, often by comparing the risk exposure to risk appetite. Risk exposure and appetite are, therefore, data-driven and can exhaustively analysed when all relevant data are made available. The use case for integrated risk management therefore requires LLM-RAG systems that preserve contextual provenance.

Management domain experts in operations and logistics oversee a locale of the enterprise and are responsible for ensuring that actual performance is aligned with planned performance. Stakeholders set objectives, budgets, and targets, and define risk appetite. Consensus for future environmental and operational conditions is developed through expert judgement with the aid of forecasting methods. Stress-testing selects a set of highly disrupted future conditions by modelling with the joint behaviours of key external drivers and an internal model of tactical operations.

6. Challenges, Risks, and Mitigations

Large language models can unintentionally reveal personally identifiable information (PII) present in their training data. Since these models are used over a read-only API, there is no possibility of training data filtering. While PII for training DataCUDA remain anonymised, the model can still result in similar problems. Information shared by enterprise employees, consultants, and contractors for collaboration purposes may also be leaked. Other DataCUDA data sources may contain sensitive information as well. Consequently, user input and model output are written to a compliance log for periodic review by an internal compliance team. A data classification framework allows for PII/tagging detection. Any information flagged is then redacted with an external service. The same framework is also leveraged for regulatory alignment. For instance, dataflows related to GDPR and HIPAA are protected per legal requirements.

Hallucinated information generated by LLMs is another major concern. An additional layer that monitors model reliability provides an estimate for how much the model should be trusted on a given inference. The system uses text-similarity algorithms from Google Natural Language Process to compute semantic similarity scores for the prompt-data pairs. The training DataCUDA contain labelled samples (hallucinations or detections). The LLM is periodically updated with new data and monitored for excessive hallucinations to guarantee a baseline level of reliability. Responses with low confidence estimates are flagged for inspection. A yellow alert triggers a review of all flagged output. A red alert indicates an uncommon halo of hallucinated responses and induces human validation imminently before the response is output.

Since LLMs are black-box systems, the inference results may seem whimsical. DocuGen helps address this aspect by generating a document to explain decision-making behind the content it produces. Statistical analysis monitors the quality of responses at a lower level and helps in the documentation process. Employees using the model also help generate the supporting congratulatory message to improve trust and transparency further.

Equation 3: Dense retrieval score and hybrid retrieval (sparse + dense)

Dense score (typically cosine)

$$c(q,d) = \cos(E(q), E(d))$$

Step-by-step hybrid score

Sparse scores (BM25) and dense cosine scores live on different scales, so:

1. **Normalize** each score to $[0, 1]$:

$$s' = \text{norm}(s_{\text{BM25}}), c' = \text{norm}(c)$$

2. **Weighted combination** with trade-off $\alpha \in [0,1]$:

$$\text{Hybrid}(q,d) = \alpha s' + (1-\alpha)c'$$

- Higher αfavors exact term match (precision on keywords).
- Lower αfavors semantic match (recall + paraphrases).

6.1. Data Privacy and Security

The use of external data to prompt generative capabilities introduces the potential for unauthorized interactions or information exposures. Some organizations choose not to leverage a generative engine because of legal and regulatory concerns. Other organizations operate in industries where hallucinations carry high costs and burdens, discouraging acceptance of the technology. Natural language models trained on public data combine the knowledge of billions of web pages in many contexts but often present fragments that sound credible but are not real. Anything or anybody imaginable can also be depicted visually or sonically. Abrupt hallucinations, improper bounds, unrealistic style, and inaccurate content can alarm users. A context-preserving system

senses, analyzes, and monetizes prompts in real time to overcome these challenges. When properly integrated into workflows, spelling out margins, scoping shifts, and performing all controls, the technology meets contextual prompt privacy and security requirements.

Compliance and custodianship functions are performed by an agent monitoring all generative tasks. The agent acts as a hold before any transmission, fulfilling all organizational policies assigned to it. Time is allocated to manually or automatically verify inputs and outputs. On-line learning mechanisms provide up-to-date checks as the adoption user community grows. Whenever risk or compliance management requires an on-line check, the model processing the request automatically enters “human-in-the-loop” mode. Proper integration with an established user support activity allows for low-latency assistance in these conditions. The integration of other risk-management or compliance gates certainly raises response times but, when correctly tuned, decreases the workload for decision institutions while balancing operational and market risk together with the constrained appetite.

6.2. Hallucination and Reliability

Concerns about the reliability of generative AI, especially regarding the appearance of hallucinations, are central to the governance of any system embedding such capabilities. As ChatGPT commonly notes when asked about its accuracy, it is “best not to trust [the model’s] output without verification.” Nevertheless, no retrieval-augmented generative AI system is immune to generating fictional output, either when facts are fabricated, duplicated, or distorted or when the result is illogical but language-like.

A challenging characteristic of hallucinations in response to retrieval-augmented queries is that a significant percentage arise from legitimate facts, leading to trust across queries despite a substantial false statement rate. Consequently, accuracy must be considered not just in a simple binary sense but also in terms of the expected quality of the answer based on the accuracy sampling that indicates the proportion of reliable answers. An out-of-the-box ChatGPT trained with the GPT-4 model typically generates factually good answers to online research topics, and the LLM-RAG system prototype can similarly exploit external web data.

Nevertheless, for critical tasks such as management decision support and compliance where hallucination risks cannot be tolerated, careful validation of relevant prompt-answer pairs needs to be performed for the deployed LLM-RAG system. Hallucination monitoring by human users can also contribute to curbing unreliability on operations and logistics tasks by enabling querying of generative replies and corrections when necessary. Query-based hallucination monitoring should be triggered for generation of non-obvious or uncommon replies such as reasoning tasks (inferred probabilistically based on how rare such prompting has been) or other prompts that not been observed or tagged by users in testing. To minimize frequent checking, the tagging should also note if the reply is deemed trustworthy or not; tagging of untrustworthy responses should be also associated with the diagnosis of underlying cause or escape branch (e.g., inaccurate contextual memory).

6.3. Interpretability and Trust

Interpretation and trust represent ongoing challenges for Generative AI technology in general and LLM-RAG systems in particular. Text generation is often presented as a black box, with little insight on how the outputs are generated, particularly in RAG configurations. Information incompleteness, inaccuracies, or outright fabrication or denial of foundational events—termed

"hallucination"—erodes user confidence. LLM-temperament testing—where prompt content is altered to demonstrate desired response patterns—has been proposed as a tool to address Hallucination but this only partially mitigates the risk.

Mitigating these shortcomings is crucial in a management decision-support context. Critical business operation processes and decisions rely upon injected prompt queries to generate output that is contextually relevant, useful, noise-free, precise, and trustworthy. In enterprise settings, users therefore expect RAG output to have logical consistency with proven and trusted core data and documents. Moreover, users must also have trust in the LLM-RAG system, particularly when it comes to governance and control. To facilitate both, audit and documentation processes have been integrated into the LLM-RAG concept, allowing for real-time oversight of generated LLM output and decision-support prompts.

7. Evaluation and Validation

Evaluation encompasses experimentation, qualitative assessment of implementation suitability, and an assessment of the integration's impact on the authoring experience within decision-support use cases. Consistency and Responsiveness are used to measure integration fidelity and generation quality, respectively.

Investigating context preservation during retrieval-augmented generation performance utilizes pipeline response latency as a proxy for context effectiveness. Inputs originate via two routes: intent prompts directly based on data indexed during the implementation; and management-driven What-If prompts addressing Risk & Compliance functions. The simulation uses a text-based interface to minimize device noise and optimize message length while controlling for task complexity in latency evaluations.

Evaluating support for the Enterprise Knowledge Workflow Framework builds on a real-time experimental decision-support setup within Operations & Logistics. Here, a fictitious Chemicals organization, producing and distributing a specialist chemical product line to third-party manufacturers, attempts to use simulation tooling to develop and elaborate forecasts in advance of materials-safety audits for that product class. Simulation is followed by a synthesis of engagement feedback and foundry tooling learning.

7.1. Experimental Setup

The evaluation comprises an experimental setup for an end-to-end usage scenario, an assessment of retrieval-quality characteristics, and benchmarking against several text-generation models. Deployment is in the Azure cloud on infrastructure suitable for low-latency real-time processing. Documents supporting operations and logistics management (e.g., S&OP, sales forecasting, inventory replenishment, freight cost estimation, material sourcing, and procurement) are sourced from Capgemini Research, Gartner Research, Gartner, Forrester, and Edgar Schein. Historical reports, forecasts, and requirements specifications are collected from a large engineering firm. The supply-chain domain is augmented by Thomson Reuters Regulatory Intelligence Risk Data: Risk Management, Responding to Global Risk, and Communications. Production prompts query categories, supported response types, and example queries.

The core language model is Azure OpenAI Service GPT-35-Turbo. An ensemble of retrieval-augmented-generation models—Llama 2, Llama 2-Chat, Mistral 7B, and LLaMA 2-Chat—integrates with Wit.ai, Flan T5-small, and Mistral 7B-Chat. PaLM 2,

ISSN: 3067-283X
AMERICAN JOURNAL OF ANALYTICS AND ARTIFICIAL INTELLIGENCE
VOLUME: 04 ISSUE: 01
RECEIVED: JANUARY 28 REVISED: FEBRUARY 04 ACCEPTED: FEBRUARY 23 PUBLISHED: MARCH 17

Claude, and ultra-large language models (ULLMs)—GPT-4, Bloomz-70B-MT (multilingual task-oriented), and MPT-30B—facilitate human-in-the-loop evaluation. Quality and reliability are controlled via audience selection (close domain), dialog-context utilization, and reference data curation (document retrieval). Performance is assessed on generation quality, factual accuracy, and risk level. Expert auditors provide external provenance-monitoring evaluations.

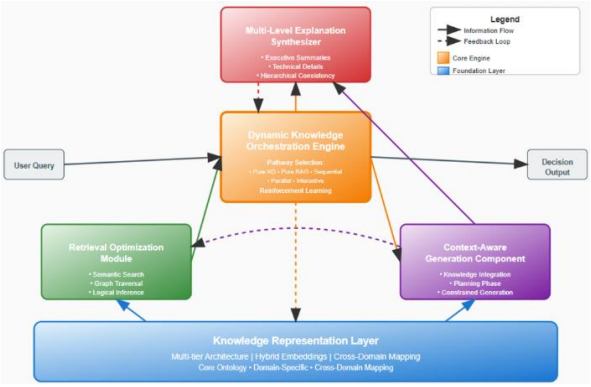


Fig 4: Experimental Setup of Context-Preserving Generative AI

7.2. Benchmarking and Results

A wide spectrum of benchmarks assesses LLM capabilities across diverse topics and tasks. Standard datasets includeSQuAD v1 and v2 (question answering); BoolQ, C3, and Hellaswag (multiple-choice); CoQA and QuAC (conversational QA); and CNN-Daily Mail, Gigaword, and XSum (summarization). Other challenging datasets for large parameter non-autoregressive models include WorldTree (textual inference) and LAMBADA (language modeling). BERT and its descendants excel in simulated open-book QA, semantic textual similarity, word sense disambiguation, paraphrase identification, and entailment tasks but struggle elsewhere, with marked performance drops for zero-shot tasks. Prompting shows effectiveness in few-shot scenarios, with considerable complexity in designing implicit prompts. Generative CV models achieve state-of-the-art results in domain-specific tasks such as satellite image classification and pedestrian trajectory generation. Prompt-based zero-shot tests reveal performance drops across vision tasks.

A non-exhaustive selection of representative LLMs, settings, and results on few-shot web-based QA is considered, encompassing different model classes and scales. Three standard April Fool's release models—GPT-4, Jurassic-2-German, Text-DaVinci-003—and FinGPT, an open-specialist advisor tuned on Wall Street data, are combined with the Core Dataset (En-Fr-Sp) for RAG in the banking domain. Robustness of the hybrid approach is illustrated with common 314-BT extracted from the SuperGLUE benchmark, employing the de-facto finetuned T5 style task formulation becuase it combines versatility, difficulty, and best performance. Full diverse predictions over Intel CPU with six 256-core nodes in 30 minutes.

7.3. Human-in-the-Loop Assessment

Use of language models still presents various challenges—even risk factors—so human-in-the-loop assessment can provide additional comfort. Models can hallucinate or behave undesirably in apparently basic scenarios. Tuning, validations, rules,

controls, and safeguards can mitigate such issues, particularly but not exclusively in business-critical environments. Understanding and addressing those risks motivates work in this space.

A proof-of-concept prototype that combines existing data sources and helps make key business decisions is ready. It implements various safeguards to limit key-holding risks, either by operationalizing or supporting task execution. Human monitoring detects biases and issues, adds a second opinion, and reduces reliance on model accuracy—the double-check option covers any hallucinations using the underlying dataset. Moreover, the system tracks and records historic generated content. An initial study evaluates components of the real-time LLM-RAG prototype—for generative AI in management decision support—using business domain practitioners and students in an operations management context. Building participants' expertise with the LLM-RAG prototype precedes evaluation.

8. Conclusion

Management decision support for real-time business processes relies on effective retrieval of up-to-date context information. The proposed context-preserving LLM-RAG framework for Enterprise Knowledge Workflows ensures that prompts for latent LLMs are augmented by the most relevant data before inference—and that context provenance is recorded for governance externally, alongside the archived response.

Automated support for management decisions helps organizations meet business objectives efficiently, adapting operations to react to unexpected events such as operational upsets or geopolitical shocks. Some enterprise functions and activities require more than a debrief of past information; they require suitable knowledge to inform real-time actions and exigent decision-making that adhere to policy. Technologies aligning with these needs must meet low latency and compliance constraints while maintaining a direct connection to the enterprise's knowledge base. Various aspects are covered: the kind of context that needs to be efficiently captured, updated, and accessible; policy considerations that govern the use of AI in the enterprise; the real-time nature of management action and strategies for validating AI outputs; approaches to ensure safe use of these tools, privacy, and regulatory alignment; use cases from operations, strategy, and risk-and-compliance perspectives; use scenarios for GFP's operational hub; and identified risk-and-compliance contexts.

8.1. Future Trends

Entering into the future integration of generative artificial intelligence in enterprise workflows, the attention of researchers shifts to the potential of Generative AI systems in the domain of natural language understanding and generation. With the rapid development of foundation models and their popularization through apps like ChatGPT a new wave of Data-Intensive Machine Learning technology, ready to revolutionize various knowledge industry tasks. However, the main approaches provide reliable output only if high-quality fine-tuning data is available, therefore Enterprise Knowledge Domains still need a sound method to govern hallucination and ensure the results used by end-users are accurate. While RAG approaches improve the retrieval of processors with a knowledge base of snippets of Information Graphs, they serve more as enhancing layers due to their native lack of provenance management, floating memory capability and Real Time Latency control process.

To create a powerful automated assistant for enterprise agents with Latency, Accuracy, Governance and Compliance constraints, the first challenge was described to be Context Preservation and then bridged to Enterprise Knowledge Workflows. Decision

Support Use Cases were highlighted and the LLM-RAG concept presented, framed as preserving context, provenance and real-time responds. The main idea is to match the hole context background assisting a managing person decision with a open question to its Langauage Model ChatGPT-3.5-RAG System in such way when sending the prompt , the model receive a big contextual embedding with all the important facts considered relevant in a human perspective sense, as well as preserving the training to ensure audit and roll back capability when needed.

9. References

[1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

[2] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-T. (2020). Dense passage retrieval for open-domain question answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 6769–6781.

[3] Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2022). Unsupervised dense information retrieval with contrastive learning. Transactions of the Association for Computational Linguistics, 10, 1507–1524.

[4] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv.

[5] Fan, W., Ding, Y., Ning, Z., Liang, S., Wang, S., Li, H., Yin, D., & Tang, J. (2024). A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. Proceedings of the ACM on Management of Data, 2(2), 1–28.

[6] Gupta, S., Sharma, A., Singh, S., & Agarwal, A. (2024). A comprehensive survey of retrieval-augmented generation (RAG): Evolution, architectures, and applications. arXiv.

[7] Asai, A., Wu, Z., Wang, Y., Hajishirzi, H., & Yih, W.-T. (2023). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. arXiv.

[8] Luo, M., Zhang, Y., Zhu, Z., Chen, D., & Li, J. (2023). RARR: Researching and revising answers with retrieval. arXiv.

[9] Shuster, K., Piktus, A., Petroni, F., Kiela, D., Lewis, P., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. arXiv.

[10] Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Sutawika, L., & others. (2023). Augmented language models: A survey. arXiv.

[11] Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., & Dolan, B. (2020). DialogPT: Large-scale generative pre-training for conversational response generation. Proceedings of ACL 2020: System Demonstrations, 270–278.

[12] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv.

[13] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., others. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv.

[14] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., others. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

[15] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., & others. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv.

[16] Ziegler, D. M., Stiennon, N., Wu, J., Brown, T., Radford, A., Amodei, D., Christiano, P. F., & Irving, G. (2019). Fine-tuning language models from human preferences. arXiv.

[17] Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Stoica, I., & others. (2023). Efficient memory management for large language model serving with PagedAttention. Proceedings of the 29th ACM Symposium on Operating Systems Principles, 735–750.

[18] Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. Proceedings of ACL-IJCNLP 2021, 4582–4597.

[[19] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations (ICLR).

[20] Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. Proceedings of EMNLP 2021, 3045–3059.

[21] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186.

[22] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., others. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

[23] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.

[24] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. International Conference on Learning Representations (ICLR).

[25] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems, 35, 22199–22213.

[26] Nair, V., Schumacher, E., Tiwary, R., Singh, A., Zhang, Y., & others. (2023). Datasets and evaluation for retrieval augmented generation: A survey. arXiv.

[27] Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., others. (2022). Improving language models by retrieving from trillions of tokens. International Conference on Machine Learning (ICML), 2206–2240.

[28] Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). REALM: Retrieval-augmented language model pre-training. Proceedings of ICML 2020, 3929–3938.

[29] Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. Proceedings of SIGIR 2020, 39–48.

[30] Wang, L., Zhang, Y., & others. (2023). RAG-based enterprise question answering: Architectures, evaluation, and deployment lessons. arXiv.

[31] Zhu, Z., Luo, M., Li, J., & Chen, D. (2024). Evaluating citation grounding and factual consistency in retrieval-augmented generation. arXiv.

[32] United States National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology.

[33] European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

[34] Shah, N., Jain, S., & others. (2024). Secure retrieval-augmented generation for enterprise knowledge: Access control, redaction, and auditability. arXiv.

[35] Chen, L., Zaharia, M., & Zou, J. (2023). How is ChatGPT's behavior changing over time? arXiv.

[36] Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. Proceedings of ACL 2020, 1906–1919.

[37] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.

[38] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., & others. (2022). OPT: Open pre-trained transformer language models. arXiv.

[39] Hope, T., Portenoy, J., Vasan, K., Borchardt, J., Horvitz, E., Weld, D. S., & others. (2022). SciFact: A benchmark for scientific claim verification. Proceedings of AAAI 2022, 12301–12308.